

CLiC-it 2019

The Sixth Italian Conference on Computational Linguistics

Proceedings of the Conference

November 13-15, 2019

Copyright ©2019 for the individual papers by the papers' authors.
Copyright ©2019 for the volume as a collection by its editors.
This volume and its papers are published under the Creative Commons License Attribution 4.0 International (CC BY 4.0).

These proceedings have been published in the CEUR Workshop Proceedings series.
The original papers are available at: <https://ceur-ws.org/Vol-2481>.
The papers are mirrored in the ACL Anthology.

ISBN 979-1-280-13600-8

Table of Contents

<i>Preface</i>	
Raffaella Bernardi, Roberto Navigli and Giovanni Semeraro	1
<i>Visually-Grounded Dialogue Models: Past, Present, and Future</i>	
Raquel Fernández	11
<i>Impossible Languages and the Architecture of Human Grammars</i>	
Andrea Moro	12
<i>Recognizing and Reducing Bias in NLP Applications</i>	
Dirk Hovy	13
<i>Prerequisite or Not Prerequisite? That's the Problem! An NLP-based Approach for Concept Prerequisite Learning</i>	
Chiara Alzetta, Alessio Miaschi, Giovanni Adorni, Felice Dell'Orletta, Frosina Koceva, Samuele Passalacqua and Ilaria Torre	14
<i>An Empirical Analysis of Linguistic, Typographic, and Structural Features in Simplified German Texts</i>	
Alessia Battisti, Sarah Ebling and Martin Volk	22
<i>Almawave-SLU: A New Dataset for SLU in Italian</i>	
Valentina Bellomaria, Giuseppe Castellucci, Andrea Favalli and Raniero Romagnoli	29
<i>Nove Anni di jTEI: What's New?(Nine Years of jTEI: What's New?)</i>	
Federico Boschetti, Gabriella Pardelli and Giulia Venturi	35
<i>BullyFrame: Cyberbullying Meets FrameNet</i>	
Silvia Brambilla, Alessio Palmero Aprosio and Stefano Menini	41
<i>Lost in Text. A Cross-Genre Analysis of Linguistic Phenomena within Text</i>	
Chiara Buongiovanni, Francesco Gracci, Dominique Brunato and Felice Dell'Orletta	49
<i>Annotating Shakespeare's Sonnets with Appraisal Theory to Detect Irony</i>	
Nicolò Busetto and Rodolfo Delmonte	55
<i>Embeddings Shifts as Proxies for Different Word Use in Italian Newspapers</i>	
Michele Cafagna, Lorenzo De Mattei and Malvina Nissim	63
<i>Suitable Doesn't Mean Attractive. Human-Based Evaluation of Automatically Generated Headlines</i>	
Michele Cafagna, Lorenzo De Mattei, Davide Bacciu and Malvina Nissim	70
<i>There and Back Again: Cross-Lingual Transfer Learning for Event Detection</i>	
Tommaso Caselli and Ahmet Üstün	77
<i>PESInet: Automatic Recognition of Italian Statements, Questions, and Exclamations With Neural Networks</i>	
Sonia Cenceschi, Roberto Tedesco, Licia Sbattella, Davide Losio and Mauro Luchetti	85
<i>What Makes a Review helpful? Predicting the Helpfulness of Italian TripAdvisor Reviews</i>	
Giulia Chiriatti, Dominique Brunato, Felice Dell'Orletta and Giulia Venturi	92
<i>Is This an Effective Way to Annotate Irony Activators?</i>	
Alessandra Teresa Cignarella, Manuela Sanguinetti, Cristina Bosco and Paolo Rosso	98

<i>Robospierre, an Artificial Intelligence to Solve “La Ghigliottina”</i>	
Nicola Cirillo, Chiara Pericolo and Pasquale Tufano	106
<i>From Sartre to Frege in Three Steps: A* Search for Enriching Semantic Text Similarity Measures</i>	
Davide Colla, Marco Leontino, Enrico Mensa and Daniele P. Radicioni	113
<i>Is “manovra” Really “del popolo”? Linguistic Insights into Twitter Reactions to the Annual Italian Budget Law</i>	
Claudia Roberta Combei	121
<i>Cross-Platform Evaluation for Italian Hate Speech Detection</i>	
Michele Corazza, Stefano Menini, Elena Cabrio, Sara Tonelli and Serena Villata	129
<i>An Open Science System for Text Mining</i>	
Gianpaolo Coro, Giancarlo Panichi and Pasquale Pagano	136
<i>Detecting Irony in Shakespeare’s Sonnets with SPARSAR</i>	
Rodolfo Delmonte and Nicolò Busetto	143
<i>Towards an Italian Learner Treebank in Universal Dependencies</i>	
Elisa Di Nuovo, Cristina Bosco, Alessandro Mazzei and Manuela Sanguinetti	151
<i>Building an Italian Written-Spoken Parallel Corpus: a Pilot Study</i>	
Elisa Dominutti, Lucia Pifferi, Felice Dell’Orletta, Simonetta Montemagni and Valeria Quochi	159
<i>Italian and English Sentence Simplification: How Many Differences?</i>	
Martina Fieromonte, Dominique Brunato, Felice Dell’Orletta and Giulia Venturi	166
<i>Error Analysis in a Hate Speech Detection Task: The Case of HaSpeeDe-TW at EVALITA 2018</i>	
Chiara Francesconi, Cristina Bosco, Fabio Poletto and Manuela Sanguinetti	172
<i>Nunc Est Aestimandum: Towards an Evaluation of the Latin WordNet</i>	
Greta Franzini, Andrea Peverelli, Paolo Ruffolo, Marco Passarotti, Helena Sanna, Edoardo Signoroni, Viviana Ventura and Federica Zampedri	179
<i>Iride: an Industrial Perspective on Production Grade End to End Dialog System</i>	
Cristina Giannone, Valentina Bellomaria, Andrea Favalli and Raniero Romagnoli	187
<i>When Lexicon-Grammar Meets Open Information Extraction: a Computational Experiment for Italian Sentences</i>	
Raffaele Guarasci, Emanuele Damiano, Aniello Minutolo and Massimo Esposito	194
<i>Are Subtitling Corpora really Subtitle-like?</i>	
Alina Karakanta, Matteo Negri and Marco Turchi	201
<i>Asymmetries in Extraction From Nominal Copular Sentences: a Challenging Case Study for NLP Tools</i>	
Paolo Lorusso, Matteo Greco, Cristiano Chesi and Andrea Moro	207
<i>Objective Frequency Values of Canonical and Syntactically Modified Idioms: Preliminary Normative Data</i>	
Azzurra Mancuso and Alessandro Laudanna	215
<i>Gender Detection and Stylistic Differences and Similarities between Males and Females in a Dream Tales Blog</i>	
Raffaele Manna, Antonio Pascucci and Johanna Monti	221

<i>CROATPAS: A Resource of Corpus-derived Typed Predicate Argument Structures for Croatian</i> Costanza Marini and Elisabetta Jezek	228
<i>Enhancing a Text Summarization System with ELMo</i> Claudio Mastronardo and Fabio Tamburini	236
<i>KIParla Corpus: A New Resource for Spoken Italian</i> Caterina Mauri, Silvia Ballarè, Eugenio Gorla, Massimo Cerruti and Francesco Suriano	243
<i>Evaluating Speech Synthesis on Mathematical Sentences</i> Alessandro Mazzei, Michele Monticone and Cristian Bernareggi	250
<i>Automated Short Answer Grading: A Simple Solution for a Difficult Task</i> Stefano Menini, Sara Tonelli, Giovanni De Gasperis and Pierpaolo Vittorini	257
<i>Games for Learning Old and Special Alphabets – The Case Study of Gamifying Mrežnik</i> Josip Mihaljević	264
<i>Text Frame Detector: Slot Filling Based On Domain Knowledge Bases</i> Martina Miliani, Lucia C. Passaro and Alessandro Lenci	270
<i>Defining Action Types: on the roles of Thematic Structures and Argument Alternations</i> Massimo Moneglia, Alessandro Panunzi and Rossella Varvara	278
<i>HateChecker: a Tool to Automatically Detect Hater Users in Online Social Networks</i> Cataldo Musto, Angelo Sansonetti, Marco Polignano, Giovanni Semeraro and Marco Stranisci	285
<i>The Contribution of Embeddings to Sentiment Analysis on YouTube</i> Moniek Nieuwenhuis and Malvina Nissim	291
<i>A Novel Integrated Industrial Approach with Cobots in the Age of Industry 4.0 through Conversational Interaction and Computer Vision</i> Andrea Pazienza, Nicola Macchiarulo, Felice Vitulano, Antonio Fiorentini, Marco Cammisà, Leonardo Rigutini, Ernesto Di Iorio, Achille Globo and Antonio Trevisi	298
<i>Annotating Hate Speech: Three Schemes at Comparison</i> Fabio Poletto, Valerio Basile, Cristina Bosco, Viviana Patti and Marco Stranisci	304
<i>ALBERTo: Italian BERT Language Understanding Model for NLP Challenging Tasks Based on Tweets</i> Marco Polignano, Pierpaolo Basile, Marco de Gemmis, Giovanni Semeraro and Valerio Basile	312
<i>Evaluating the MuMe Dialogue System with the IDIAL protocol</i> Aureliano Porporato, Alessandro Mazzei, Daniele P. Radicioni and Rosa Meo	318
<i>To be Fair: a Case for Cognitively-Inspired Models of Meaning</i> Simon Preissner and Aurélie Herbelot	325
<i>The Impact of Self-Interaction Attention on the Extraction of Drug-Drug Interactions</i> Luca Putelli, Alfonso Emilio Gerevini, Alberto Lavelli and Ivan Serina	332
<i>Enriching Open Multilingual Wordnets with Morphological Features</i> Stefania Racioppa and Thierry Declerck	339
<i>A Comparison of Representation Models in a Non-Conventional Semantic Similarity Scenario</i> Andrea Amelio Ravelli, Oier Lopez de Lacalle and Eneko Agirre	345

<i>How Much Competence Is There in Performance? Assessing the Distributional Hypothesis in Word Bigrams</i>	
Johann Seltsmann, Luca Ducceschi and Aurélie Herbelot	353
<i>Jointly Learning to See, Ask, Decide when to Stop, and then GuessWhat</i>	
Ravi Shekhar, Alberto Testoni, Raquel Fernández and Raffaella Bernardi	361
<i>Creating a Multilingual Terminological Resource using Linked Data: the case of Archaeological Domain in the Italian language</i>	
Giulia Speranza, Carola Carlino and Sina Ahmadi	367
<i>Vir is to Moderatus as Mulier is to Intemperans - Lemma Embeddings for Latin</i>	
Rachele Sprugnoli, Marco Passarotti and Giovanni Moretti	374
<i>Valutazione umana di Google Traduttore e DeepL per le traduzioni di testi giornalistici dall'inglese verso l'italiano (Human evaluation of Google Translator and DeepL for translations of journalistic texts from English into Italian)</i>	
Mirko Tamosanis	381
<i>Prendo la Parola in Questo Consesso Mondiale: A Multi-Genre 20th Century Corpus in the Political Domain</i>	
Sara Tonelli, Rachele Sprugnoli and Giovanni Moretti	388
<i>Reflexives, Impersonals and Their Kin: a Classification Problem</i>	
Kledia Topciu and Cristiano Chesi	396
<i>Annotation and Analysis of the PoliModal Corpus of Political Interviews</i>	
Daniela Trotta, Sara Tonelli, Alessio Palmero Aprosio and Annibale Elia	403
<i>Analyses of Literary Texts by Using Statistical Inference Methods</i>	
Mehmet Can Yavuz	410
<i>Neural Semantic Role Labeling using Verb Sense Disambiguation</i>	
Domenico Alfano, Roberto Abbruzzese and Donato Cappetta	417
<i>Kronos-it: a Dataset for the Italian Semantic Change Detection Task</i>	
Pierpaolo Basile, Giovanni Semeraro and Annalina Caputo	423
<i>How do Physiotherapists and Patients talk? Developing and annotating RiMotivAzione dialogue corpus.</i>	
Andrea Bolioli, Francesca Alloatti, Mariafrancesca Guadalupi, Roberta Iolanda Lanzi, Giorgia Pregnotato and Andrea Turolla	429
<i>Standardizing Language with Word Embeddings and Language Modeling in Reports of Near Misses in Seveso Industries</i>	
Simone Bruno, Silvia Maria Ansaldi, Patrizia Agnello and Fabio Massimo Zanzotto	436
<i>Computational Linguistics Against Hate: Hate Speech Detection and Visualization on Social Media in the Contro L'OdioProject</i>	
Arthur T. E. Capozzi, Mirko Lai, Valerio Basile, Cataldo Musto, Marco Polignano, Fabio Poletto, Manuela Sanguinetti, Cristina Bosco, Viviana Patti, Giancarlo Ruffo, Giovanni Semeraro and Marco Stranisci	442
<i>Supporting Journalism by Combining Neural Language Generation and Knowledge Graphs</i>	
Marco Cremaschi, Federico Bianchi, Andrea Maurino and Andrea Primo Pierotti	448

<i>Deep Bidirectional Transformers for Italian Question Answering</i>	
Danilo Croce, Giorgio Brandi and Roberto Basili	454
<i>Applying Psychology of Persuasion to Conversational Agents through Reinforcement Learning: an Exploratory Study</i>	
Francesca Di Massimo, Valentina Carfora, Patrizia Catellani and Marco Piastra	460
<i>WebIsAGraph: A Very Large Hypernymy Graph from a Web Corpus</i>	
Stefano Faralli, Irene Finocchi, Simone Paolo Ponzetto and Paola Velardi	466
<i>CULTURE as a ‘Liquid’ Modern Word. Evidence from Synchronic and Diachronic Language Resources</i>	
Maristella Gatto	472
<i>A Dataset of Real Dialogues for Conversational Recommender Systems</i>	
Andrea Iovine, Fedelucio Narducci and Marco de Gemmis	478
<i>Quanti anni hai? Age Identification for Italian</i>	
Aleksandra Maslennikova, Paolo Labruna, Andrea Cimino and Felice Dell’Orletta	484
<i>Multi-task Learning Applied to Biomedical Named Entity Recognition Task</i>	
Tahir Mehmood, Alfonso Gerevini, Alberto Lavelli and Ivan Serina	490
<i>Mining Italian Short Argumentative Texts</i>	
Ivan Namor, Pietro Totis, Samuele Garda and Manfred Stede	496
<i>Fixing Comma Splices in Italian with BERT</i>	
Daniele Puccinelli, Silvia Demartini and Renée E. D’Aoust	502
<i>A Comparative Study of Models for Answer Sentence Selection</i>	
Federico Rossetto, Alessio Gravina, Silvia Severini and Giuseppe Attardi	508
<i>An Italian Question Answering System for Structured Data based on Controlled Natural Languages</i>	
Lucia Siciliani, Pierpaolo Basile, Giovanni Semeraro and Matteo Mennitti	514
<i>The Tenuousness of Lemmatization in Lexicon-based Sentiment Analysis</i>	
Marco Vassallo, Giuliano Gabrieli, Valerio Basile and Cristina Bosco	520



Preface

It is our great pleasure to welcome you to CLiC-it 2019 (clic2019.di.uniba.it/), the Sixth Italian Conference on Computational Linguistics, held between November 13th and 15th in Bari, hosted and locally organized by Università degli Studi di Bari Aldo Moro.

The CLiC-it conference series is an initiative of the Italian Association for Computational Linguistics (AILC) which, after six years of activity, has clearly established itself as the premier national forum for research and development in the fields of Computational Linguistics and Natural Language Processing, where leading researchers and practitioners from academia and industry meet to share their research results, experiences, and challenges.

The maturity of the conference is reflected by the quality of the submitted works. We would like to take this opportunity to warmly thank all the authors for submitting their original research. This year CLiC-it received 82 submissions, confirming its increasing trend (from 64 submissions in 2015 to 70 in 2018).

The Program Committee worked very hard to ensure that every paper received at least two careful and fair reviews, with the 69.51% of the papers which received three or even more reviews. This process finally led to the acceptance of 20 papers for oral presentation and 55 papers for poster presentation, with a global acceptance rate of 91.46% motivated by the inclusive spirit of the conference.

That process involved 34 Area Chairs and 209 Program Committee members. They were assisted by 4 additional reviewers. We are extremely grateful to all the PC members and reviewers for producing 238 detailed and insightful reviews.

The conference is also receiving considerable attention from the international community, with 26 (31.71%) submitted papers showing at least one author affiliated to a foreign institution, of which 24 accepted (32%). This amounts to a total of 41 authors over 252 (16.33%) affiliated to 14

foreign countries: Croatia, Denmark, France, Germany, Ireland, Luxembourg, Malta, Netherlands, Romania, Russia, Spain, Switzerland, Turkey, and United States.

Regardless of the format of presentation, all accepted papers are included in the proceedings equally and are available as open access publication. In line with previous editions, the conference is organised around thematic areas managed by two chairs per area.

In addition to the technical program, this year we have two invited talks and a tutorial on different topics, showing the interdisciplinary spirit of our research community. We are very grateful to both Raquel Fernández (University of Amsterdam) for agreeing to share with the Italian Computational Linguistics community her knowledge on visually grounded dialogue models, and to Andrea Moro (Scuola Universitaria Superiore IUSS Pavia) for sharing his expertise on the architecture of human grammars, as well as to Dirk Hovy for his tutorial on the problem of bias in Natural Language Processing applications.

As in the previous edition of the conference, we organised a special track called "Research Communications", encouraging authors of articles published in 2019 at outstanding international conferences in our field to submit short abstracts of their work. Research communications are not published in the proceedings, but they are orally presented within a dedicated session at the conference, in order to enforce dissemination of excellence in research. We received 10 submissions and could include 6 of them in the program.

Finally, the program includes a panel discussion on Ethical issues in Natural Language Processing chaired by Alessandro Lenci (University of Pisa). The goal of the panel is to foster a discussion on some key ethical topics in NLP research and applications, with a focus on their impact on the Italian community. Themes of the panel include negative stereotypes in data-driven computational models; sustainability of data- and resource-intense NLP; the impact of NLP technology in digital society; privacy and NLP, among other crucial questions.

Traditionally, around one half of the participants at CLiC-it are young postdocs, PhD students, or even undergraduate students. Following the tradition of past years, a prize will be given to the best paper among those whose first author is a student. This year, the best paper will be selected among 14 oral papers and 30 papers presented as posters.

Moreover, during the conference we award the prize for the best Master Thesis (Laurea Magistrale) in Computational Linguistics, defended at an

Italian University between August 1st 2018 and July 31st 2019. This special prize is also endorsed by AILC. We received 6 candidate theses, which have been evaluated by a special jury. The prize will be awarded at the conference by a member of the jury.

Even if CLiC-it is a medium size conference, organizing this annual meeting requires major effort from many people. This conference would not have been possible without the dedication, devotion and hard work of the members of the Local Organising Committee and of the Student Volunteers, who offered their time and energies during the past last year to contribute to the success of the event. We are also extremely grateful to our Program Committee members for producing a lot of detailed and insightful reviews, as well as to the Area Chairs who assisted the Program Chairs in their duties. All these people are named in the following pages.

In addition to the contributions mentioned above, we also gratefully acknowledge the support from endorsing organisations and institutions and from all of our sponsors, who generously provided funds and services that are crucial for the realisation of this event. Special thanks are also due to the University of Bari Aldo Moro for its support in the organisation of the event, as well as to our media partner Start Magazine.

Please join us at CLiC-it 2019 to interact with experts from academia and industry on topics related to Computational Linguistics and Natural Language Processing, and to experience and share new research findings, best practices, state-of-the-art systems and applications. We hope that, as in the past, this year's conference will be intellectually stimulating, and that you will take home many new ideas and methods that will help extend your own research.

Raffaella Bernardi, Roberto Navigli, Giovanni Semeraro
CLiC-it 2019 Conference and Program Chairs

Organizing Committee

Conference and Program Chairs

Raffaella Bernardi, Università degli Studi di Trento

Roberto Navigli, Università degli Studi di Roma La Sapienza

Giovanni Semeraro, Università degli Studi di Bari Aldo Moro

Area Chairs

Dialogue, Discourse and Natural Language Generation

- Alessandro Mazzei, Università degli Studi di Torino
- Marco Guerini, Fondazione Bruno Kessler

Explainability of Deep Learning models for NLP

- Danilo Croce, Università degli Studi di Roma Tor Vergata
- Aurelie Herbelot, Università degli Studi di Trento

Information Extraction, Information Retrieval and Question Answering

- Raffaele Perego, ISTI - CNR
- Stefano Mizzaro, Università degli Studi di Udine

Knowledge Representation

- Enrico Franconi, Libera Università di Bolzano
- Diego Reforgiato, Università degli Studi di Cagliari

Language Resources and Evaluation

- Elisabetta Jezek, Università degli Studi di Pavia
- Cristina Bosco, Università degli Studi di Torino

Lexical and Sentence-level Semantics

- Alessandro Panunzi, Università degli Studi di Firenze

- Rocco Tripodi, Università Ca' Foscari di Venezia

Linguistic Issues in CL and NLP

- Marco Passarotti, Università Cattolica del Sacro Cuore, Milano
- Malvina Nissim, Università di Groningen

Linguistic Theories, Cognitive Modelling and Psycholinguistics

- Marco Marelli, Università degli Studi di Milano-Bicocca
- Francesco Vespignani, Università degli Studi di Trento

Machine Translation and Multilinguality

- Luisa Bentivogli, Fondazione Bruno Kessler
- Johanna Monti, Università degli Studi di Napoli L'Orientale

Morphology and Syntax Processing

- Fabio Tamburini, FICLIT - Alma mater studiorum Università di Bologna
- Cristiano Chesi, Ne.T.S.IUSS Center for Neurolinguistics and Theoretical Syntax, Pavia

NLP for Digital Humanities

- Federico Boschetti, Istituto di Linguistica Computazionale "A. Zampoli" (ILC), CNR di Pisa
- Rachele Sprugnoli, Università Cattolica del Sacro Cuore, Milano

NLP for Web and Social Media

- Serena Villata, CNRS - Laboratoire d'Informatique, Signaux et Systèmes de Sophia-Antipolis
- Viviana Patti, Università degli Studi di Torino

Pragmatics and Creativity

- Federica Cavicchio, Università degli Studi di Padova
- Carlo Strapparava, Fondazione Bruno Kessler

Research and Industrial NLP Applications

- Francesca Bonin, IBM Research AI
- Alessandro Moschitti, Amazon

Replicable and Reproducible methods

- Pierpaolo Basile, Università degli Studi di Bari Aldo Moro
- Giorgio Maria Di Nunzio, Università degli Studi di Padova

Spoken Language Processing and Automatic Speech Understanding

- Francesco Cutugno, Università degli Studi di Napoli Federico II
- Alessandro Vietti, Libera Università di Bolzano

Vision, Robotics, Multimodal and Grounding

- Tatiana Tommasi, Istituto Italiano di Tecnologia, Milano
- Raffaella Folgieri, Università degli Studi di Milano

Local Organisers from the University of Bari Aldo Moro

Pierpaolo Basile

Marco de Gemmis

Andrea Iovine

Pasquale Lops

Cataldo Musto

Fedelucio Narducci

Nicole Novielli

Marco Polignano

Gaetano Rossiello

Lucia Siciliani

Vincenzo Tamburrano

Student Volunteers

Giuseppe Colavito

Paolo Gasparro

Eleonora Ghizzota

Daniela Grassi

Lorenzo Loconte

Program Committee members and Reviewers

Alishahi Afra, Laura Aina, Mehwish Alam, Giambattista Amati, Oscar Araque, Luigi Asprino, Giuseppe Attardi, Mattia Atzeni, Vevake Balaraman, Simone Balloccu, Valentina Bambini, Eduard Barbu, Pierpaolo Basile, Valerio Basile, Roberto Basili, Andrea Bellandi, Luisa Bentivogli, Monica Berti, Marianna Bolognesi, Francesca Bonin, Federico Boschetti, Cristina Bosco, Antonio Branco, Pauli J Brattico, Dominique Brunato, Cristina Burani, Maria Grazia Busa, Davide Buscaldi, Marina Buzzoni, José G. C. de Souza, Elena Cabrio, Basilio Calderone, Charles Callaway, Nicoletta Calzolari, Emanuela Campisi, Lea Canales, Francesco Cangemi, Annalina Caputo, Tommaso Caselli, Giuseppe Castellucci, Federica Cavicchio, Giuseppe Giovanni Antonio Celano, Fabio Celli, Loredana Cerrato, Mauro Cettolo, Cristiano Chesi, Francesca Chiusaroli, Grzegorz Chrupała, Andrea Cimino, Michael Cochez, Giovanni Colavizza, Simone Conia, Sergio Consoli, Anna Corazza, Gianpaolo Coro, Piero Cosi, Gregory Crane, Alice Cravotta, Davide Crepaldi, Fabio Crestani, Danilo Croce, Francesco Cutugno, Francesca D’Errico, Giovanni Da San Martino, Rossana Damiano, Marco de Gemmis, Daniele De Massari, Thierry Declerck, Dario Del Fante, Angelo Mario Del Grosso, Marco Del Tredici, Felice Dell’Orletta, Claudio Delli Bovi, Danilo Dessi, Barbara Di Eugenio, Mattia Antonino Di Gangi, Maria Di Maro, Giorgio Maria Di Nunzio, Mauro Dragoni, Maud Ehrmann, Andrea Esuli, Kilian Evang, Stefano Faralli, Dimeji Farri, Anna Feltracco, Marcello Ferro, Nicola Ferro, Simone Filice, Antske Fokkens, Raffaella Folgieri, Enrico Franconi, Diego Frassinelli, Francesca Frontini, Aldo Gangemi, Albert Gatt, Lorenzo Gatti, Emiliano Giovannetti, Alessandro Giuliani, Marco Guerini, Christian Hardmeier, Sadiq Hasan, Dag Haug, Rim Helaoui, Monique Hendriks, Aurelie Herbelot, Amac Herdagdelen, Delia Irazu Hernandez Farias, Ignacio Iacobacci, Carlos A. Iglesias, Diana Inkpen, Elisabetta Jezek, Charles Jochim, Alina Karakanta, Ana Kostadinovska, Adamantios Koumpis, E J Krahmer, Sandra Kuebler, Alexander Kuhnle, Jacek Kustra, Surafel Melaku Lakew, Alberto Lavelli, Gianluca Lebani, Alessandro Lenci, Eleonora Litta, Giorgia Lodi, Samuel Louvan, Claudio Lucchese, Marco Maggini, Simone Magnolini, Paolo Mairano, Maria Maistro, Francesco Mambrini, Alice Marascu, Diego Marcheggiani, Marco Marelli, Mirko Marras, Claudia Marzi, Alessandro Mazzei, Massimo Melucci, Stefano Menini, V. Menkovski, Stefano Mizzaro, Massimo Moneglia, Johanna Monti, Alessandro Moschitti, Claudio Mulatti, Maria Teresa Musacchio, Cataldo Musto, Federico Nanni, Franco

Maria Nardini, Fedelucio Narducci, Costanza Navarretta, Vincent Ng, Massimo Nicosia, Malvina Nissim, Nicole Novielli, Andrea Nuzzolese, Antonio Origlia, Salvatore Orlando, Francesco Osborne, Petya Osenova, Alessio Palmero Aprosio, Ludovica Pannito, Alessandro Panunzi, Patrick Paroubek, Tommaso Pasini, Lucia Passaro, Marco Passarotti, Viviana Patti, Steffen Pauws, Raffaele Perego, Diego Pescarini, Sandro Pezzelle, Paola Pietrandrea, Vito Pirrelli, Massimo Poesio, Isabella Poggi, Marco Polignano, Edoardo Maria Ponti, Marten Postma, Valentina Presutti, Valeria Quochi, Daniele Radicioni, Alessandro Raganato, Diego Reforgiato, Corentin Ribeyre, Daniele Riboni, Bruce Robertson, Matteo Romanello, Salvatore Romeo, Francesco Ronzano, Paolo Rosso, Agata Rotondi, Alessandro Russo, Irene Russo, Bogdan Sacaleanu, Harald Sack, Manuela Sanguinetti, Marco S. G. Senaldi, Lucia Siciliani, Fabrizio Silvestri, Maria Simi, Luca Soldaini, Claudia Soria, Rachele Sprugnoli, Jacopo Staiano, Ieva Staliunaite, R. Stephens, Carlo Strapparava, Francesca Strik Lievers, Simone Sulpizio, Fabio Tamburini, Amirhossein Tebbifakhr, Maurizio Tesconi, Tatiana Tommasi, Sara Tonelli, Rocco Tripodi, Enrica Troiano, Marco Turchi, Antonio Uva, Dieter Van de Craen, Marieke van Erp, Rossella Varvara, Giulia Venturi, Francesco Vespignani, Federica Vezzani, Alessandro Vietti, Laure Vieu, Serena Villata, Marco Viviani, Pieter Vos, Ivan Vulić, Tobias Wirth, Charalampos Xanthopoulos, Fabio Massimo Zanzotto and Enrico Zovato.

CLiC-it 2019 is endorsed by



Sponsors

Gold



Silver



Bronze



Media Partner



Visually-Grounded Dialogue Models: Past, Present, and Future

Raquel Fernández
University of Amsterdam, The Netherlands
`raquel.fernandez@uva.nl`

Abstract

The past few years have seen an increasing interest in developing neural-network-based agents for visually-grounded dialogue, where the conversation participants communicate about visual content. I will start by discussing how visual grounding can be integrated with traditional task-oriented dialogue system components. Most current work in the field focuses on reporting numeric results solely based on task success. I will argue that we can gain more insight by (i) analysing the linguistic output of alternative systems and (ii) probing the representations they learn. I will also introduce a new dialogue dataset we have developed using a data-collection setup designed to investigate linguistic common ground as it accumulates during visually-grounded interaction.

Impossible Languages and the Architecture of Human Grammars

Andrea Moro

Scuola Universitaria Superiore IUSS Pavia, Italia

`andrea.moro@iusspavia.it`

Abstract

Every human language meets a set of formal principles such as recursion. Are the boundaries of Babel cultural, conventional, accidental or neurobiological? By testing the brains network activations to the acquisition of artificial “impossible languages” with neuroimaging techniques it has been possible to provide strong evidence in favor of a neurobiological explanation. Along with network activations, the first experiments at deciphering the neuronal electrophysiological code underlying language are illustrated, in particular those recording the “sound of thoughts” in inner speech.

Recognizing and Reducing Bias in NLP Applications

Dirk Hovy Università Bocconi, Italia
`dirk.hovy@unibocconi.it`

Abstract

As NLP technology becomes used in ever more settings, it has ever more impact on the lives of people all around the world. As NLP practitioners, we have become increasingly aware that we have the responsibility to evaluate the effects of our research and prevent or at least mitigate harmful outcomes. This is true for academic researchers, government labs, and industry developers. However, without experience of how to recognize and engage with the many ethical conundrums in NLP, it is easy to become overwhelmed and remain inactive. One of the most central ethical issues in NLP is the impact of hidden biases that affect performance unevenly, and thereby disadvantage certain user groups.

This tutorial aims to empower NLP practitioners with the tools spot these biases, and a number of other common ethical pitfalls of our practice. We will cover both high-level strategies, as well as go through specific case sample exercises. This is a highly interactive workshop with room for debate and questions from the attendees. The workshop will cover the following broad topics:

- Biases: Understanding the different ways in which biases affect NLP data, models, and input representations, including including strategies to test for and reduce bias in all of them.
- Dual Use: Learning to anticipate how a system could be repurposed for harmful or negative purposes, rather than its intended goal.
- Privacy: Protecting the privacy of users both in corpus construction and model building.

Prerequisite or Not Prerequisite? That's the Problem!

An NLP-based Approach for Concept Prerequisites Learning

Chiara Alzetta^{*◇}, Alessio Miaschi^{*◇}, Giovanni Adorni^{*}, Felice Dell'Orletta[◇],
Frosina Koceva^{*}, Samuele Passalacqua^{*}, Ilaria Torre^{*}

^{*}DIBRIS, Università degli Studi di Genova, ^{*}Dipartimento di Informatica, Università di Pisa,

[◇]Istituto di Linguistica Computazionale "Antonio Zampolli", Pisa - ItaliaNLP Lab

{chiara.alzetta, frosina.koceva}@edu.unige.it, alessio.miaschi@phd.unipi.it,
samuele.passalacqua@dibris.unige.it, {ilaria.torre, adorni}@unige.it,
felice.dellorletta@ilc.cnr.it

Abstract

English. This paper presents a method for prerequisite learning classification between educational concepts. The proposed system was developed by adapting a classification algorithm designed for sequencing Learning Objects to the task of ordering concepts from a computer science textbook. In order to apply the system to the new task, for each concept we automatically created a learning unit from the textbook using two criteria based on concept occurrences and burst intervals. Results are promising and suggest that further improvements could highly benefit the results.¹

Italiano. Il presente articolo descrive una strategia per l'identificazione di prerequisiti fra concetti didattici. Il sistema proposto è stato realizzato adattando un algoritmo per ordinamento di Learning Objects al compito di ordinamento di concetti estratti da un libro di testo di informatica. Per adeguare il sistema al nuovo scenario, per ogni concetto stata automaticamente creata una unità di apprendimento a partire dal libro di testo selezionando i contenuti sulla base di due differenti criteri: basandosi sull'occorrenza del concetto e sugli intervalli di burst. I risultati sono promettenti e lasciano intuire la possibilità di ulteriori miglioramenti.

1 Introduction

Personalised learning paths creation is an active research topic in the field of education (Chen,

2009; Kurilovas et al., 2015; Almasri et al., 2019). The most fundamental issue behind this task is the need to understand how educational concepts are pedagogically related to each other: what information one has to study/know first in order to understand a given topic. In this paper we focus on such relations, i.e. *prerequisite relations*, between educational concepts of a textbook in English and we present a method for their automatic identification. Here, we define *concepts* all the relevant topics extracted from the textbook and we represent them as single or multi word terms.

Automatic prerequisite extraction is a task deeply rooted in the field of education, whose results can be easily integrated in many different contexts, such as curriculum planning (Agrawal et al., 2016), course sequencing (Vuong et al., 2011), reading list generation (Gordon et al., 2017), automatic assessment (Wang and Liu, 2016), domain ontology construction (Zouaq et al., 2007; Larranaga et al., 2014) and automatic educational content creation (Lu et al., 2019). Several methods have been devised to extract prerequisite relations (Liang et al., 2015; Pan et al., 2017a; Liang et al., 2018b), however they were mainly focused on educational materials already enriched with some sort of explicit relations, such as Wikipedia pages, course materials or learning objects (LOs). More challenging is identifying prerequisites when no such relations are given and textual content is the only available resource.

In 2019, we proposed two methods to identify prerequisite relations between concepts without using external knowledge or even pre-defined relations. The former method (Adorni et al., 2019) is based on burst analysis and temporal reasoning on concepts occurrence, while the latter (Miaschi et al., 2019) uses deep learning for learning object ordering. Both these methods extract prerequisite relations from textual educational materials without using any form of structured information.

¹Copyright © 2019 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

In this work, we adapt the system for learning object ordering described in Miaschi et al. (2019) to the task of sequencing concepts in a textbook according to their prerequisite relations. For training and testing our system we relied on a new version of PRET (Alzetta et al., 2018), a gold dataset manually annotated with prerequisite relations between educational concepts. Moreover, since the classifier was designed to acquire learning objects as input, we automatically created a learning unit² for each concept according to two different criteria: (i) considering all sentences showing an occurrence of the concept, (ii) considering burst intervals (Kleinberg, 2003) of each concept extracted according to the strategy of Adorni et al. (2019).

The remainder of the paper is organised as follows. First, we present related work (Section 2) and the dataset used for the experiments (Section 3). Section 4.1 presents the classifier, while Burst analysis is described in Section 4.2 and the experimental settings in Section 4.3. Results and discussion are reported in Section 4.4, while error analysis is illustrated in Section 5. Section 6 concludes the paper.

Our Contribution. In this paper: (i) we use a deep learning-based approach for prerequisite relation extraction between educational concepts of a textbook; (ii) we test the impact of creating learning units for each concept according to different criteria and without relying on any explicit structured information, such as Wikipedia hyperlinks; (iii) we show the effectiveness of our approach on real educational materials.

2 Related Work

Datasets annotated with prerequisite relations are built mainly considering two types of data: course materials, acquired from MOOCs (Chaplot et al., 2016; Pan et al., 2017a; Pan et al., 2017b; Gasparetti et al., 2018; Roy et al., 2018) or university websites (Liang et al., 2017; Li et al., 2019), and educational materials in a broader sense, such as scientific databases (Gordon et al., 2017), learning objects (Talukdar and Cohen, 2012; Gasparetti et al., 2018) and textbooks (Wang et al., 2016). The most common approach for prerequisite annotation is to ask experts to evaluate all possible

pairs generated from the combination of selected concepts (Chaplot et al., 2016; Wang et al., 2016; Li et al., 2019) or a random sample of that set (Pan et al., 2017b; Gordon et al., 2017; Gasparetti et al., 2018). The dataset presented by Wang et al. (2016) is the one we consider most closely related to ours, since it shows prerequisite relations between relevant concepts extracted from a textbook. However, in their dataset a matching with a Wikipedia page was a strict requirement for concept selection. Contrary to previous works, we asked experts to build the concept pairs if a prerequisite relation was observed while reading a textbook, regardless the existence of a corresponding Wikipedia page for the concepts. Hence we allowed for more subjectivity, without restricting experts’ evaluation to a predefined list of items.

For what concerns prerequisite learning approaches, initial work in this field relied on graph analysis (Vassileva, 1997; Brusilovsky and Vassileva, 2002) or, more recently, on link-based metrics inferred from the Wikipedia graph of hyperlinks between pages (Liang et al., 2015). Talukdar and Cohen (2012) made the first attempt to apply machine learning techniques to prerequisite prediction: hyperlinks, hierarchical category structure and edits of Wikipedia pages are the features of a MaxEnt classifier. Similarly, Gasparetti et al. (2018) use Wikipedia hierarchical category structure and hyperlinks. Similarly to our approach, (Liang et al., 2018a; Liang et al., 2018b) integrated text-based features for prerequisite learning, but reported graph-based features as more informative.

Contrary to the above methods, we assign a higher informative value to the textual content referring to a concept and we use this only to extract the features for the classifier. Moreover, we combine the classifier with the burst algorithm (Kleinberg, 2003), which selects the most relevant textual content related to a concept from the textual material. This choice makes our method suitable for prerequisite learning on educational contents also when structured graph information is not available.

3 Dataset

For our experiments we relied on a novel version of PRET dataset (Alzetta et al., 2018), PRET 2.0, a dataset manually annotated with prerequisite relation between educational concepts extracted from

²Learning unit is meant here as learning content, with no reference to units of learning in curricula and tables of content.

a chapter of a computer science textbook written in English (Brookshear and Brylow, 2015).

In this novel version, five experts were asked to re-annotate the same text indicating any prerequisite concept of each relevant term appearing in the text. The set of relevant terms was extracted with the same automatic strategy described in Alzetta et al. (2018), but this time the list was manually validated by three experts in order to identify a commonly agreed set of concepts, which resulted in a terminology of 132 concepts. Besides these terms, each expert could independently add new concepts to the terminology when annotating the text if he/she regards them as relevant. Consequently, experts produced different sets of concept pairs annotated with prerequisite relations since 221 new concepts were manually added during the annotation process.

The final gold dataset results from the combination of all annotations, thus considering as positive pairs (i.e. showing a prerequisite relation) all pairs of concepts annotated by at least one expert. The manual annotation resulted in 25 pairs annotated by all five experts, 46 annotated by four experts, 83 by three, 214 by two and 698 by only one annotator, for a total of 1,066 pairs.

2,349 transitive pairs were also automatically generated and added to the dataset: if a prerequisite relation exists between concepts A and B and between concepts B and C, we add a positive relation between A and C to increase the coherence of annotation. In order to obtain a balanced dataset for training our deep learning system, negative pairs were automatically created by randomly pairing concepts and adding them as negative examples if they were missing in the dataset. Overall, the final dataset consists of 353 concepts and 6,768 relations.

4 Method and Experiments

In this Section we present our approach for learning prerequisites between educational concepts. We trained and tested the same deep learning model on three datasets generated from PRET 2.0 that vary with respect to the criterion used for retrieving textual content of each concept in the dataset. As a result, we were able to study performance variations of the classifier given different input data.

Task. We tackle the problem of concept prerequisite learning as a task of automatic binary classi-

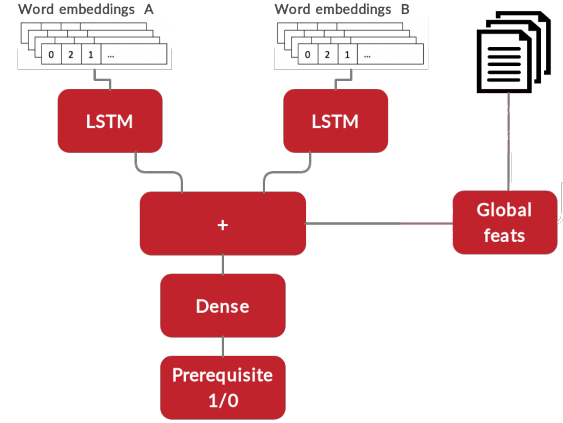


Figure 1: Method workflow.

fication of concept pairs: given a pair of concepts (A,B), we predict whether or not concept B is a prerequisite of concept A.

4.1 Classifier

The system used to predict whether or not two concepts show a prerequisite relation is the deep learning architecture described in Miaschi et al. (2019). Specifically, we relied on the model which uses pre-trained word embeddings (WE) and global features automatically extracted from the dataset.

The system architecture (see Figure 1) is composed of two LSTM-based sub-networks with 64 units, whose outputs are concatenated and joined with a set of global features. The input of the two LSTM-based sub-networks corresponds to the pre-trained WE of concept A and B respectively. The output layer consists of a single Dense unit with sigmoid activation function. The pre-trained WE were computed using an English lexicon of 128 dimensions built using the ukWac corpus (Baroni et al., 2009). Global features were devised to extract linguistic information from learning units of both concepts in a pair, such as mentions to the other concept of the pair or the Jaccard similarity between textual contents of the two learning units.

For the complete list of global features, refer to Miaschi et al. (2019).

4.2 Burst Analysis

Burst analysis is based on the assumption that a phenomenon might become particularly relevant in a certain period along a time series, most likely because its occurrence rises above a certain thresh-

old. Such periods of increased activity of the phenomenon are called "burst intervals" and can be modelled by means of a two state automaton in which the phenomenon is in the first state if it has a low occurrence, but then it moves to the second state if its occurrence rises above a certain threshold, and eventually it goes back to the first state if its occurrence goes below the threshold (Kleinberg, 2003).

Given its nature, this kind of analysis is highly employed for detecting events from data streams (Fung et al., 2005; Takahashi et al., 2012; Kleinberg, 2016). When applied to textual data – e.g., for text clustering (He et al., 2007), summarization (Subasic and Berendt, 2010) or relation extraction (Yoon et al., 2014; Lee et al., 2015) – the linear progression of the text acts as the time series, hence burst intervals correspond to sequences of sentences where a given term is particularly relevant. In Adorni et al. (2019) burst analysis was used to detect the bursting intervals of concepts along a textbook chapter: for each term, the burst algorithm identified a unique or multiple burst intervals of various length (i.e. a different number of sentences involved in each interval). Temporal reasoning (Allen, 1983) was then employed to find prerequisite relations between concepts.

In this work we use the burst intervals retrieved as described in Adorni et al. (2019) to select relevant content of the textbook for each concept. Our intuition is that burst intervals should capture the most informative portions of text for each concept from the entire textbook content. Note that for this experiment we only used the bursts detected with the first phase of the algorithm described in (Adorni et al., 2019), i.e. the temporal reasoning is not employed here.

4.3 Experimental Settings

Since our deep learning model was designed to find prerequisite relations between learning objects, we had to adapt our classification algorithm to the task we deal with in this work, namely ordering concepts from a textbook. To this aim, we created learning units for each concept of PRET 2.0 dataset and we used them as input for the classifier.

In order to verify the impact of different input data, we tested different strategies for the creation of learning units. Hence, content related to each concept was retrieved according to two different

Model	Emb. Dim.	F-Score	Accuracy
Occurrence	5	73.75	69.65
	10	74.79	70.36
	15	73.7	69.19
	30	73.11	67.97
	avg	73.84	69.30
Burst Intervals	5	71.75	65.54
	10	73.91	69.49
	15	72.97	67.77
	30	71.37	65.06
	avg	72.5	66.96
Most Relevant Burst Interval	5	73.06	67.8
	10	72.04	66.52
	15	71.58	64.43
	30	71.49	64.48
	avg	72.04	65.80
Baseline		66.66	50

Table 1: Classification F-Score and Accuracy values for the three models with varying number of sentences considered for lexical features. Average and baseline values are also reported.

criteria: (1) considering all sentences where a certain concept occurs (Occurrence Model); (2) considering burst intervals for each concept. The latter is further divided into two cases depending on the appearing order of burst intervals: (i) burst intervals reflect their linear order along the text (Burst Intervals Model); (ii) burst intervals are re-ordered, having the most relevant burst interval as first (Most Relevant Burst Interval Model). The most relevant burst interval is defined as the first burst interval that exceeds the average length of all the bursts of that concept (Adorni et al., 2019; Passalacqua et al., 2019).

The resulting datasets show different learning unit dimensions: Burst Intervals models produce learning units with an average length of 534 tokens, while those considered for the Occurrence Model are smaller, with 250 tokens on average. While global features consider the entire content of the learning unit, for all models WE are computed only for the first n sentences. We tried different length of n : 5, 10, 15 and 30.

Results in terms of F-Score and accuracy were compared against a Zero Rule algorithm baseline.

4.4 Experiments Results and Discussion

Results reported in Table 1 show satisfying performances of our system that outperforms the baseline in all configurations. Best results are obtained by the Occurrence Model using 10 sentences to compute lexical features. In general, computing the WE on 10 sentences or less allows to obtain

better performances in all settings. This could be due to the fact that the definition of a concept and its contextualisation with respect to other concepts are generally discussed by the author of the book when the concept is first mentioned in the text. Thus, sentences containing the first occurrences of the term seem to be the most informative for this task. To assess this hypothesis, we manually inspected sentences containing the first mention of each concept. The analysis revealed that 36.3% of the observed sentences contained a concept definition, thus supporting our intuition that the first mention is relevant for concept contextualisation.

The results obtained using the Burst Interval Model are slightly worse, although comparable, probably because, since burst intervals do not necessarily capture all the occurrences of a concept, in some cases the first mentions could be missing from the learning unit. The lowest scores are predictably those obtained using the Most Relevant Burst Interval Model: changing the order of the sentences penalises the system since the temporal order often plays an important role when a prerequisite relation is established between two concepts. Several algorithms exploit a time-based strategy for prerequisite extraction relying on the temporal nature of this relation (Sosnovsky et al., 2004; Adorni et al., 2018) and the analysis of human annotations suggests that the direction of this relation (i.e. A is prerequisite of B or vice-versa) tends to be highly correlated with the temporal order of the two concepts (Passalacqua et al., 2019). Besides, the most relevant burst is not necessarily the first burst interval for that concept and, for this reason, it could contain less relevant information about the concept and its prerequisites. Interestingly, the best results for this model are obtained considering only 5 sentences for computing WE, probably because the system has less chance of observing a lexicon related to other concepts.

If we look at the variation of accuracy values with respect to the classifier confidence (see Figure 2), we observe that our system shows an expected behaviour. In fact, at high confidences correspond high accuracy scores, while at confidence around .5 (12.66% of dataset pairs) we notice that the classifier is more unsure of its decision, obtaining results below the baseline. It should be noted also that the majority of concept pairs (25%) have been classified with a confidence value around .6, while the pairs obtaining the highest confidence

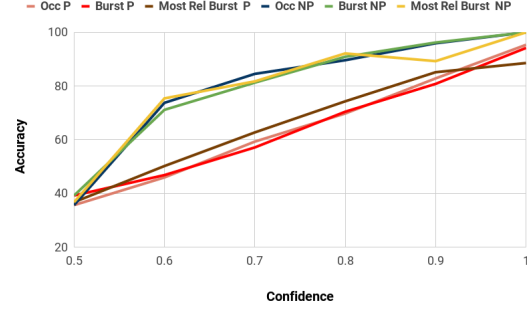


Figure 2: Variation of accuracy values wrt the classifier confidence for pairs labelled as prerequisite (*P*) and non prerequisite (*NP*) in all models considering 10 sentences to compute lexical features.

value (i.e. equal to 1) are only 1.21%.

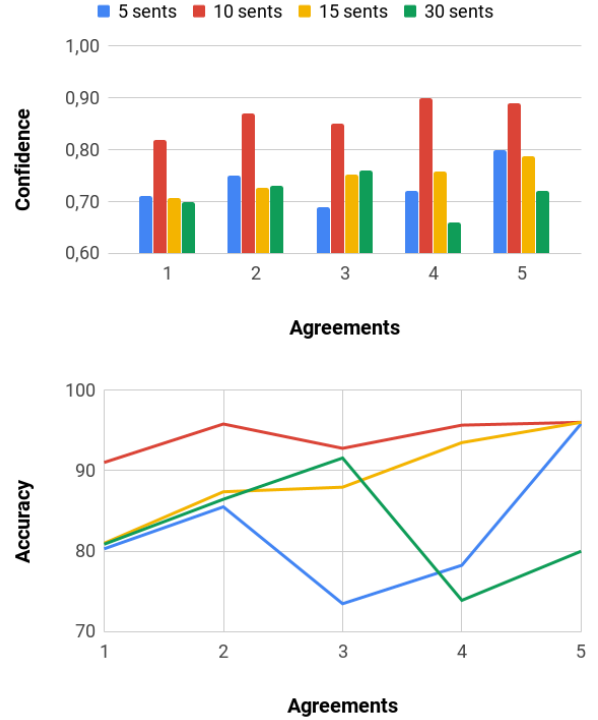


Figure 3: Variation of confidence (on top) and accuracy (on bottom) wrt the agreement value for the Occurrence Model (all possible embeddings length are considered).

The graphs in Figure 3 show the variation of confidence and accuracy values with respect to the annotators agreement. We report results only for the Occurrence Model since it is the one that obtained the best scores during classification. As we can see, the concept pairs for which all the annota-

tors agree on tend to obtain higher confidence and, consequently, the classifier shows the best performances. The only exception is the model that computes WE using the first 30 sentences, which obtains instead the best scores on the pairs annotated by only 3 experts. The reason for this behaviour will be explored in future work.

5 Error Analysis

This Section compares the results obtained by the three models (i.e. Occurrence, Burst Interval and Most Relevant Burst Interval) when considering 10 sentences for computing WE.

The overall number of pairs assigned with a wrong label by the classifier is quite similar across each setting: 1,835 pairs for the Occurrences model, 1,923 for the Burst Interval model and 2,089 for the Most Relevant Burst model. Moreover, we observe that among these pairs more than 80% were classified as “prerequisite”, suggesting that the system overestimates the prerequisite relation, assigning the label also to non-prerequisite pairs.

Focusing the analysis on relations that are annotated as prerequisites in the dataset, we observe how their prediction varies across models. 126 pairs were assigned with a wrong “non-prerequisite” label by all models showing similar average confidence values: 0.66, 0.66 and 0.62 for Occurrences, Burst and Most Relevant Burst model respectively. This result suggests that these pairs are particularly complex to classify. Conducting a deeper analysis on this subset, we notice that 85.71% (108) of the pairs are transitive pairs automatically generated (see Section 3). Such type of relations seems thus harder to classify than manually annotated ones and might require a different set of features to be recognised considering also that they represent more distant relations. Furthermore, consider that the remaining 18 pairs (14.28%) are manually annotated relations with low agreement values: 15, 2 and 1 were annotated by one, two and three annotators respectively.

6 Conclusion

In this paper we tested a deep learning model for prerequisite relation extraction in a real educational environment, using a dataset (PRET 2.0) built starting from a computer science textbook. The results demonstrated the effectiveness of our system, suggesting that it is possible to infer pre-

requisite relation out of textual educational material without using any form of structured information. Nevertheless, further work needs to be done, particularly for improving the performances of our system in a out-of-domain scenario, namely using concept pairs of a different domain during testing. Moreover, it could be useful to investigate the use of transitive relations and to study more accurately their impact on the system’s performance. In addition, in order to identify prerequisite relationships while taking into account different types of relations (e.g. transitive ones) it could be interesting to frame our task as a ranking or multi-classification task rather than a binary classification one. Further analysis is also required to investigate the effect of using different numbers of sentences for creating WE. We plan also to explore the impact of using temporal reasoning on concept pairs (Adorni et al., 2019), which has not been considered in this work.

References

- Giovanni Adorni, Felice Dell’Orletta, Frosina Kocova, Ilaria Torre, and Giulia Venturi. 2018. Extracting dependency relations from digital learning content. In *Italian Research Conference on Digital Libraries*, pages 114–119. Springer.
- Giovanni Adorni, Chiara Alzetta, Frosina Kocova, Samuele Passalacqua, and Ilaria Torre. 2019. Towards the identification of propaedeutical relations in textbooks. In *International Conference on Artificial Intelligence in Education*, pages 1–13. Springer.
- Rakesh Agrawal, Behzad Golshan, and Evangelos Papalexakis. 2016. Toward data-driven design of educational courses: A feasibility study. *Journal of Educational Data Mining*, 8(1):1–21.
- James F Allen. 1983. Maintaining knowledge about temporal intervals. *Communications of the ACM*, 26(11).
- Abdelbaset Almasri, Adel Ahmed, Naser Almasri, Yousef S Abu Sultan, and Ahmed Y Mahmoud. 2019. Intelligent tutoring systems survey for the period 2000-2018. *IJARW, International Journal of Academic Engineering Research (IJAER)*, 3 (5):21-37.
- Chiara Alzetta, Frosina Kocova, Samuele Passalacqua, Ilaria Torre, and Giovanni Adorni. 2018. Pret: Prerequisite-enriched terminology. a case study on educational texts. In *Proceedings of the Fifth Italian Conference on Computational Linguistics CLiC-it 2018*.
- Marco Baroni, Silvia Bernardini, Adriano Ferraresi, and Eros Zanchetta. 2009. The wacky wide

- web: a collection of very large linguistically processed web-crawled corpora. *Language resources and evaluation*, 43(3):209–226.
- Glenn Brookshear and Dennis Brylow. 2015. *Computer Science: An Overview, Global Edition*, chapter 4 Networking and the Internet. Pearson Education Limited.
- Peter Brusilovsky and Julita Vassileva. 2002. Course sequencing techniques for large-scale web-based education. *Int. Journal of Continuing Engineering Education and Life-long Learning*.
- Devendra Singh Chaplot, Yiming Yang, Jaime G Carbonell, and Kenneth R Koedinger. 2016. Data-driven automated induction of prerequisite structure graphs. In *EDM*, pages 318–323.
- Chih-Ming Chen. 2009. Ontology-based concept map for planning a personalised learning path. *British Journal of Educational Technology*, 40(6):1028–1058.
- Gabriel Pui Cheong Fung, Jeffrey Xu Yu, Philip S Yu, and Hongjun Lu. 2005. Parameter free bursty events detection in text streams. In *Proceedings of the 31st international conference on Very large data bases*, pages 181–192.
- Fabio Gaspiretti, Carlo De Medio, Carla Limongelli, Filippo Sciarrone, and Marco Temperini. 2018. Prerequisites between learning objects: Automatic extraction based on a machine learning approach. *Telematics and Informatics*, 35(3):595–610.
- Jonathan Gordon, Stephen Aguilar, Emily Sheng, and Gully Burns. 2017. Structured generation of technical reading lists. In *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 261–270.
- Qi He, Kuiyu Chang, Ee-Peng Lim, and Jun Zhang. 2007. Bursty feature representation for clustering text streams. In *Proceedings of the 2007 SIAM International Conference on Data Mining*, pages 491–496.
- Jon Kleinberg. 2003. Bursty and hierarchical structure in streams. *Data Mining and Knowledge Discovery*, 7(4):373–397.
- Jon Kleinberg. 2016. Temporal dynamics of on-line information streams. In *Data Stream Management*, pages 221–238. Springer.
- Eugenijus Kurilovas, Inga Zilinskiene, and Valentina Dagiene. 2015. Recommending suitable learning paths according to learners preferences: Experimental research results. *Computers in Human Behavior*, 51:945–951.
- Mikel Larranaga, Angel Conde, Inaki Calvo, Jon A Elorriaga, and Ana Arruarte. 2014. Automatic generation of the domain module from electronic textbooks: method and validation. *IEEE transactions on knowledge and data engineering*, 26(1):69–82.
- Seulki Lee, Youkyoung Park, and Wan C Yoon. 2015. Burst analysis for automatic concept map creation with a single document. *Expert Systems with Applications*, 42(22):8817–8829.
- Irene Li, Alexander R Fabbri, Robert R Tung, and Dragomir R Radev. 2019. What should i learn first: Introducing lecturebank for nlp education and prerequisite chain learning. *Proceedings of AAAI 2019*.
- Chen Liang, Zhaohui Wu, Wenyi Huang, and C Lee Giles. 2015. Measuring prerequisite relations among concepts. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1668–1674.
- Chen Liang, Jianbo Ye, Zhaohui Wu, Bart Pursel, and C Lee Giles. 2017. Recovering concept prerequisite relations from university course dependencies. In *AAAI*, pages 4786–4791.
- Chen Liang, Jianbo Ye, Shuting Wang, Bart Pursel, and C Lee Giles. 2018a. Investigating active learning for concept prerequisite learning. *Proc. EAAI*.
- Chen Liang, Jianbo Ye, Han Zhao, Bart Pursel, and C Lee Giles. 2018b. Active learning of strict partial orders: A case study on concept prerequisite relations. *arXiv preprint arXiv:1801.06481*.
- Weiming Lu, Pengkun Ma, Jiale Yu, Yangfan Zhou, and Baogang Wei. 2019. Metro maps for efficient knowledge learning by summarizing massive electronic textbooks. *International Journal on Document Analysis and Recognition (IJ DAR)*, pages 1–13.
- Alessio Miaschi, Chiara Alzetta, Franco Alberto Cardillo, and Felice Dell’Orletta. 2019. Linguistically-driven strategy for concept prerequisites learning on italian. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 285–295.
- Liangming Pan, Chengjiang Li, Juanzi Li, and Jie Tang. 2017a. Prerequisite relation learning for concepts in moocs. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1447–1456.
- Liangming Pan, Xiaochen Wang, Chengjiang Li, Juanzi Li, and Jie Tang. 2017b. Course concept extraction in moocs via embedding-based graph propagation. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 875–884.
- Samuele Passalacqua, Frosina Koceva, Chiara Alzetta, Ilaria Torre, and Giovanni Adorni. 2019. Visualisation analysis for exploring prerequisite relations in textbooks. *First Workshop on Intelligent Textbooks*.
- Sudeshna Roy, Meghana Madhyastha, Sheril Lawrence, and Vaibhav Rajan. 2018. Inferring concept prerequisite relations from online educational resources. *31st AAAI Conference on*

- Sergey Sosnovsky, Peter Brusilovsky, and Michael Yudelson. 2004. Supporting adaptive hypermedia authors with automated content indexing. In *Proceedings of Second International Workshop on Authoring of Adaptive and Adaptable Educational Hypermedia at the Third International Conference on Adaptive Hypermedia and Adaptive Web-Based Systems (AH'2004)*, Eindhoven, the Netherlands.
- Ilija Subasic and Bettina Berendt. 2010. From bursty patterns to bursty facts: The effectiveness of temporal text mining for news. In *Proceedings of ECAI 2010: 19th European Conference on Artificial Intelligence*, pages 517–522.
- Yusuke Takahashi, Takehito Utsuro, Masaharu Yoshioka, Noriko Kando, Tomohiro Fukuhara, Hiroshi Nakagawa, and Yoji Kiyota. 2012. Applying a burst model to detect bursty topics in a topic model. In *International Conference on NLP*, pages 239–249. Springer.
- Partha Pratim Talukdar and William W Cohen. 2012. Crowdsourced comprehension: predicting prerequisite structure in wikipedia. In *Proceedings of the Seventh Workshop on Building Educational Applications Using NLP*, pages 307–315. Association for Computational Linguistics.
- Julita Vassileva. 1997. Dynamic course generation. *Journal of computing and information technology*, 5(2):87–102.
- Annalies Vuong, Tristan Nixon, and Brendon Towle. 2011. A method for finding prerequisites within a curriculum. In *EDM*, pages 211–216.
- Shuting Wang and Lei Liu. 2016. Prerequisite concept maps extraction for automatic assessment. In *Proceedings of the 25th International Conference Companion on World Wide Web*, pages 519–521. International World Wide Web Conferences Steering Committee.
- Shuting Wang, Alexander Ororbia, Zhaohui Wu, Kyle Williams, Chen Liang, Bart Pursel, and C Lee Giles. 2016. Using prerequisites to extract concept maps from textbooks. In *Proceedings of the 25th acm international on conference on information and knowledge management*, pages 317–326. ACM.
- Wan C Yoon, Sunhee Lee, and Seulki Lee. 2014. Burst analysis of text document for automatic concept map creation. In *International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems*, pages 407–416. Springer.
- Amal Zouaq, Roger Nkambou, and Claude Frasson. 2007. An integrated approach for automatic aggregation of learning knowledge objects. *Interdisciplinary Journal of E-Learning and Learning Objects*, 3(1):135–162.

An Empirical Analysis of Linguistic, Typographic, and Structural Features in Simplified German Texts

Alessia Battisti

Sarah Ebling

Martin Volk

Institute of Computational Linguistics, University of Zurich

Andreasstrasse 15, 8050 Zurich, Switzerland

alessia.battisti@uzh.ch, {ebling|volk}@cl.uzh.ch

Abstract

English. We investigate a newly compiled corpus of simplified German texts for evidence of multiple complexity levels using unsupervised machine learning techniques. We apply linguistic features used in previous supervised machine learning research and additionally exploit structural and typographic characteristics of simplified texts. The results show a difference in complexity among the texts investigated, with optimal partitioning solutions ranging between two and four clusters. They demonstrate that both linguistic and structural/typographic features are constitutive of the clusters.

Italiano. *Esaminiamo un nuovo corpus di testi in tedesco semplificato per cercare delle evidenze relative a molteplici livelli di complessità utilizzando tecniche di apprendimento automatico non supervisionato. Appliciamo variabili linguistiche utilizzate in precedenti ricerche con apprendimento automatico supervisionato e sfruttiamo inoltre le caratteristiche strutturali e tipografiche dei testi semplificati. I risultati mostrano una differenza di complessità tra i testi analizzati, con suddivisioni ottimali variabili da due a quattro cluster. Ciò dimostra che sia le caratteristiche linguistiche sia quelle strutturali/tipografiche sono costitutive dei cluster.*

abilities. This group includes persons with cognitive impairment and learning disabilities, prelingually deaf persons, functionally illiterate persons, and foreign language learners (Bredel and Maaß, 2016). Simplified language is characterised by reduced lexical and syntactic complexity and includes images, structured layout, and explanations of difficult words. For simplified German, several guidelines exist that define which structures need to be avoided, which need to be paraphrased, and which are comprehensible (Bundesministerium für Arbeit und Soziales, 2011; Inclusion Europe, 2009; Maaß, 2015; Netzwerk Leichte Sprache, 2013).

Various countries have acknowledged simplified language as a means of inclusion that enables the target populations mentioned above to inform themselves of their legal rights and participate in society. German-speaking countries have been promoting simplified language only in the last years, in particular since the ratification of the United Nations Convention on the Rights of Persons with Disabilities (United Nations, 2006) in Austria (2008), Germany (2009), and Switzerland (2014). As a result, large amounts of texts in simplified German have become available.

More recently, simplified German has been conceptualised as a construct with multiple complexity levels (Bock, 2014; Bredel and Maaß, 2016; Kellermann, 2014). However, these proposals are merely theoretical: They are not yet operationalised, i.e., no sets of guidelines exist that distinguish the proposed levels with reference to linguistic or other features. The social franchise network *capito*,¹ a provider of simplification services as well as training courses for simplified language translators, recognises three levels of simplified German corresponding to the Common European Framework of Reference for Language (CEFR)

1 Introduction

Simplified language aims at providing comprehensible information to persons with reduced reading

Copyright © 2019 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

¹<https://www.capito.eu/> (last accessed: June 27, 2019)

(Council of Europe, 2001) levels A1, A2, and B1. Being commercially orientated, *capito* does not make its CEFR adaptation publicly available.

In this paper, we present an unsupervised machine learning (clustering) approach to analysing texts in simplified German with the aim of investigating evidence of multiple complexity levels. To the best of our knowledge, this is the first study of its kind. We apply linguistic features used in previous supervised machine learning research (classification) and additionally exploit structural and typographic characteristics of simplified texts that have been described in the literature but not incorporated into clustering and/or classification approaches in the context of simplified language.

The remainder of this paper is structured as follows: Section 2 presents the research background. Section 3 describes our approach, introducing a novel dataset (Section 3.1), the feature design and engineering (Section 3.2), the clustering experiments (Section 3.3), and a discussion thereof (Section 3.4). Section 4 offers a conclusion and an outlook on future research questions.

2 Research Background

Two natural language processing tasks deal with the concept of simplified language: automatic readability assessment and automatic text simplification. Readability assessment refers to the process of determining the level of difficulty of a text. Traditionally, this has involved taking into account readability measures based on surface features such as the number of syllables in a word or number of words in a sentence, e.g., via the Flesch Reading Ease Score (Flesch, 1948). Recently, more sophisticated models employing deeper linguistic features such as lexical, semantic, morphological, morphosyntactic, syntactic, pragmatic, discourse, psycholinguistic, and language model features have been proposed (Collins-Thompson, 2014; Dell’Orletta et al., 2014; Heimann Mühlenbock, 2013; Schwarm and Ostendorf, 2005).

Readability assessment implies the existence of multiple complexity levels. Complexity levels are identified, e.g., along school grades or levels of the CEFR (Hancke, 2013; Pilan and Volodina, 2018; Reynolds, 2016; Vajjala and Lõo, 2014).

The work presented in this paper represents a preliminary stage of the readability assessment task for simplified German in that it investigates

empirically whether different complexity levels exist in previous German simplification practice in the first place.

3 Clustering Simplified German texts

3.1 Dataset

Battisti and Ebling (2019) compiled a corpus of German/simplified German texts for use in automatic readability assessment and automatic text simplification. The corpus represents an enhancement of a parallel (German/simplified German) corpus created by Klaper et al. (2013). Compared to its predecessor, the corpus of Battisti and Ebling (2019) contains additional parallel data and newly contains monolingual-only data as well as structural and typographic information.

The authors collected PDFs and web pages from 92 different domains of public offices, translation agencies, and organisations publishing content in German and simplified German. Overall, the corpus consists of 6,217 documents (378 parallel and 5,461 monolingual). Metadata was recorded in the Open Language Archives Community (OLAC) Standard² and converted into the metadata standard CMDI of CLARIN, a European research infrastructure for language resources and technology.³ If available, information on the language level of a simplified German text (typically A1, A2, or B1) was stored in the metadata. 52 websites and 233 PDFs (amounting to approximately 26,000 sentences) have an explicit language level label.

Linguistic annotation was added automatically using ParZu (Sennrich et al., 2009) (for tokens and dependency parses), NLTK (Bird et al., 2009) (for sentence segmentation), TreeTagger (Schmid, 1995) (for part-of-speech tags and lemmas), and Zmorge (Sennrich and Kunz, 2014) (for morphological units). In addition, information on text structure (e.g., paragraphs, lines), typography (e.g., boldface, italics), and images (content, position, and dimensions) was added. The annotations were stored in the Text Corpus Format by WebLicht (TCF) developed as part of CLARIN.⁴

For the experiments reported in this paper, we

²<http://www.language-archives.org/OLAC/olacms.html> (last accessed: June 27, 2019)

³<https://www.clarin.eu/> (last accessed: June 27, 2019)

⁴<https://weblicht.sfs.uni-tuebingen.de/weblichtwiki/index.php/TheTCFFormat> (last accessed: June 27, 2019)

considered the monolingual documents of the corpus, i.e., the monolingual-only documents as well as the simplified German side of the parallel data. This amounted to 5,839 texts (193,845 sentences).

3.2 Features

In addition to constituting the first approach to investigating simplified German texts using unsupervised machine learning, the unique contribution of this paper consists of leveraging information that has been shown to be characteristic of simplified language (Arfé et al., 2018; Bock, 2018; Bredel and Maaß, 2016) but has not been incorporated into machine learning approaches involving simplified language. Specifically, we considered features derived from text structure (e.g., paragraphs, lines), typography (e.g., font type, font style), and image (content, position, and dimensions) information.

In a simplified text, typographical information, such as boldface and italics, serves as a discourse marker signalling words and phrases that require particular attention and convey different purposes (Arfé et al., 2018). Leveraging the concepts of multi-modality and multi-codality in the psychology of perception (Schnotz, 2014), images⁵ are supposed to support the text by activating previous knowledge and exemplifying the objects in the text (Bredel and Maaß, 2016).

Subset	Features	Number
1	All	115
2	Surface	26
3	Deeper	89
4	Lexical + semantic	17
5	Morphological + syntactic	72

Table 1: Subsets of feature combinations.

Altogether, the feature set comprised 115 features arranged into five feature groups, as shown in Table 1. Subset 3 (“Deeper”) consisted of lexical, semantic, morphological, and syntactic features. “Surface” is short for surface, structural, and typographic features.

Surface, structural, and typographic features: We took advantage of the structural and typographic information included in the corpus

⁵For the sake of simplicity, the term “images” here subsumes pictures, pictograms, photographs, graphics, and maps.

(cf. Section 3.1) and introduced as features the number of images, paragraphs, lines, words of a specific font type and style, and adherence to a one-sentence-per-line rule. We additionally included the number of digits and numbers in words (Saggion, 2017), number of abbreviations and initial letters, and the number of individual punctuation marks and special characters. Among the special characters was the *Mediopunkt* (‘centred dot’), a typographic device proposed by Maaß (2015) for visually segmenting compound words. We also computed the *Läsbarhetsindex* (‘readability index’, LIX) (Björnsson, 1968).⁶

Lexical and semantic features: This group included features for lexical richness, lexical variation (e.g., nominal ratio, noun/pronoun ratio, bilogarithmic TTR (Vajjala and Meurers, 2012)), word frequency based on the German reference corpus DeReKo (Lüngen, 2017), and lists of words classified at different perceptive levels (Glaboniat et al., 2005). We also included question words and named entities, which may strain the reading comprehension process if the target reader does not have the appropriate knowledge.

Morphological, morphosyntactic, and syntactic features: In this group, we included particles, prepositions, demonstrative and personal pronouns, and (separately) first-, second-, and third-person pronouns. We additionally counted adverbs, modal verbs, subjunctions, and conjunctions. We added genitive attributes in relation to *von*+dative constructions.⁷ We additionally included the number of negative forms, the presence of pre- and post-modifiers, and impersonal constructions. We took advantage of the verbal morphology and included verbal mood- and tense-based features (Dell’Orletta et al., 2011). We also considered direct vs. indirect speech constructions, the types of subordinate clauses as well as features based on word and sentence order.

⁶ $LIX = N_w / N_s + (W \times 100) / N_w$, where N_w is the number of words, N_s is the number of sentences, and W is the percentage of tokens longer than six characters.

⁷In German, the genitive attribute can be substituted by a *von*+dative construction. Importantly, this is a case of simplified German conflicting with the grammar of Standard German, which encourages the use of the former construction.

3.3 Experiments and Results

3.3.1 Method

We applied agglomerative hierarchical clustering. We used the `scipy`⁸ toolkit alongside with models recursively created with the `scikit-learn`⁹ library. The data matrix was created using the cosine similarity metric and the average linkage function. Because of the significant variation in length of the documents, we normalised the features by dividing the values by the length of each document expressed in tokens. We then performed principal component analysis (PCA) to diminish the sparseness of the data matrix and avoid the curse-of-dimensionality trap. In a second experiment, we applied feature agglomeration instead of PCA prior to clustering. Feature agglomeration allows for a straightforward interpretation of the results.

Given the lack of a ground truth for our data, we evaluated the experiments using the following metrics: silhouette score, Calinski-Harabasz index, and Elbow method. These metrics were also used to choose the optimal number of clusters.

3.3.2 Results

Table 2 shows the results of the first three iterations of our clustering approach after the feature agglomeration step. We observed that a value between 2 and 4 (inclusive) represented a good clustering solution for the whole corpus according to the metrics. A dendrogram corroborated these results (cf. Figure 1).

Upon inspection of the clusters, we found the main differences to be due to the following features: number of nouns, number of verbs, number of paragraphs, adherence to one-sentence-per-line rule, number of interrogative clauses, number of different fonts, and number of words in bold. Considering the mean ratio of the features in a two-cluster solution, Cluster 1 displayed a higher frequency of nouns (0.31 vs. 0.24) and adjectives (0.9 vs. 0.6) and a lower frequency of verbs (0.13 vs. 0.17) than Cluster 2, which in turn included a slightly higher rate of images (0.008 vs. 0.004).

3.4 Discussion

The inverse proportion of the mean ratios concerning nouns and verbs (cf. Section 3.3.2) suggested

that Cluster 1 included texts focusing on objects or concepts, since verbs (events, actions, etc.) had been turned into nouns (concepts, things, etc.) following the linguistic process of nominalisation, while the linguistic structure of texts in Cluster 2 was simpler.

Figure 2 visualises the box plots of six of the surface features of Subset 2 (number of full stops, number of commas, adherence to one-sentence-per-line rule, number of paragraphs, number of different fonts, number of images) based on the three-cluster solution suggested by the agglomerative hierarchical approach. The first cluster consisted of texts that followed the one-sentence-per-line rule, featured a low frequency of commas, and a high number of paragraphs. These characteristics are crucial properties of simplified texts. Our findings further emphasise the importance of distinguishing among different types of punctuation marks in the context of simplified language: while for commas, a low frequency is indicative of textual simplicity, the reverse is true for full stops. Texts included in Cluster 1 did not contain images. This outcome relates to the results of a more recent study by Bock (2018), according to which images should be used with caution even in simplified German texts to avoid the potential of distraction and cognitive overload.

4 Conclusion and Outlook

In this paper, we have presented the first approach to investigating simplified German texts by means of unsupervised machine learning techniques as a basis for future readability assessment studies on this language variety. In addition, we have introduced novel features that have been described in the literature but not incorporated into machine learning (clustering and/or classification) approaches in the context of simplified language, notably: number of images, number of paragraphs, number of lines, number of words of a specific font type, and adherence to a one-sentence-per-line rule. Our findings provide evidence that existing texts are not simplified at a unique complexity level of German. We have demonstrated that features based on structural information are capable of accounting for the different complexity levels found.

As a next step, we will use the results of the experiments presented in this paper to establish a framework of inductively generated complexity

⁸<https://www.scipy.org/> (last accessed: June 27, 2019)

⁹<https://scikit-learn.org/stable/> (last accessed: June 27, 2019)

	Subset 1		Subset 2		Subset 3		Subset 4		Subset 5	
	Sil	CH	Sil	CH	Sil	CH	Sil	CH	Sil	CH
2	0.601	3867.1	0.373	1135.2	0.675	5214.2	0.693	3593.9	0.695	5463.2
3	0.532	2476.2	0.372	1266.3	0.617	3329.5	0.55	1824.8	0.572	3273.9
4	0.456	1698.3	0.493	1417.6	0.592	2572.7	0.505	1248.9	0.51	2517.8

Table 2: Comparison of the silhouette scores (Sil) and Calinski-Harabasz indices (CH) after feature agglomeration on all data samples.

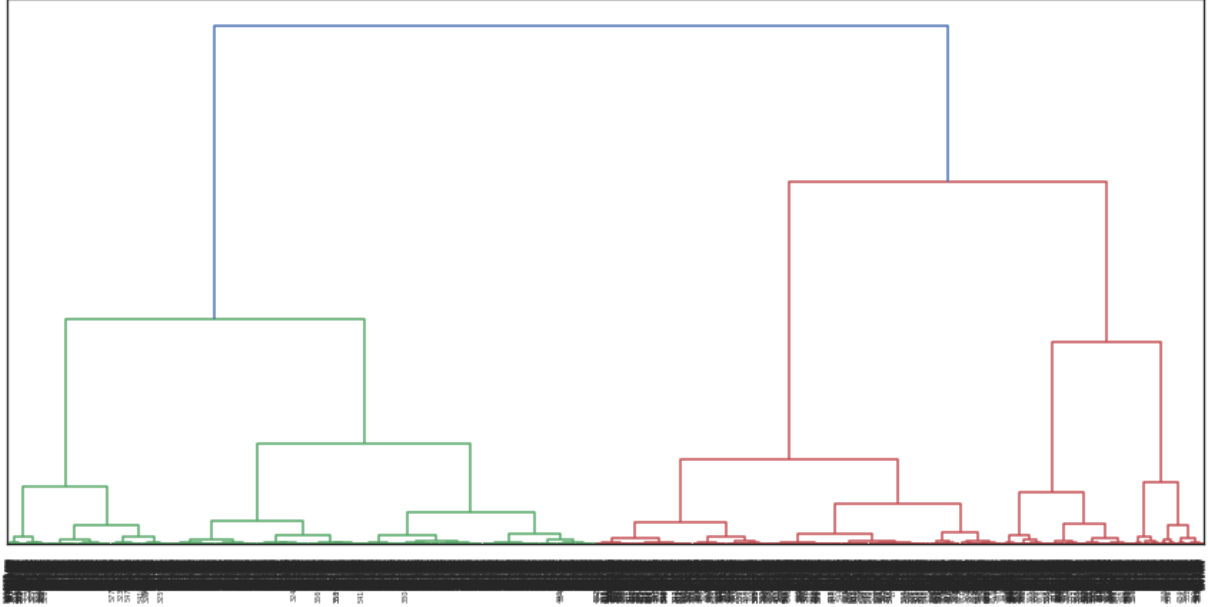


Figure 1: Dendrogram of the texts considering agglomerated features of Subset 1.

levels. This framework will serve as the basis for readability assessment in the context of simplified German. Knowledge derived from our study can also inform automatic and manual approaches to simplification of German.

References

- Barbara Arfé, Lucia Mason, and Inmaculada Fajardo. 2018. Simplifying informational text structure for struggling readers. *Reading and Writing*, 31(9):2191–2210.
- Alessia Battisti and Sarah Ebling. 2019. A corpus for automatic readability assessment and text simplification of german. arXiv:1909.09067.
- Steven Bird, Edward Loper, and Ewan Klein. 2009. *Natural Language Processing with Python*. O’Reilly Media Inc.
- Carl-Hugo Björnsson. 1968. *Läsbarhet*. Liber, Stockholm.
- Bettina M. Bock. 2014. “Leichte Sprache”: Abgrenzung, Beschreibung und Problemstellungen aus Sicht der Linguistik. *Sprache barrierefrei gestalten*, pages 17–51.
- Bettina M. Bock. 2018. “Leichte Sprache” - Kein Regelwerk. Sprachwissenschaftliche Ergebnisse und Praxisempfehlungen aus dem LeiSA-Projekt. Technical report, Universität Leipzig.
- Ursula Bredel and Christiane Maaß. 2016. *Leichte Sprache: Theoretische Grundlagen. Orientierung für die Praxis*. Duden, Berlin.
- Bundesministerium für Arbeit und Soziales. 2011. Verordnung zur Schaffung barrierefreier Informationstechnik nach dem Behindertengleichstellungsgesetz (Barrierefreie-Informationstechnik-Verordnung-BITV 2.0). Technical Report Teil 1.
- Kevyn Collins-Thompson. 2014. Computational assessment of text readability. A survey of current and future research. *ITL International Journal of Applied Linguistics*, 165(2):97–135.
- Council of Europe. 2001. *Common European Framework of Reference for Languages: Learning, teaching, assessment*. Cambridge University Press, Cambridge.

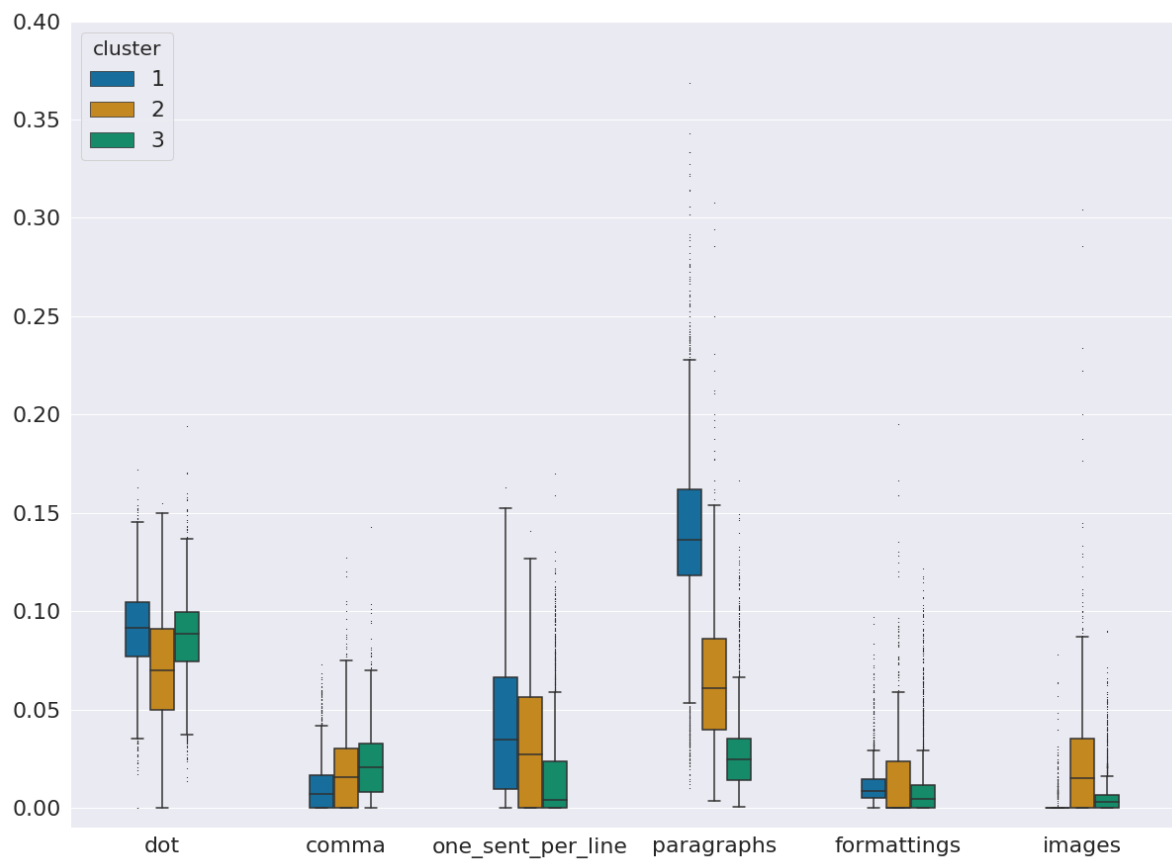


Figure 2: Six features of Subset 2.

- Felice Dell’Orletta, Simonetta Montemagni, and Giulia Venturi. 2011. READ-IT: Assessing readability of Italian texts with a view to text simplification. In *Proceedings of the Second Workshop on Speech and Language Processing for Assistive Technologies*, pages 73–83, Edinburgh, Scotland, UK. Association for Computational Linguistics.
- Felice Dell’Orletta, Martijn Wieling, Giulia Venturi, Andrea Cimino, and Simonetta Montemagni. 2014. Assessing the readability of sentences: Which corpora and features? In *Proceedings of the Ninth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 163–173, Baltimore, Maryland, June. Association for Computational Linguistics.
- Rudolph Flesch. 1948. A new readability yardstick. *Journal of Applied Psychology*, 32:221–233.
- Manuela Glaboniat, Martin Müller, Paul Rusch, Helen Schmitz, and Lukas Wertenschlag. 2005. *Profile Deutsch*. Klett Langenscheidt, Berlin/Munich, Germany.
- Julia Hancke. 2013. Automatic Prediction of CEFR Proficiency Levels Based on Linguistic Features of Learner Language. Master’s thesis, University of Tübingen, Germany.
- Katarina Heimann Mühlenbock. 2013. *I see what you mean: Assessing readability for specific target groups*. Ph.D. thesis, University of Gothenburg.
- Inclusion Europe. 2009. Information für alle: Europäische Regeln, wie man Informationen leicht lesbar und leicht verständlich macht. Technical report, Inclusion Europe.
- Gudrun Kellermann. 2014. Leichte und Einfache Sprache Versuch einer Definition. In *Aus Politik und Zeitgeschichte*, volume 64, pages 9–11.
- David Klaper, Sarah Ebling, and Martin Volk. 2013. Building a German/Simple German parallel corpus for automatic text simplification. In *ACL Workshop on Predicting and Improving Text Readability for Target Reader Populations*, pages 11–19, Sofia, Bulgaria.
- Harald Lungen. 2017. DEREKO - Das Deutsche Referenzkorpus. *Zeitschrift für Germanistische Linguistik*.
- C. Maaß. 2015. *Leichte Sprache: Das Regelbuch*. Barrierefreie Kommunikation. Lit Verlag.
- Netzwerk Leichte Sprache. 2013. Die Regeln für Leichte Sprache. Technical report.
- Ildiko Pílan and Elena Volodina. 2018. Investigating the importance of linguistic complexity features across different datasets related to language learning. In *Proceedings of the Workshop on Linguistic Complexity and Natural Language Processing*, pages 49–58, Santa Fe, New-Mexico.
- Robert Reynolds. 2016. Insights from Russian second language readability classification: complexity-dependent training requirements, and feature evaluation of multiple categories. In *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 289–300, San Diego, California.
- Horacio Saggion. 2017. *Automatic Text Simplification*. Morgan & Claypool Publishers.
- Helmut Schmid. 1995. Improvements in part-of-speech tagging with an application to German. In *Proceedings of the EACL’95 SIGDAT Workshop*, pages 47–50, Dublin, Ireland.
- Wolfgang Schnitz, 2014. *An Integrated Model of Text and Picture Comprehension*, pages 72–103. Cambridge University Press, second edition.
- Sarah E. Schwarm and Mari Ostendorf. 2005. Reading level assessment using support vector machines and statistical language models. In *Proceedings of the 43rd Annual meeting of the Association for Computational Linguistics*, pages 523–530.
- Rico Sennrich and Beat Kunz. 2014. Zmorge: A German Morphological Lexicon Extracted from Wiktionary. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation*, pages 1063–1067, Reykjavik, Iceland. European Language Resources Association.
- Rico Sennrich, Gerold Schneider, Martin Volk, and Martin Warin. 2009. A new hybrid dependency parser for German. In *Proceedings of the Biennial GSCL Conference*, pages 115–124, Potsdam.
- United Nations. 2006. Convention on the Rights of Persons with Disabilities and Optional Protocol.
- Sowmya Vajjala and Kaidi Lõo. 2014. Automatic CEFR level prediction for Estonian learner text. In *Proceedings of the third workshop on NLP for computer-assisted language learning*, volume 107, pages 113–127, Uppsala, Sweden.
- Sowmya Vajjala and Detmar Meurers. 2012. On Improving the Accuracy of Readability Classification using Insights from Second Language Acquisition. In *Proceedings of the 7th workshop on building educational applications using NLP*, pages 163–173, Montreal, Canada.

Almawave-SLU: a New Dataset for SLU in Italian

Valentina Bellomaria, Giuseppe Castellucci, Andrea Favalli and Raniero Romagnoli

Language Technology Lab
Almawave srl
[first name initial].[last name]@almawave.it

Abstract

The widespread use of conversational and question answering systems made it necessary improve the performances of speaker intent detection and understanding of related semantic slots, i.e., Spoken Language Understanding (SLU). Often, these tasks are approached with supervised learning methods, which needs considerable labeled datasets. This paper¹ presents the first Italian dataset for SLU in voice assistants scenario. It is the product of a semi-automatic procedure and is used as a benchmark of various open source and commercial systems.

1 Introduction

Conversational interfaces, e.g., Google’s Home or Amazon’s Alexa, are becoming pervasive in daily life. As an important part of any conversation, language understanding aims at extracting the meaning a partner is trying to convey. Spoken Language Understanding (SLU) plays a fundamental role in such a scenario. Generally speaking, in SLU a spoken utterance is first transcribed, then semantic information is extracted. Language understanding, i.e., extracting a semantic “frame” from a transcribed user utterance, typically involves: i) Intent Detection (ID) and ii) Slot Filling (SF) (Tur et al., 2010). The former makes the classification of a user utterance into an intent, i.e., the purpose of the user. The latter finds what are the “arguments” of such intent. As an example, let us consider Figure 1, where the user asks for playing a song (Intent=PlayMusic) (*with or without you*, Slot=song) of an artist (*U2*, Slot=artist). Usually, supervised learning methods are adopted

play	with	or	without	you	by	U2
riproduci	with	or	without	you	degli	U2
O	B-song	I-song	I-song	I-song	O	B-artist

Figure 1: An example of Slot Filling in IOB format for a sentence with intent *PlayMusic*.

for SLU. Their efficacy strongly depends on the availability of labeled data. There are various approaches to the production of labeled data, depending on the intricacy of the problem, on the characteristics of the data, and on the available resources (e.g., annotators, time and budget). When the reuse of existing public data is not feasible, manual labeling should be accomplished, eventually by automating part of the labeling process.

In this work, we present the first public dataset for the Italian language for SLU. It is generated by a semi-automatic procedure from an existing English dataset annotated with intents and slots. We have translated the sentences into Italian and reported the annotations based on a token span algorithm. Then, the translation, spans and consistency of the entities in Italian have been manually validated. Finally, the dataset is used as benchmark for NLU systems. In particular, we will compare a recent state-of-the-art (SOTA) approach (Castellucci et al., 2019) with Rasa (ras, 2019) taken from the open source world, IBM Watson Assistant (wat, 2019), Google DialogFlow (dia, 2019) and, finally, Microsoft LUIS (msl, 2019), some commercial solutions in use.

Following, in section 2 related works will be discussed; In section 3 the dataset generation will be discussed. Section 4 we will present the experiments. Finally, in section 5 we will draw the conclusions.

2 Related Work

SLU has been addressed in the Natural Language Processing community mainly in the English lan-

¹Copyright © 2019 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

guage. A well-known dataset used to demonstrate and benchmark various NLU algorithms is Airline Travel Information System (ATIS) (Hemphill et al., 1990) dataset, which consists of spoken queries on flight related information. In (Braun et al., 2017) three dataset for Intent classification task were presented. *AskUbuntu Corpus* and *Web Application Corpus* were extracted from Stack-Exchange and the third one, i.e., *Chatbot Corpus*, was originated from a Telegram chatbot. The newer multi-intent dataset SNIPS (Coucke et al., 2018) is the starting point for the work presented in this paper. An alternative approach to manual or semi-automatic labeling is the one proposed by the data scientists of the Snorkel project with Snorkel Drybell (Bach et al., 2018) that aims at automating the labeling through the use of data programming. Other works have explored the possibility of creating datasets in a language starting from datasets in other languages, such as (Jabaian et al., 2010) and (Stepanov et al., 2013). Regarding the Italian language two main works can be pointed out (Raymond et al., 2008; Vanzo et al., 2016). Our work differs mainly in the application domain (i.e., we focus on the voice assistants scenario). In particular, (Raymond et al., 2008) mainly focuses on dialogues in a customer service scenario; (Vanzo et al., 2016) focuses on Human-Robot interaction.

3 Almawave-SLU: A new dataset for Italian SLU

We created the new dataset ² starting from the SNIPS dataset (Coucke et al., 2018), which is in English. It contains 14,484 annotated examples³ with respect to 7 intents and 39 slots. In table 1 an excerpt of the dataset is shown. We started from this dataset as: i) it contains a reasonable amount of examples; ii) it is multi-domain; iii) we believe it could represent a more realistic setting in today’s voice assistants scenario.

We performed a semi-automatic procedure consisting of two phases: an automatic translation with contextual alignment of intents and slots; a manual validation of the translations and annotations. The resulting dataset, i.e., Almawave-SLU, has fewer training examples, a total of 7,142 and the same number of validation and test examples of the original dataset. Again, 7

intents and 39 slots have been annotated. Table 2 shows the distribution of examples for each intent.

3.1 Translation and Annotation

In a first phase, we translated each English example in Italian by using the Translator Text API: part of the Microsoft Azure Cognitive Services. In order to create a more valuable resource in Italian, we also performed an automatic substitution of the names of movies, movie theatres, books, restaurants and of the locations with some Italian counterpart. First, we collected from the Web a set E of about 20,000 Italian versions of such entities; then, we substituted each entity in the sentences of the dataset with one randomly chosen from E .

After the translation, an automatic annotation was performed. The intent associated with the English sentence has been copied to its Italian counterpart. Slots have been transferred by aligning the source and target tokens⁴ and by copying the corresponding slot annotation. In case of exceptions, e.g., multiple alignments on the same token or missing alignment, we left the token without annotation.

3.2 Human Revision

In a second phase, the dataset was divided into 6 different sets, each containing about 1,190 sentences. Each set was assigned to 2 annotators⁵, and each was asked to review the translation from English to Italian and the reliability of the automatic annotation. The guideline was to consider a valid annotation when both the alignment and the semantic slots were correct. Moreover, also a semantic consistency check was performed: e.g., served dish and restaurant type or city and region or song and singer. The 2 annotators have been used to cross-check the annotations, in order to provide more reliable revisions. When the 2 annotators disagreed, the annotations have been validated by a third different annotator.

During the validation phase some interesting phenomena emerged. ⁶ For example, there have been cases of inconsistency between the restaurant name and the type of served dish when the name of the restaurant mentioned the kind of food served, e.g., "*Prenota un tavolo da Pizza Party per mangiare noodles*". There were also wrong associations between the type of restaurant and service

²The Almawave-SLU dataset is available for download. To obtain it, please send an e-mail to the authors.

³There are 13084, 700 and 700 for training, validation and test, respectively.

⁴The alignment was provided by the Translator API.

⁵A total of 6 annotators were available.

⁶Some inconsistencies were in the original dataset

AddToPlaylist	Add the song virales de siempre by the cary brothers to my gym playlist.
BookRestaurant	I want to book a top-rated brasserie for 7 people.
GetWeather	What kind of weather will be in Ukraine one minute from now?
PlayMusic	Play Subconscious Lobotomy from Jennifer Paull.
RateBook	Rate The children of Niobe 1 out of 6 points.
SearchCreativeWork	Looking for a creative work called Plant Ecology
SearchScreeningEvent	Is Bartok the Magnificent playing at seven AM?

Table 1: Examples from the SNIPS dataset. The first column indicates the intent, the second columns contains an example.

requested, e.g., *"Prenota nell'area piscina per 4 persone in un camion-ristorante"*. A truck restaurant is actually a van equipped for fast-food in the street. Again, among the cases of unlikely associations resulting from automatic replacement, the inconsistency between temperatures and cities is mentioned, in cases like "snow in the Sahara". Another type of problem occurred when the same slot was used to identify very different objects. For example, for the intent *SearchCreativeWork*, the slot *object_name* was used for paintings, games, movies, etc... We can observe and analyze a couple of examples for this intent: *Can you find me the work, The Curse of Oak Island ?* and *Can you find me, Hey Man ?*. The first example contains *The Curse of Oak Island*, that is a television series and the second refers to *Hey Man* that is a music album, but both are labeled as *object_name*, where the *object_type* are different and not specified. In all these cases, the annotators were asked to correct the sentences and the annotations, accordingly. Again, in the case of *BookRestaurant* intent a manual revision was made when in the same sentence the city and state coexist: to make the data more relevant to the Italian language, the region relative to the city is changed, e.g., *"I need a table for 5 at a highly rated gastropub in Saint Paul, MN"* is translated and adapted for Italian in *"Vorrei prenotare un tavolo per 5 in un gastropub molto apprezzato a Biella, Piemonte"*.

	Train	Train-R	Valid	Test
AddToPlayList	744	185	100	124
BookRestaurant	967	250	100	92
GetWeather	791	195	100	104
PlayMusic	972	240	100	86
RateBook	765	181	100	80
SearchCreativeWork	752	172	100	107
SearchScreeningEvent	751	202	100	107

Table 2: Almayave-SLU Datasets statistics. Train-R is the reduced training set.

3.3 Automatic Translation Analysis

In many cases, machine translation lacked context awareness: this isn't an easy task due to phenomena as polysemy, homonymy, metaphors and idioms. There can be problems of lexical ambiguities when a word has more than one meaning and can produce wrong interpretations. For example, the verb "to play" can mean "spend time doing enjoyable things", such as "using toys and taking part in games", "perform music" or "perform the part of a character".

Human intervention occurred to maintain the meaning of the text dependent on cultural and situational contexts. Different translation errors were modified by the annotators. For example, the automatic translation of the sentence *Play Have You Met Miss Jones by Nicole from Google Music.* was *Gioca hai incontrato Miss Jones di Nicole da Google Music.*, but the correct Italian version is *Riproduci Have You Met Miss Jones di Nicole da Google Music.*. In this case the wrong translation of the verb *play* causes a meaningless sentence.

Often, translation errors are due to the presence of prepositions, that have the same function in Italian as they do in English. Unfortunately, these cannot be directly translated. Each preposition is represented by a group of related senses, some of which are very close and similar while others are rather weak and distant. For example, the Italian preposition "di" can have six different English counterparts – of, by, about, from, at, and than. For example, in the SNIPS dataset the sentence *I need a table for 2 on feb. 18 at Main Deli Steak House* was translated as *Ho bisogno di un tavolo per 2 su Feb. 18 presso Main Deli Steak House.* Here, the translation of "on" is wrong: the correct Italian version should translate it as "il". Another example with wrong preposition translation is the sentence *"What will the weather be one month from now in Chad ?"*, the automatic translation of "one month from now" is "un mese da ora" but the correct translation is "tra un mese".

Common errors were in the translation of temporal expression, that are different between Italian and English. For example the translation of the sentence “*Book a table in Fiji for zero a.m*” was “*Prenotare un tavolo in Fiji per zero a.m*” but in Italian “zero a.m” is “mezzanotte”.

Other errors were specific of some intents, as they tend to have more slangs. For example, the translation of *GetWeather*’s sentences was problematic because the main verb is often misinterpreted, while in the sentences related to the intent *BookRestaurant* a frequent failure occurred on the interpretation of prepositions. For example, the sentence “*Will it get chilly in North Creek Forest?*” was translated as “*Otterrà freddo in North Creek Forest?*”, while the correct translation is “*Farà freddo a North CreekForest?*”. In this case, the system misinterpreted the context, assigning to “get” the wrong meaning.

4 Benchmarking SLU Systems

Nowadays, there are several human-machine interacting platforms, commercial and open source. Machine learning algorithms enable these systems to understand natural language utterances, match them to intents, and extract structured data. We decided to use the Almagest-SLU dataset with the following SLU systems.

4.1 SLU Systems

RASA. RASA (ras, 2019) is an open source alternative to popular NLP tools for the classification of intentions and the extraction of entities. Rasa contains a set of high-level APIs to produce a language parser through the use of NLP and ML libraries, via the configuration of the pipeline and embeddings. It seems to be very fast to train, does not require great computing power and, despite this, it seems to get excellent results.

LUIS. Language Understanding service (msl, 2019) allows the construction of applications that can receive input in natural language and extract the meaning from it through the use of Machine Learning algorithms. LUIS was chosen as it provides also an easy-to-use graphical interface dedicated to less experienced users. For this system the computation is completely done remotely and no configuration is needed.

Watson Assistant. IBM’s Watson Assistant (wat, 2019) is a white label cloud service that al-

lows software developers to embed a virtual assistant, that use Watson AI machine learning and NLU, in their software. Watson Assistant allows customers to protect information gathered through user interaction in a private cloud. It was chosen because it was conceived for an industrial market and for its long tradition in this task.

DialogFlow. Dialogflow (dia, 2019) is a Google service to build engaging voice and text-based conversational interfaces, powered by a natural language understanding (NLU) engine. Dialogflow makes it easy to connect the bot service to a number of channels and runs on Google Cloud Platform, so it can scale to hundreds of millions of users. DialogFlow was chosen due to its wide distribution and ease of use of the interface.

Bert-Joint. It is a SOTA approach to SLU adopting a joint Deep Learning architecture in an attention-based recurrent frameworks (Castellucci et al., 2019). It exploits the successful Bidirectional Encoder Representations from Transformers (BERT) model to pre-train language representations. In (Castellucci et al., 2019), the authors extend the BERT model in order to perform the two tasks of ID and SF jointly. In particular, two classifiers are trained jointly on top of the BERT representations by means of a specific loss function.

4.2 Experimental Setup

Almagest-SLU has been used for training and evaluation of Rasa, Luis, Watson Assistant, DialogFlow and Bert-Joint. Another evaluation is made on 3 different training datasets, i.e Train-R, of reduced dimensions with respect to the Almagest-SLU, each about 1,400 sentences equally distributed on intent.

The train/validation/test split used for the evaluations is 5,742 (1,400 for Train-R), 700 and 700, respectively. Regarding Rasa, we used version 1.0.7, and we adopted the standard “supervised embeddings” pipeline, since it is recommended in the official documentation. This pipeline consists of a *WhiteSpaceTokenizer*, that was modified to avoid the filter of punctuation tokens, a *Regex Featurizer*, a *Conditional Random Field* to extract entities, a *Bag-of-words Featurizer* and an *Intent Classifier*. LUIS was tested against the api v2.0, and the loading of data to train the system with LUIS APP VERSION 0.1. Unfortunately Watson

System	Eval-1 with Train set			Eval-2 with Train-R set		
	Intent	Slot	Sentence	Intent	Slot	Sentence
Rasa	96.42	85.40	65.76	93.84	78.58	52.25
LUIS	95.99	79.47	50.57	94.46	72.51	35.53
Watson Assistant	96.56	-	-	95.03	-	-
Dialogflow	95.56	74.62	46.16	93.60	65.23	36.68
Bert-Joint	97.6	90.0	77.1	96.13	83.04	65.23

Table 3: Overall scores for Intent and Slot

Assistant supports only English models for the annotations of contextual entities, i.e. slots; therefore, we have only measured the intents⁷. Regarding DialogFlow, a “Standard” (free) utility has been created with API version 2; the python library “dialogflow” has been used for the predictions.⁸ DialogFlow allows the choice between pure ML mode (“ML only”) and hybrid rule-based and ML mode (“match mode”). We chosen ML mode. Regarding the BERT-Joint system, a pre-trained BERT model is adopted, which is available on the BERT authors website⁹. This model is composed of 12-layer and the size of the hidden state is 768. The multi-head self-attention is composed of 12 heads for a total of 110M parameters. As suggested in (Castellucci et al., 2019), we adopted a dropout strategy applied to the final hidden states before the intent/slot classifiers. We tuned the following hyper-parameters over the validation set: (i) number of epochs among (5, 10, 20, 50); (ii) Dropout keep probability among (0.5, 0.7 and 0.9). We adopted the Adam optimizer (Kingma and Ba, 2015) with parameters $\beta_1 = 0.9$, $\beta_2 = 0.999$, L2 weight decay 0.01 and learning rate $2e-5$ over batches of size 64.

4.3 Experimental Results

In table 3 the performances of the systems are shown. The SF performance is the F1 while the ID and Sentence performances are measured with the accuracy. We also show an evaluation carried out with models trained on three different split of reduced size derived from the whole dataset. The reported value is the average of measurements obtained separately on the entire test dataset.

⁷Refer to Table 3. Entity feature support details at <https://cloud.ibm.com/docs/services/assistant?topic=assistant-language-support>

⁸<https://cloud.google.com/dialogflow/docs/reference/rest/v2/projects.agent.intents#Part>

⁹https://storage.googleapis.com/bert_models/2018_11_23/multi_cased_L-12_H-768_A-12.zip

Regarding the ID task, all models are performing similarly, but Bert-Joint F1 score is slightly higher than others. For SF task, notice that there are significant differences between LUIS, DialogFlow and Rasa performances.

Finally, Bert-Joint achieved the top score on joint classification, in the assessments with the two different sizes of the dataset. The adaptation of nominal entities in Italian may have amplified the problem for the other models.

5 Conclusion

The contributions of this work are two-fold: first, we presented and released the first Italian SLU dataset (Almawave-SLU) in the voice assistants context. It is composed of 7,142 sentences annotated with respect to intents and slots, almost equally distributed on the 7 different intents. The effort spent on the construction of this new resource, according to the semi-automatic procedure described, is about 24 FTE¹⁰, with an average production of about 300 examples per day. We consider this effort lower than typical efforts to create linguistic resources from scratch.

Second, we compared some of the most popular NLU services with this data. The results show they all have similar features and performances. However, compared to another specific architecture for SLU, i.e., Bert-Joint, they perform worse. It was expected and it demonstrates the Almawave-SLU can be a valuable dataset to train and test SLU systems on the Italian language. In future, we hope to continuously improve the data and to extend the dataset.

6 Acknowledgment

The authors would like to thank to David Alessandrini, Silvana De Benedictis, Raffaele Mazzocca, Roberto Pellegrini and Federico Wolenski for the support in the annotation, revision and evaluation phases.

¹⁰Full Time Equivalent

References

- Stephen H. Bach, Daniel Rodriguez, Yintao Liu, Chong Luo, Haidong Shao, Cassandra Xia, Souvik Sen, Alexander J. Ratner, Braden Hancock, Houman Alborzi, Rahul Kuchhal, Christopher Re, and Rob Malkin. 2018. Snorkel drybell: A case study in deploying weak supervision at industrial scale. *CoRR*, abs/1812.00417.
- Daniel Braun, Adrian Hernandez-Mendez, Florian Matthes, and Manfred Langen. 2017. Evaluating natural language understanding services for conversational question answering systems. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 174–185, Saarbrücken, Germany, August. Association for Computational Linguistics.
- Giuseppe Castellucci, Valentina Bellomaria, Andrea Favalli, and Raniero Romagnoli. 2019. Multilingual intent detection and slot filling in a joint bert-based model. *CoRR*, abs/1907.02884.
- Alice Coucke, Alaa Saade, Adrien Ball, Theodore Bluche, Alexandre Caulier, David Leroy, Clement Doumouro, Thibault Gisselbrecht, Francesco Caltagirone, Thibaut Lavril, Mael Primet, and Joseph Dureau. 2018. Snips voice platform: an embedded spoken language understanding system for private-by-design voice interfaces. *CoRR*, abs/1805.10190.
2019. Google dialogflow. <https://dialogflow.com>.
- Charles T. Hemphill, John J. Godfrey, and George R. Doddington. 1990. The atis spoken language systems pilot corpus. In *Speech and Natural Language: Proceedings of a Workshop Held at Hidden Valley*.
- Bassam Jabaian, Laurent Besacier, and Fabrice Lefèvre. 2010. Investigating multiple approaches for slu portability to a new language. In *INTER-SPEECH*.
- Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.
2019. Microsoft luis on azure. <https://azure.microsoft.com/it-it/services/cognitive-services/language-understanding-intelligent-service/>.
2019. Rasa: Open source conversational ai. <https://rasa.com/>.
- Christian Raymond, Kepa Joseba Rodriguez, and Giuseppe Riccardi. 2008. Active annotation in the LUNA Italian corpus of spontaneous dialogues. In *LREC 2008*.
- Evgeny Stepanov, Ilya Kashkarev, Ali Orkan Bayer, Giuseppe Riccardi, and Arindam Ghosh. 2013. Language style and domain adaptation for cross-language slu porting. pages 144–149, 12.
- G. Tur, D. Hakkani-Tur, and L. Heck. 2010. What is left to be understood in atis? In *2010 IEEE Spoken Language Technology Workshop*, pages 19–24, Dec.
- Andrea Vanzo, Danilo Croce, Giuseppe Castellucci, Roberto Basili, and Daniele Nardi. 2016. Spoken language understanding for service robotics in italian. In Giovanni Adorni, Stefano Cagnoni, Marco Gori, and Marco Maratea, editors, *AI*IA 2016 Advances in Artificial Intelligence*, pages 477–489, Cham. Springer International Publishing.
2019. Ibm watson assistant v1. <https://cloud.ibm.com/apidocs/assistant>.

Nove Anni di jTEI: What's New?

Federico Boschetti^{1,2}

Gabriella Pardelli¹

Giulia Venturi¹

1 Istituto di Linguistica Computazionale “A. Zampolli”, CNR / Pisa

2 Digital and Public Humanities Center – Università Ca’ Foscari / Venezia

{nome.cognome}@ilc.cnr.it

Abstract

English. This paper illustrates methods and tools to study the development of research topics in the TEI community across the years. For this purpose, automatic terminology extraction technologies were exploited.

Italiano. Questo contributo illustra metodi e strumenti per studiare il cambiamento diacronico degli interessi di ricerca della comunità TEI grazie all’uso di metodi di estrazione automatica della terminologia da corpora di dominio.¹

1 Introduzione

Questo contributo nasce dall’intento di studiare con metodi di *distant reading* jTEI: il Journal of the Text Encoding Initiative (<https://journals.openedition.org/jtei>), perché è una rivista che rappresenta un ponte interessante fra la comunità delle Digital Humanities e la comunità della Linguistica Computazionale.

Come indicato da Schreibman (2011), jTEI nasce nel 2011 dopo tre anni di gestazione con l’intento di pubblicare *selected papers* dei convegni annuali (i volumi 1-2, 4, 6, 8-10) e numeri monotematici su argomenti di rilevanza per la comunità TEI (il volume 3 dedicato alla linguistica e il volume 5 dedicato alle infrastrutture). Schreibman (2014) dichiara inoltre che il volume 7, il primo frutto di una *open call*, tocca “contemporary meta concerns within the community”.

Un tassello del settore delle Digital Humanities viene rilevato in questo studio attraverso l’analisi diacronica di termini estratti dagli articoli pubblicati in jTEI dal 2011 al 2019. Lo scopo è quello

di andare a identificare termini mono- e polirematici tipici del dominio, spia dell’orientamento tematico delle attività di ricerca della comunità TEI. Oggi lo studio delle comunità sta diventando infatti centrale per comprendere e interpretare per i vari domini la direzione scientifica nonché il genere, gli stakeholder e le possibili connessioni tra comunità. Solo per fare un esempio, dalla lettura degli indici dell’estrazione del jTEI Corpus, la comunità scientifica che ruota intorno a TEI sembra non voglia usare il sostantivo *computer* e l’aggettivo *computational*, preferendo usare invece l’aggettivo *digital* combinato con una miriade di sostantivi (come ad es. *editions, humanities, text, resources, age, archive, objects, facsimile, library, tools*) in linea con gli usi della più ampia comunità delle Digital Humanities, ma non della Linguistica Computazionale.

2 Background

Questo contributo prosegue sulla linea degli studi dedicati a riviste e comunità con interessi interdisciplinari di informatica e discipline linguistiche, storico-filologiche o letterarie. In particolare, per lo studio dell’evoluzione terminologica nelle Scienze Umane e Sociali si veda Tuzzi (2018); per lo studio delle comunità della Linguistica Computazionale e delle Digital Humanities si veda Sprugnoli et al. (2019) e Pardelli et al. (2019); per lo studio della comunità della Tecnologia della Lingua e delle Risorse Linguistiche si vedano Mariani et al. (2014), Francopoulo et al. (2016), Soria et al. (2014), Bartolini et al. (2018) e Del gratta et al. (2018); per lo studio della comunità internazionale di Grey Literature si veda Pardelli et al. (2017).

Le soluzioni sin ad oggi messe a punto nell’ambito dell’estrazione automatica di terminologia da corpora di dominio sono molteplici e di diversa natura. Sebbene differiscano rispetto alle metriche utilizzate, alcuni obiettivi sono condivisi e riguar-

¹Copyright ©2019 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

dano principalmente due aspetti legati alla difficoltà di definire strategie per: *i*) risolvere il problema legato al fatto che il confine tra terminologia di dominio e lingua comune non sempre è così netto (Cabr , 1999) e *ii*) delineare dei criteri comuni nella definizione di unit  terminologica polirematica (Ramisch, 2015), dal momento che esse rappresentano pi  della met  del vocabolario di un madre-lingua (Jackendoff, 1997). La metodologia proposta in questo contributo suggerisce una strategia per superare entrambi tali aspetti problematici. Come descritto in Bonin et al. (2010), la soluzione proposta si basa su di una originale combinazione di filtri linguistici e statistici che permettono di *i*) discriminare la terminologia di dominio dalla lingua comune impiegando metriche statistiche che pesano la rilevanza dei termini estratti all'interno del corpus di acquisizione (corpus di dominio) rispetto ad un corpus di riferimento (corpus rappresentativo della lingua comune, tipicamente una collezione di articoli di giornale); *ii*) estrarre unit  polirematiche anche nei casi in cui la corrispondente testa lessicale non sia stata precedentemente individuata come unit  monorematica specifica del dominio. L'intuizione   di considerarle come elementi 'unici' costituiti da sequenze di categorie morfosintattiche (vedi Sezione 3.2). Ci  permette di suggerire una risposta all'osservazione che "non sempre la settorialit  di un LC [lessema complesso]   connessa con l'esistenza di accezioni speciali dei membri componenti, ma pu  derivare dal fatto che il LC assume in determinati contesti un significato globale speciale" (De Mauro and Voghera, 1996).

3 Metodo

3.1 Descrizione e preparazione del corpus

Gli articoli della rivista sono reperibili online sia in .pdf che in .xhtml e, per i numeri pi  recenti, anche in .xml (TEI-XML). Il corpus su cui si basa la nostra indagine parte dall'estrazione del *plain text* dall'XHTML, una volta escluso il contenuto metatestuale e paratestuale. La Tabella 1 mostra la composizione del corpus.

3.2 Estrazione terminologica

Per studiare la variazione terminologica avvenuta nel corso degli anni di pubblicazione della rivista abbiamo adottato due metodi complementari: il primo basato sull'indicizzazione del corpus tramite la terminologia estratta in modo non supervi-

Volume	#Articoli	#Parole	Lungh. media
1	6	21,480	4,198 parole
2	8	26,469	3,308 parole
3	7	38,327	5,475 parole
4	8	29,431	3,678 parole
5	7	24,921	3,560 parole
6	6	21,681	3,613 parole
7	5	26,528	5,305 parole
8	16	70,025	4,376 parole
9	6	23,897	3,982 parole
10	6	31,992	5,332 parole
TOT.	75	314,751	

Tabella 1: Composizione del corpus e lunghezza media degli articoli.

sionato e il secondo basato sull'indicizzazione dello stesso corpus tramite parole chiave fornite dagli autori come metadati degli articoli.

Il processo di estrazione terminologica non supervisionata   stato realizzato grazie a *Text-to-Knowledge (T2K)* (Dell'Orletta et al., 2014), piattaforma di estrazione e organizzazione della conoscenza da corpora multilingui di dominio basata su tecnologie di Natural Language Processing sviluppata da ILC-CNR e ampiamente validata in diversi contesti applicativi². T2K, costruito su di un'originale combinazione di sistemi a regole e algoritmi basati su metodi di apprendimento automatico, consente di estrarre da una collezione di testi linguisticamente annotati entit  rilevanti anche quando esse non sono presenti in una risorsa semantico-lessicale di dominio a disposizione. Ci  permette di far fronte e superare il tradizionale collo di bottiglia che si incontra in ogni compito di analisi semantica del testo, quello ci  di rendere esplicito il collegamento tra la realizzazione linguistica dell'informazione e la rappresentazione esplicita dell'informazione stessa.

Allo scopo pertanto di individuare ed estrarre elementi informativi nuovi rispetto a quelli presenti nel repertorio delle parole chiave a disposizione, il corpus   stato linguisticamente annotato a diversi livelli di analisi. A partire dal testo annotato a livello morfosintattico grazie al Parts-Of-Speech tagger descritto in Dell'Orletta (2009), sono state individuate le unit  terminologiche candidate all'estrazione. La metodologia, descritta in Bonin et al. (2010), consente di individuare potenziali unit  monorematiche e polirematiche impiegando una combinazione di filtri linguistici e statistici configurabili rispetto agli

²<http://www.italianlp.it/demo/t2k-text-to-knowledge/>

obiettivi di ricerca. Allo scopo della nostra indagine, i filtri linguistici sono stati configurati in modo da individuare all'interno del corpus di acquisizione: *i*) le potenziali unità monorematiche, sulla base della categoria morfo-sintattica assegnata (tipicamente 'sostantivo'); *ii*) le potenziali unità polirematiche, sulla base di una serie di sequenze di categorie morfo-sintattiche rappresentative di diversi tipi di modificazione nominale. Ad esempio, da una sequenza come 'aggettivo+sostantivo' sono individuate polirematiche quali *critical edition*, *lexical entry*, *cultural heritage*; da una sequenza 'sostantivo+sostantivo' sono individuati potenziali termini quali *TEI standard*, *manuscript material*, *knowledge representation*; per arrivare a sequenze più complesse come 'sostantivo+preposizione+sostantivo' sulla base della quale sono stati individuati termini quali *string of text*, *editions of letters* o sequenze 'sostantivo+preposizione+aggettivo+sostantivo' che permette di rintracciare un termine come *DTABf for printed texts*, *evaluation of digital scholarship* o 'aggettivo+aggettivo+sostantivo' realizzazione linguistica di un termine come *historical financial records*. I filtri statistici, applicati alla lista di termini candidati all'estrazione, consentono di ordinare tali termini sulla base della loro rilevanza all'interno del corpus di acquisizione, attribuendo loro un valore di significatività stabilita sulla base del C-NC Value (Frantzi and Ananiadou, 1999), una delle misure più utilizzate nei sistemi di estrazione terminologica.

In linea con gli obiettivi di ricerca del nostro studio, i termini così estratti sono stati impiegati dal modulo di indicizzazione di T2K per rintracciare all'interno dell'intera collezione di articoli del *JTEI* i singoli contesti nei quali i termini compaiono. Grazie a questo processo è stato possibile condurre l'indagine sulla variazione diacronica dei termini nelle diverse annate della rivista, consentendo di studiare l'evoluzione di tendenze di ricerca e tematiche di studio.

3.3 Trattamento delle parole chiave

Sono state prese in considerazione le parole chiave che gli autori stessi hanno indicato fra i metadati. Sul totale degli articoli raccolti le parole chiave distinte sono 259.

3.4 Mann-Kendall Trend Test

Per esplorare le variazioni significative d'impiego dei termini e delle parole chiave nell'in-

tervallo temporale osservato, è stato scelto il Mann-Kendall trend test, disponibile nel pacchetto trend di R (<https://bit.ly/30bWRkd>). Considerando il numero esiguo di dati disponibili per ciascun termine (o parola chiave) si è scelta quindi una statistica non parametrica sufficientemente affidabile anche con un numero di misurazioni inferiori a dieci. Per motivi di omogeneità dei dati, sono stati presi in considerazione soltanto i sette numeri della rivista riguardanti atti di convegni presi in successione cronologica, come si può vedere nelle Figure 3 e 4. I dati su cui si è applicato l'MK Test sono stati preparati in formato tabellare sia per i termini estratti automaticamente, sia per le parole chiave indicate dagli autori, disponendo su ciascuna riga un termine (o una parola chiave), su ciascuna colonna un numero della rivista e in ciascuna cella la relativa frequenza percentuale. L'MK Test fornisce un valore positivo per trend crescenti e un valore negativo per trend decrescenti. Per lo studio dei risultati sono stati presi in considerazione soltanto gli esiti con $p\text{-value} < 0.05$.

4 Risultati

4.1 Studio dei profili degli autori

Dall'analisi dei trend terminologici i numeri della rivista non dedicati ad atti dei convegni TEI (3, 5 e 7) sono stati esclusi anche perché i profili degli autori stessi hanno carattere di eccezione. Per il monitoraggio, gli autori sono stati classificati in base alla loro presenza o meno in riviste o atti di convegno di Linguistica Computazionale (con contributi o con menzioni in bibliografia). Come si può vedere in Fig. 1, il numero dedicato a TEI e linguistica (3) e il numero aperto (7) hanno attratto un numero elevato di linguisti computazionali. Sorprendentemente invece il numero dedicato alle infrastrutture TEI (5) non ha avuto la stessa attrattiva.

4.2 Dati relativi ai termini estratti

I risultati discussi in quanto segue fanno riferimento ai primi 500 termini circa mono- e polirematici estratti, con una frequenza di occorrenza ≥ 3 . La Tabella 2 riporta un estratto della lista dei primi 25 termini estratti dall'intero corpus, ordinati per rilevanza statistica e accompagnati dalla frequenza assoluta nel corpus. Per ogni termine, T2K permette di estrarre il lemma e la forma prototipica, cioè la variante linguistica più frequente del

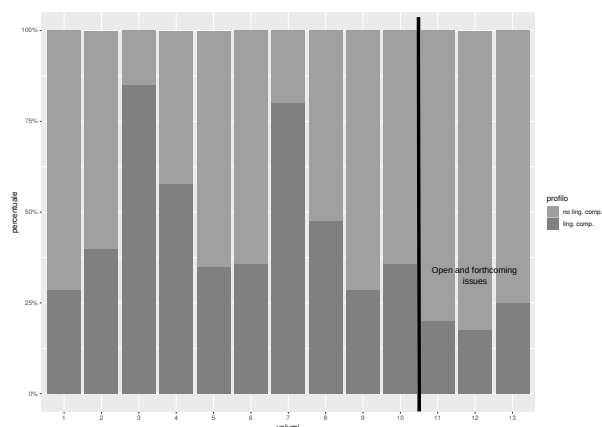


Figura 1: Autori che non hanno pubblicazioni in ambito di linguistica computazionale (no lc) e autori che ne hanno (lc)

lemma all'interno della collezione documentale di partenza.

Come introdotto nella Sezione 3.2, la fase di indicizzazione ha permesso di calcolare la distribuzione dei termini all'interno dei singoli articoli mettendo in evidenza eventuali differenze nell'uso di uno stesso termine. La Figura 2 mostra ad esempio come, sul totale di occorrenze di parole polirematiche estratte che contengono l'aggettivo *digital*, ogni volume sia caratterizzato da distribuzioni percentuali diverse. Alcuni termini possono considerarsi poco specifici come *digital age*, *digital form*, *digital resources*, *digital tools*, *digital projects*, *digital medium*. Non pochi termini risultano essere tuttavia puntuali e peculiari del settore, tra questi sono stati estratti nell'arco temporale *digital archive*, *digital critical editions*, *digital document*, *digital editions*, *digital Humanities*, *digital images*, *digital library*, *digital objects*, *digital scholarship*, *digital text*. Il grafico permette di leggere la modulazione diacronica dei termini introdotti dagli autori e riconoscibili nel settore delle Digital Humanities. Ad esempio, possiamo notare come il termine *Digital Humanities* è il termine che ha un significato più ampio e accoglie gli altri termini peculiari. Esso è pertanto sempre presente nei dieci volumi anche se la frequenza di occorrenza risulta essere altalenante. Un momento di prosperità di questo termine risulta circoscritto al volume 6 del 2013.

4.3 Distribuzione delle parole chiave nel testo

Abbiamo verificato la distribuzione delle parole chiave nel corpo degli articoli e ciò ci ha permes-

Forma prototipica	Lemma	Frequenza
TEI	TEI	2597
text	text	1261
element	element	934
project	project	485
user	user	455
document	document	421
manuscript	manuscript	396
XML	XML	393
annotation	annotation	292
TEI Guidelines	TEI Guidelines	166
edition	edition	253
tools	tool	249
information	information	248
content	content	224
language	language	221
object	object	219
source	source	214
TEI P5	TEI P5	132
TEI Consortium	TEI consortium	98
TEI documents	TEI document	91
digital editions	digital edition	89
TEI XML	TEI XML	85
TEI community	TEI community	71
manuscript description	manuscript description	54
digital humanities	digital humanity	53

Tabella 2: I primi 25 termini estratti dall'intero corpus.

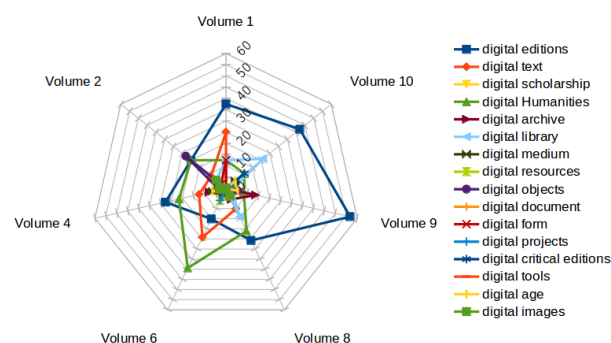


Figura 2: Distribuzione percentuale di termini polirematici estratti che contengono l'aggettivo *digital*.

so di individuare, fra le complessive 259, 32 parole chiave usate esclusivamente come metadati, e quindi che non occorrono mai nel testo, come ad esempio *bibliographical standards*, *collaborative workflow*, *TEI corpora* e 227 impiegate invece anche all'interno del testo (ad esempio *forums*).

Un'asimmetria degna di nota riguarda le sequenze aggettivo+sostantivo *critical edition* e *scholarly edition* (entrambe parole chiave) in composizione con *digital*. Mentre infatti gli autori hanno indicato nei metadati degli articoli *digital scholarly edition* come parola chiave autonoma, hanno tra-

lasciato invece *digital critical edition*, benché sia termine polirematico estratto da T2K e in alcuni articoli cooccorra *digital scholarly edition*.

4.4 Risultati dell'MK Test

Lo studio delle variazioni d'impiego dei termini al fine di identificare delle tendenze significative ha prodotto i seguenti risultati con trend crescente: *different types*, *@corresp attribute*, *open data*, *TEI Correspondence SIG*, *research questions*, *work in progress*, *Berlin-Brandenburg Academy of Sciences*, *bibliographic references*, *TEI model*, *TEI Simple*, *case study*, *TEI XML*; e i seguenti risultati con trend decrescente: *author's note*, *literary texts*, *manuscript material*, *TEI users*, *humanities research*, *TEI-encoded documents*.

Se si escludono termini isolati oppure legati a tecnologie specifiche o a particolari gruppi di ricerca, i dati sembrano far emergere una tendenza interessante. Come si può vedere in Fig. 3, aumenta l'impiego di termini condivisi con le altre scienze con basi sperimentali, fra cui le scienze del linguaggio di cui la Linguistica Computazionale fa parte, come *research questions*, *case study* e *open data*, mentre diminuisce l'impiego di termini specifici delle discipline umanistiche, come *literary texts*, *manuscript material* e *humanities research*.

Infine, lo studio delle variazioni d'impiego significative delle parole chiave indicate come metadati dagli autori stessi (Fig. 4) mostra il crescente interesse verso il web semantico (*sense* è largamente impiegato in contesti relativi alla codifica di ontologie) e verso progetti volti a rendere TEI maggiormente usabile come *TEI Simple* (<https://tei-c.org/2014/09/10/tei-simple>). Scende invece drasticamente l'impiego di parole chiave che esprimono tecnologie o concetti ormai assodati e condivisi, come *Unicode* e *community*, parola quest'ultima comprensibilmente dominante nel primo numero della rivista.

5 Conclusione

Recuperare un campione del trend delle attività di ricerca di un particolare settore scientifico, come quelle delle Digital Humanities attraverso il jTEI, può essere stimolante per comprendere gli ambiti indagati dai vari autori nell'arco temporale di dieci anni. In particolare la disponibilità di catturare oggi, articoli open access crea opportunità per l'analisi di comunità scientifiche che nel pas-

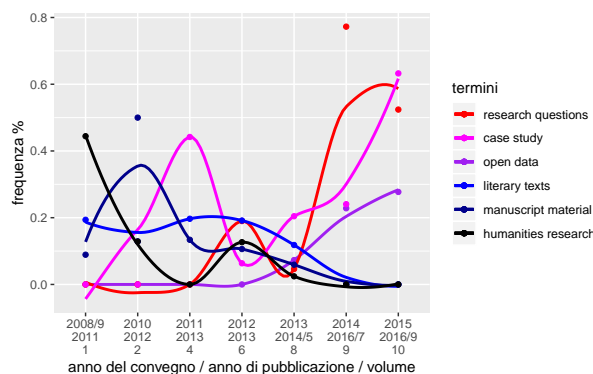


Figura 3: Trends dei termini

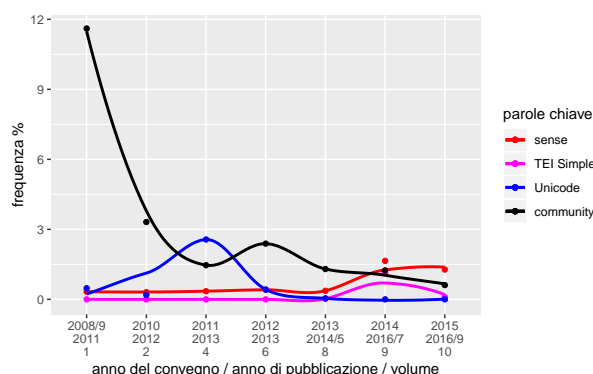


Figura 4: Trends delle parole chiave

sato non era concepibile. Il lavoro svolto rappresenta una prima esperienza di recupero informativo e di analisi per studiare il trend della comunità scientifica delle Digital Humanities attraverso una rivista ad essa dedicata, il jTEI. Pensiamo altresì che sia fondamentale ampliare le nostre fonti con altre tipologie di riferimento: come blog, forum, atti di conferenze nazionali e internazionali e riviste. Nell'analisi uno sguardo sarà rivolto anche agli autori per comprendere connessioni e estraneità tra la linguistica computazionale e le Digital Humanities.

References

- R. Bartolini, S. Goggi, M. Monachini and G. Paredelli. 2018. The LREC Workshops Map. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 201)*, ELRA, Paris, pp. 557-562. <https://aclweb.org/anthology/papers/L/L18/L18-1088/>
- F. Bonin, F. Dell'Orletta, S. Montemagni and G. Venturi. 2010. A Contrastive Approach to Multi-word Extraction from Domain-specific Corpora. *Proceedings of 7th Edition of International Conference on*

- Language Resources and Evaluation (LREC 2010)*, 17-23 May, Valletta, Malta.
- M. T. Cabré. 1999. The terminology. Theory, methods and applications. John Benjamins Publishing Company.
- R. Del Gratta, S. Goggi, G. Pardelli and N. Calzolari. 2018. LREMap, a Song of Resources and Evaluation. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. ELRA, Paris, pp. 1275-1281. <https://www.aclweb.org/anthology/L18-1203>
- F. Dell'Orletta. 2009. Ensemble system for Part-of-Speech tagging. *Proceedings of Evalita'09, Evaluation of NLP and Speech Tools for Italian*, Reggio Emilia, December.
- F. Dell'Orletta, G. Venturi, A. Cimino and S. Montemagni. 2014. T2K: a System for Automatically Extracting and Organizing Knowledge from Texts. *Proceedings of 9th Edition of International Conference on Language Resources and Evaluation (LREC 2014)*, 26-31 May, Reykjavik, Iceland.
- T. De Mauro and M. Voghera. 1996. Scala mobile. Un punto di vista sui lessemi complessi. P. Benincà et al. (eds.), *Italiano e dialetti nel tempo. Saggi di grammatica per Giulio C. Lepschy*, Roma, Bulzoni, pp. 99-131.
- G. Francopoulo, J. Mariani and P. Paroubek. 2016. A Study of Reuse and Plagiarism in LREC papers. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, ELRA, Paris, pp. 1890-1897. <https://www.aclweb.org/anthology/L16-1298>
- K. Frantzi and S. Ananiadou. 1999. The C-value / NC Value domain independent method for multiword term extraction. *Journal of Natural Language Processing*, 6(3):145-179.
- R. Jackendoff. 1997. Twistin' the night away. *Language*, 73, pp. 534-559.
- J. Mariani, P. Paroubek, G. Francopoulo and O. Hamon. 2014. Rediscovering 15 Years of Discoveries in Language Resources and Evaluation: The LREC Anthology Analysis. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)*, ELRA, Paris, pp. 4632-4669. http://www.lrec-conf.org/proceedings/lrec2014/pdf/1228_Paper.pdf
- G. Pardelli, S. Goggi and F. Boschetti. 2019. Strolling around the dawn of Digital Humanities. *Book of Abstract for the 8th Annual Conference AIUCD 2019*, pp. 261-264.
- G. Pardelli, S. Goggi, R. Bartolini, I. Russo and M. Monachini. 2017. A Geographical Visualization of GL Communities: A Snapshot. In *Eighteenth International Conference on Grey Literature: Leveraging Diversity in Grey Literature*, Washington, November 28-29, 2016. Edited by Dominic Farace and Jerry Frantzen, TransAtlantic-Amsterdam, 18, pp. 109-113.
- T. Pohlert. 2018. *Non-Parametric Trend Tests and Change-Point Detection*, CRAN. <https://bit.ly/30bWRkd>,
- C. Ramisch. 2015. Multiword expressions acquisition: A generic and open framework. New York: Springer.
- S. Schreibman. 2011. Editorial Introduction to the First Issue. *Journal of the Text Encoding Initiative*, 1. <http://journals.openedition.org/jtei/229>
- S. Schreibman. 2014. Editorial Introduction to Issue 7 of the Journal of the Text Encoding Initiative. *Journal of the Text Encoding Initiative*, 7. <http://journals.openedition.org/jtei/1046>
- C. Soria, N. Calzolari, M. Monachini, V. Quochi, N. Bel, K. Choukri, M. Mariani, J. Odiijk and S. Piperidis. 2014. The language resource Strategic Agenda: the FLReNet synthesis of community recommendations. *Language Resources and Evaluation*, December 2014, 48 (4), pp. 753-775. <https://link.springer.com/article/10.1007/s10579-014-9279-y>
- R. Sprugnoli, G. Pardelli, F. Boschetti and R. Del Gratta. 2019. Un'Analisi Multidimensionale della Ricerca Italiana nel Campo delle Digital Humanities e della Linguistica Computazionale. *Umanistica Digitale*, ISSN 2532-8816, 5, pp. 59-89. <https://umanisticadigitale.unibo.it/article/view/8581>
- A. Tuzzi. 2018. Tracing the Life Cycle of Ideas in the Humanities and Social Sciences. New York: Springer.

BullyFrame: Cyberbullying Meets FrameNet

Silvia Brambilla[‡], Alessio Palmero Aprosio[†], Stefano Menini[†]

[†]FBK (Trento), [‡]University of Bologna

silvia.brambilla2@unibo.it

{aprosio,menini}@fbk.eu

Abstract

English. This paper presents BullyFrame, a dataset of cyberbullying interactions collected from WhatsApp conversations in Italian and annotated with FrameNet semantic frames. We will describe the creation of the dataset discussing the problematic aspects found in the annotation process, such as the lack of coverage of FrameNet for the annotation of texts extracted from social media. Finally, we present a preliminary study that describes the relations between the frames and the cyberbullying-related annotation of the original dataset.¹

Italiano. *Questo studio presenta BullyFrame, un dataset di conversazioni WhatsApp in italiano contenenti episodi di cyberbullismo e annotate secondo i frame semantici di FrameNet. Verrà descritta la creazione del dataset discutendo gli aspetti problematici incontrati nel processo di annotazione, come ad esempio i limiti di copertura di FrameNet per l'annotazione di testi estratti da social media. Infine, presentiamo uno studio preliminare che descrive le relazioni tra l'annotazione di FrameNet e quella del dataset originale, relativa al cyberbullismo.*

1 Introduction

The semantic analysis of a text involves the classification of predicates into a set of events, for which it is important to determine who did what, when and where. For example, in the sentence “In 1912, the Titanic hit an iceberg on its first trip across the

Atlantic”, the verb “hit” represents the event, “Titanic” is the main actor of that event, “1912” and “Atlantic” indicate when and where it took place, and so on. The process of extracting the semantic roles and relations in a sentence is called Semantic Role Labeling (SRL), and, in the last years, both resources listing possible events and corpora have been annotated with this kind of information. Examples of such datasets are FrameNet (Ruppenhofer et al., 2006) and PropBank (Palmer et al., 2005). Given the availability of these resources, over the years SRL has gained more attention and has become an important task in computational linguistics, with a growing number of works and evaluations (QasemiZadeh et al., 2019; Basili et al., 2012).

Unfortunately, the vast majority of annotated datasets relies mainly on newswire and narrative texts, and their coverage turns out to be inadequate when it comes to annotate more specific domains, such as, for instance, football domain (Torrent et al., 2014) or medicine domain (Tan et al., 2011).

Aside from that, over the last decades, ICT technologies and communication habits underwent profound changes, with the greatest part of text production in the world coming from social networks and being usually written in non-standard language.² This kind of communication is of fundamental importance, in particular for teenagers’ social life. For instance, according to the last report by the Italian Statistical Institute (ISTAT, 2014) in Italy 82.6 of children aged 11-17 use the mobile phone every day. The use of these new technologies, however, leads also to some undesirable side effects, as the proliferation of hate speech and the digitization of traditional forms of harassment, also known as cyberbullying.

Many studies (O’Moore and Kirkham, 2001; Fekkes et al., 2006; Farag et al., 2019) have high-

¹Copyright © 2019 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

²<https://www.domo.com/learn/data-never-sleeps-6>

lighted that cyberbullying can have a negative impact on the victims' psychological and emotional well-being and that, in extreme cases, it can lead to self-harm and suicidal thoughts. For this reason, some strategies have been implemented to detect and contrast this phenomenon (Van Hee et al., 2018; Zhao et al., 2016; Menini et al., 2019), but none of them makes use of SRL, and no resources on this topic based on frame semantics have been developed yet. We therefore developed BullyFrame, a dataset annotated with frame semantic annotation, where the messages are taken from a corpus of data on cyberbullying interaction in Italian, gathered through a WhatsApp experimentation with lower secondary school students (Sprugnoli et al., 2018). Our work leads to the release of the annotated corpus (see Section 3), and constitutes a feasibility study, that investigates the potential lacks of FrameNet - resource that does not claim to be exhaustive in its coverage - for the annotation of online chats that, in addition to their non-standard nature, contain offensive language and informal expressions. We show, for instance, that some frames are completely missing, such as those regarding sexual orientation, as discussed in Section 4. In other cases, FrameNet provides frames whose purpose is similar to the needed one, but cannot fit perfectly the meaning of the sentence. For example, the frame "Offenses" refers to acts that violate a legal code, but it is not used for marking offenses (or bad words) between two users, e.g. "idiotia" ("idiot" - currently tagged as `Mental_property`), "stronzetta" ("asshole" - left currently with no annotation). Similarly, a sentence like "Ti ricordo che io ho ballato con Kledi" ("I remind you that I danced with Kledi") cannot be correctly annotated, as neither `Evoking`, nor `Reminder` or `Remembering_*` frames are able to capture the meaning of someone who reminds something to another person.

In Section 5, we also provide a comparison study to highlight relations between the newly-released frame annotation and the existing one regarding the type of cyberbullying expression. Results show that some of them are strictly connected (even when it is not immediate to understand).

Finally, in Section 6 we present Framy, a frame annotation tool that works as a web server and that has been used for annotating BullyFrame.

2 Related Work

The work presented in this paper spans topics from different research areas. As for the methodology, we deal with issues related to the annotation of Italian texts with FrameNet and frame annotation on social media texts. Then, as case study, we focus on the cyberbullying domain, where we witness a growing interest and a large number of novel works over the last few years.

The FrameNet database is a resource originally developed for the English language that has proven to be largely portable over different languages. This because its frames appear to be mostly language independent, as pointed out by Gilardi and Baker (2018). Nevertheless, some language specific differences can arise both at the level of frames themselves (coarse-grained level) and at the level of frame elements (FEs) (fine-grained level) (Lönneker-Rodman, 2007). As an example it is possible to recall the works of Candido et al. (2014) on French, of Ohara (2012) on Japanese and of Subirats and Sato (2004) on Spanish. In all the three languages the creation of a FrameNet-like resource required to add new frames or FEs or modify already existing ones, for instance in French some frames needed to be merged, while others needed to be split into two subframes.

For the Italian language, we rely on previous researches, carried out at the Universities of Bologna and Roma Tor Vergata (Basili et al., 2017; Vanzo et al., 2017), Fondazione Bruno Kessler in Trento (Tonelli et al., 2009; Tonelli and Pianta, 2009; Tonelli, 2010) and Pisa (Johnson and Lenci, 2011), that investigated the creation of an Italian FrameNet and first annotated Italian texts with frames.

Gerrard et al. (2017) outline how frame annotation of texts extracted from social media could be challenging because of the differences between social media data and the kind of data on which FrameNet is built, i.e. edited and well-formed sentences. For this reason as for today only few studies annotated social media texts with frame information (Kim and Hovy, 2006; Gerrard et al., 2017; ElSherief et al., 2018) even if it proved to be useful for example in identifying opinions with their holder and topic (Kim and Hovy, 2006) or in deepening the analysis of Directed and Generalized hate speech (ElSherief et al., 2018).

Works on cyberbullying try to detect and pre-

vent the phenomenon exploiting different methodologies and techniques. In particular, a dataset extracting data from Facebook has been developed at University of Pisa (Del Vigna et al., 2017), while at the University of Turin a similar corpus has been created from Twitter (Sanguinetti et al., 2018). Dinakar et al. (2011) build individual topic-sensitive binary classifiers, Van Hee et al. (2018) perform classification based on n-grams and specific features as the presence of aggressive and subjective language, while Zhao et al. (2016) apply different weights to pre-defined insulting words using them as bullying features combined with bag-of-words and latent semantic features for their classifier.

As for today, at the best of our knowledge, there are not research works that studied the possible interconnections between cyberbullying and frames.

3 Dataset Description

For the annotation of the frames related to cyberbullying we use as starting point the dataset from Sprugnoli et al. (2018). The dataset presents a collection of WhatsApp chats written by 12-13 years old students simulating instances of cyberbullying in specific scenarios.

The text of the chats is provided with annotations about *i*) the *role* of who is writing (i.e. Victim, Bully, or supporter of one of the two sides) and *ii*) labels with the *type* of offense that can be found on each message (in particular, the labels include: Threat or blackmail, General Insult, Body Shame, Sexism, Racism, Curse or Exclusion, Insult Attacking Relatives, Harmless Sexual Talk, Defamation, Sexual Harassment, Defense, Encouragement to the Harassment, and Other).

The dataset consists of 10 chats, for a total of 2192 messages (14,600 tokens) and includes 1,203 cyberbullying expressions, corresponding to 6,000 tokens.

Starting from this, we fully annotated the sentences referring to FrameNet 1.7: the resulting annotation is available for download from the resource website.³ It is released under the Creative Commons Attribution-ShareAlike 4.0 International license.⁴

A total of 2,458 frames and 2,769 frame element have been annotated on 1,558 sentences. The remaining 1,211 sentences cannot be annotated,

mainly because no corresponding frames can be found (1,180 sentences), or because there was a picture instead (19 sentences), or finally because the messages have been deleted by the user (12 sentences). Table 1 (a) shows statistics on how many frames have been annotated for each sentence. Regarding the coverage, a total of 268 unique frames and 696 unique frame elements have been found in the dataset. Table 1 (b) shows the most frequent frames that have been annotated. Finally, Table 1 (c) shows statistics on how many frame elements are annotated for each frame.

4 Frame Annotation

In order to investigate possible connections between frames and cyberbullying we annotated all the sentences of the dataset with frames and frame elements referring to the 1.7 version of FrameNet. In each sentence we tried to annotate all the possible evoked frames alongside with their frame elements.

When annotating the sentences we have to face some problems that, due to the nature of this dataset, to the differences between English and Italian, and to the nature of FrameNet itself, is not complete but that is constantly updated and enlarged.

Problematic aspects can be found on three different levels: Frames layer, Frame Elements layer and Frame Evoking Elements layer.

Frames layer: We found that some of the concepts that were evoked by lexical units (LUs) were not present in FrameNet. The missing frames could be:

- a) Concepts that are new to FrameNet and that are linked to the particular nature of the text. This is the case for instance of frames that occur often in conversations or in oral communication. These concepts are often not present in FrameNet, but frequent in our dataset since it includes interactions between participants and is close to oral communication. For example we found that FrameNet does not have a frame that covers “greetings”, evoked in sentences such as:

“Ciao ci sentiamo domani” (Bye, we’ll talk tomorrow)

“Hahahah esatto ciao e buon allenamento” (Hahahah, exactly bye and have a good training)

³<https://github.com/dhfbk/bullyframe>

⁴<https://creativecommons.org/licenses/by-sa/4.0/>

Frames	Sentences	Frequency	Frame	Frame elements	Frames
8	2	167	Silencing	4	7
7	2	138	Desirability	3	118
6	7	109	Statement	2	633
5	8	108	Correctness	1	1121
4	46	107	Cause_emotion	0	332
3	132	97	Desiring		
2	406	87	Awareness		
1	955	83	Opinion		
0	603	73	Capability		
Pic	19	69	Intentionally_act		
Del	12				

Table 1: These three tables show: (a) the number of sentences with the corresponding amount of frame found in them; (b) the frequencies of the top 10 frames; (c) the frequencies of frame elements for each frame annotation.

“Buongiorno a tutti!” (Have a good day, everybody!)

“Sì e tu vai a giocare a rugby” (Yes, and you go play rugby)

- b) Concepts that are new to FrameNet and that are linked to abusive language and cyberbullying. For example we found that bullies often refer to people’s sexual orientation as an insult such as in:

“Crede di essere figo facendo il gay a danza” (He thinks he looks cool acting like a gay when he dances)

“Manco fossi gay 🤔🤔🤔” (What am I, gay? 🤔🤔🤔)

“Sei così effeminato che intorno a te ci sono più finocchi che in un orto” (You are so effeminate that around you there are more pansies than in a garden)

However, a frame that covers this concept is missing in FrameNet.

- c) Concepts that are new to FrameNet, but that are not specifically linked to the nature of the text nor to abusive language or cyberbullying. For example in FrameNet are missing frames related with “sports” and similar activities:

“Anche tu fai calcio” (You play football as well)

“Lui non fa danza classica” (He does not do ballet)

- d) Concepts that are not new to FrameNet corresponding to holes in the FrameNet hierarchy. For example FrameNet has a frame for *Silencing*, a frame for *Becoming_silent* but it does not have a frame for *Being_silent*.

Frame Elements layer: We found that not only frames were missing but that it was also possible to find missing FEs.

For example it appears to be missing the FE *Reason* for the frame *Statement*, useful for annotating sentences such as:

“Lo diciamo per il tuo bene” (We say that for your own sake)

here *“Per il tuo bene” (For your own sake)* expresses the motivation for which the speaker makes his statement and could be labeled as *Reason*.

Another example can be the frame *Ingestion* for which a FE *Quantity*, for annotating the quantity of the ingestibles eaten, appears to be missing. For example, in the sentence:

“Non mangiare tanto o diventi ancora più obeso” (Do not eat a lot or you will get even fatter)

the FE label `Quantity` would be perfectly fitting for annotating the adverb *“tanto”* (a lot).

Frame-Evoking Elements layer: Problems linked to the fact that in the sentences we tagged we find that not only words or multiword expressions (MWEs) evoke frames but that also other elements. In particular we found that frames can be evoked also by:

- a) **Constructions:** For example in the sentences *“Di sicuro un cane è più bravo di lui”* (A dog is better than him for sure) or *“Noi siamo più forti di te”* (We are stronger than you) the frame `Surpassing` is evoked by the construction *“essere più X di Y”* (To be Xer than Y) rather than by a word or a multiword expression.

- b) **Emoji:** For example, in the sentence

“Ma tu sei già una 🐞” (But you are already a 🐞)

the *“Pile of Poo”* emoji evokes the frame `Desirability`.

Aside from these three problematic layers, we found that for a considerable amount of messages it was not possible to add any frame annotation because of problems of different nature. More specifically we found that:

- a) Some messages are only made of punctuation marks, mostly ellipsis, exclamation points and question marks.
- b) Some messages are made of interjections or discourse markers and it is, thus, not possible to identify any frame evoking element:

“Ooooooooooooooooooooo 🤔”

“Ahahahahahahahahahahahahah”

- c) In some other cases there are sentences that have been split into two or more messages. In these cases it is often possible to find messages in which no frame is evoked, but that constitute a FE of a frame evoked in the bigger sentence that has been split.

For example, the sentence:

“Ma noi verremmo con i nostri bei cori” (But we would come with our nice chant)

has been split into two different messages *“Ma noi verremmo”* (But we would come) and *“Con i nostri bei cori”* (With our nice chants). The first message can be annotated with the frame `Arriving` while the second message could only be annotated as the `Arriving` frame element `Depictive`.

The sentence:

“Neanche hai capito che è una citazione di Battiato ” (You didn’t even understand that this is a quote from Battiato)

have been split into *“Neanche hai capito che è una citazione”* (You didn’t even understand that it is a quote) and *“Di Battiato”* (From Battiato). In the first message, the LU *“capire.v”* (*understand.v*) evokes the frame `Awareness`, and *“Che è una citazione”* (That it is a quote) instantiates its frame element `Content`, whereas the second message can only be considered as a part of it.

- d) Some messages contain only affirmative and negative expressions, i.e. *“Yes”* or *“No”*.
- e) Other messages only repeat a word or a group of words of the previous message or anticipate one word or a group of words that will be part of the subsequent message:

“Tu”, “Tu che sei un maschio”
(You, You that are a boy)

- f) Finally there are messages that only aim to correct a word or a letter previously misspelled:

*“Ai scritto”, “*Hai”* (You wrote)

*“Bravo Bul”, “*Bullo”* (Good bully)

A field that is particularly relevant is the semantic field of emotions. We found that FrameNet frames referring to this field have sometimes fuzzy boundaries and that it is sometimes hard to choose a frame over another. Moreover there are also some frames that seem to be missing: for example in FrameNet there is no frame that covers the concept of *“Expressing emotions”*

evoked by LUs such as “weep.v” or “cry.v” or “laugh.v”. Indeed, the first is completely missing in FN, the second is present as evoking Make_noise, Communication_noise and Vocalization, the third in present only as evoking Make_noise.

5 Annotations comparison

In order to highlight significant relations between frames and cyberbullying, we compared the frame annotation with the already existing annotation regarding the type of cyberbullying expression (see Section 3). In particular we computed their correlation using the weighted mutual information. This kind of evaluation can be useful, for instance, to predict cyberbullying conversations using tools that automatically extract semantic information with respect to frames, such as SEMAFOR (Das et al., 2014).

The results, reported in Table 2, show some interesting outcomes. Most of them are in line with what we could have expected, but some others instead reflect the limitations of FrameNet in the annotation of this kind of interactions. For example we can see that “General_insult” is related with frames such as Mental_property or Desirability, this well matches with the intuitions that those frames capture respectively expressions which denigrates the interlocutor by referring to his/her lower intelligence, e.g. “Idiota” or “Stupida” (“Idiot”, “Stupid”), or to his/her scarce desirability, e.g. “Sfigato” (“Loser/Lame”). The same can be said for the pairs “Treat_or_Blackmail” - Cause_harm and “Insult-BodyShame” - Aesthetics, where the connection between the frame and the cyberbullying type appears to be straightforward. Nevertheless there are also pairs if which the connection is hard to understand. For example “Encouragement to the Harasser” shows a strong relation with the frame Correctness. This is due, once again, to the limitations of FrameNet that lacks of some frames, in this particular case it lacks of a frame for the expressions that indicate a reinforcement of what one of the interlocutors just said such as “Esatto” (“Exactly”) or “Hai ragione” (“You are right”) that are now listed under the frame Correctness.

Bullying annotation	Frame	wMI
Curse_or_Exclusion	Silencing	0.0672
General_Insult	Desirability	0.0304
General_Insult	Mental_property	0.0227
Encourage_Harasser	Correctness	0.0177
Curse_or_Exclusion	Desiring	0.0135
Threat_or_Blackmail	Cause_harm	0.0127
Discrimination-Sexism	Suitability	0.0083
Curse_or_Exclusion	Required_event	0.0080
General_Insult	Silencing	0.0065
Insult-BodyShame	Aesthetics	0.0046

Table 2: Correlation between the new annotations of frames and the previous ones of cyberbullying types using weighted mutual information (wMI).

6 The annotation interface

The annotation on FrameNet has been performed using a tool called Framy, developed at Fondazione Bruno Kessler and freely available on Github⁵ under the Apache 2.0 license. It is written in php and needs a MySQL database to work.

The application is optimized for frame semantics annotation, and can be configured to work with every version of FrameNet. After loading the already tokenized text data using the included scripts, a human annotator can select both the lexical unit that evokes the frame and the frame elements relative to the selected words.

7 Conclusions and Future Work

In this paper, we present and release BullyFrame, an Italian resource consisting in a set of WhatsApp chats with full-text FrameNet annotations. The data, freely accessible on GitHub, increases the availability of resources in Italian. We also discuss how FrameNet lacks certain frames, as it cannot cover some expressions used mainly in the social media language. Finally, we describe Framy, a free tool that supports the manual annotation of texts w.r.t. FrameNet.

In the future, we want to extend this dataset by including other text resources, and extend FrameNet coverage for the social media domain, to deal with informal expressions and emojis.

Acknowledgments

This work has been supported by the European Commission project Hatemeter (REC-DISC-AG-

⁵<https://github.com/dhfbk/framy>

2016, action grants 2016: European citizenship rights, anti-discrimination, preventing and combating intolerance).

References

- Roberto Basili, Diego De Cao, Alessandro Lenci, Alessandro Moschitti, and Giulia Venturi. 2012. EvalIta 2011: The Frame Labeling over Italian Texts Task. In *International Workshop on Evaluation of Natural Language and Speech Tool for Italian*, pages 195–204. Springer.
- Roberto Basili, Silvia Brambilla, Danilo Croce, and Fabio Tamburini. 2017. Developing a large scale FrameNet for Italian: the IFrameNet experience. *CLiC-it 2017 11-12 December 2017, Rome*, page 59.
- Marie Candito, Pascal Amsili, Lucie Barque, Farah Benamara Zitoune, Gaël De Chalendar, Marianne Djemaa, Pauline Haas, Richard Huyghe, Yvette Yannick Mathieu, Philippe Muller, et al. 2014. Developing a French Framenet: Methodology and first results.
- Dipanjan Das, Desai Chen, André FT Martins, Nathan Schneider, and Noah A Smith. 2014. Frame-semantic parsing. *Computational linguistics*, 40(1):9–56.
- Fabio Del Vigna, Andrea Cimino, Felice Dell’Orletta, Marinella Petrocchi, and Maurizio Tesconi. 2017. Hate me, hate me not: Hate speech detection on Facebook.
- Karthik Dinakar, Roi Reichart, and Henry Lieberman. 2011. Modeling the detection of textual cyberbullying. In *fifth international AAAI conference on weblogs and social media*.
- Mai ElSherief, Vivek Kulkarni, Dana Nguyen, William Yang Wang, and Elizabeth Belding. 2018. Hate lingo: A target-based linguistic analysis of hate speech in social media. In *Twelfth International AAAI Conference on Web and Social Media*.
- Nadine Farag, Samir Abou El-Seoud, Gerard McKee, and Ghada Hassan. 2019. Bullying hurts: A survey on non-supervised techniques for cyber-bullying detection. In *Proceedings of the 2019 8th International Conference on Software and Information Engineering*, pages 85–90. ACM.
- Minne Fekkes, Frans IM Pijpers, A Miranda Fredriks, Ton Vogels, and S Pauline Verloove-Vanhorick. 2006. Do bullied children get ill, or do ill children get bullied? a prospective cohort study on the relationship between bullying and health-related symptoms. *Pediatrics*, 117(5):1568–1574.
- David Gerrard, Martin Sykora, and Thomas Jackson. 2017. Social media analytics in museums: extracting expressions of inspiration. *Museum Management and Curatorship*, 32(3):232–250.
- Luca Gilardi and C Baker. 2018. Learning to Align across Languages: Toward Multilingual FrameNet. In *Proceedings of the International FrameNet Workshop*, pages 13–22.
- Martina Johnson and Alessandro Lenci. 2011. Verbs of visual perception in Italian FrameNet. *Constructions and Frames*, 3(1):9–45.
- Soo-Min Kim and Eduard Hovy. 2006. Extracting opinions, opinion holders, and topics expressed in online news media text. In *Proceedings of the Workshop on Sentiment and Subjectivity in Text*, pages 1–8. Association for Computational Linguistics.
- Birte Lönneker-Rodman. 2007. Multilinguality and FrameNet. *International Computer Science Institute Technical Report*.
- Stefano Menini, Giovanni Moretti, Michele Corazza, Elena Cabrio, Sara Tonelli, and Serena Villata. 2019. A System to Monitor Cyberbullying based on Message Classification and Social Network Analysis. In *Proceedings of the third Workshop on Abusive Language Online*, Florence, Italy.
- Kyoko Ohara. 2012. Semantic Annotations in Japanese FrameNet: Comparing Frames in Japanese and English. In *LREC*, pages 1559–1562. Citeseer.
- Mona O’Moore and Colin Kirkham. 2001. Self-esteem and its relationship to bullying behaviour. *Aggressive behavior*, 27(4):269–283.
- Martha Palmer, Daniel Gildea, and Paul Kingsbury. 2005. The proposition bank: An annotated corpus of semantic roles. *Computational linguistics*, 31(1):71–106.
- Behrang QasemiZadeh, Miriam RL Petruck, Regina Stodden, Laura Kallmeyer, and Marie Candito. 2019. SemEval-2019 Task 2: Unsupervised Lexical Frame Induction. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 16–30.
- Josef Ruppenhofer, Michael Ellsworth, Myriam Schwarzer-Petruck, Christopher R Johnson, and Jan Scheffczyk. 2006. FrameNet II: Extended theory and practice.
- Manuela Sanguinetti, Fabio Poletto, Cristina Bosco, Viviana Patti, and Marco Stranisci. 2018. An Italian Twitter Corpus of Hate Speech against Immigrants. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*, Miyazaki, Japan, May. European Languages Resources Association (ELRA).
- Rachele Sprugnoli, Stefano Menini, Sara Tonelli, Filippo Oncini, and Enrico Piras. 2018. Creating a WhatsApp Dataset to Study Pre-teen Cyberbullying. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 51–59.

- Carlos Subirats and Hiroaki Sato. 2004. Spanish framenet and framesql. In *4th International Conference on Language Resources and Evaluation. Workshop on Building Lexical Resources from Semantically Annotated Corpora. Lisbon (Portugal)*. Cite-seer.
- He Tan, Rajaram Kaliyaperumal, and Nirupama Benis. 2011. Building frame-based corpus on the basis of ontological domain knowledge. In *Proceedings of BioNLP 2011 Workshop*, pages 74–82. Association for Computational Linguistics.
- Sara Tonelli and Emanuele Pianta. 2009. Three issues in cross-language frame information transfer. In *Proceedings of the International Conference RANLP-2009*, pages 441–448.
- Sara Tonelli, Daniele Pighin, Claudio Giuliano, and Emanuele Pianta. 2009. Semi-automatic development of FrameNet for Italian. In *Proceedings of the FrameNet Workshop and Masterclass, Milano, Italy*.
- Sara Tonelli. 2010. Semi-automatic techniques for extending the FrameNet lexical database to new languages.
- Tiago Torrent, Maria Margarida Salomão, Fernanda Campos, Regina Braga, Ely Matos, Maucha Gamonal, Julia Gonçalves, Bruno Souza, Daniela Gomes, and Simone Peron. 2014. Copa 2014 framenet brasil: a frame-based trilingual electronic dictionary for the football world cup. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: System Demonstrations*, pages 10–14.
- Cynthia Van Hee, Gilles Jacobs, Chris Emmery, Bart Desmet, Els Lefever, Ben Verhoeven, Guy De Pauw, Walter Daelemans, and Véronique Hoste. 2018. Automatic detection of cyberbullying in social media text. *PloS one*, 13(10):e0203794.
- Andrea Vanzo, Danilo Croce, Roberto Basili, and Daniele Nardi. 2017. Structured Learning for Context-aware Spoken Language Understanding of Robotic Commands. In *Proceedings of the First Workshop on Language Grounding for Robotics*, pages 25–34.
- Rui Zhao, Anna Zhou, and Kezhi Mao. 2016. Automatic detection of cyberbullying on social networks based on bullying features. In *Proceedings of the 17th international conference on distributed computing and networking*, page 43. ACM.

Lost in Text. A Cross-Genre Analysis of Linguistic Phenomena within Text

Chiara Buongiovanni[•], Francesco Gracci[•], Dominique Brunato[◊], Felice Dell’Orletta[◊]

[•] University of Pisa

{c.buongiovanni, f.gracci}@studenti.unipi.it

[◊]Istituto di Linguistica Computazionale “Antonio Zampolli” (ILC-CNR)

ItaliaNLP Lab - www.italianlp.it

{dominique.brunato, felice.dellorletta}@ilc.cnr.it

Abstract

Moving from the assumption that formal, rather than content features, can be used to detect differences and similarities among textual genres and registers, this paper presents a new approach to the linguistic profiling methodology, which focuses on the internal parts of a text. A case study is presented showing that it is possible to model the degree of variance within texts representative of four traditional genres and two levels of complexity for each.¹

1 Introduction

The combined use of corpus-based and computational linguistics methods to investigate language variation has become an established line of research. The heart of this research is the so-called ‘linguistic profiling’, a technique in which a large number of counts of linguistic features automatically extracted from parsed corpora are used as a text profile and can then be compared to average profiles for groups of texts (van Halteren, 2004). Although it has been originally developed for authorship verification and recognition, linguistic profiling has been successfully applied to the study of genre and register variation, following Biber’s claim that “linguistic features from all levels function together as underlying dimensions of variation, with each dimension defining a different set of linguistic relations among registers” (Biber, 1993). By modeling the ‘form’ of a text through large sets of linguistic features extracted from representative corpora, it has been possible not only to enhance automatic classification of genres (Stamatatos et al., 2001), but also to get a better un-

derstanding of the impact of features in classifying genres and text varieties (Cimino et al., 2017).

This paper moves in this framework but presents a new approach of linguistic profiling, in which the unit of analysis is not the document as a whole entity, but the internal parts in which it is articulated. In this respect, our perspective is similar to the one proposed by Crossley et al. (2011), who developed a supervised classification method based on linguistically motivated features to discriminate paragraphs with a specific rhetorical purpose within English students’ essays. However, differently from that work, we focus on Italian and enlarge the analysis to four traditional textual genres and two levels of language complexity for each. The aim is i) to explore to what extent the internal structure of a text can be modeled via linguistic features automatically extracted from texts and ii) to study whether the variance across different parts of a text changes according to genre and level of complexity within genre.

2 Corpora and approach

Our investigation was carried out on four genres: Journalism, Educational writing, Scientific prose and Narrative. For each genre, we selected the two corpora described in Brunato and Dell’Orletta (2017), which represent a ‘complex’ and a ‘simple’ language variety for that genre, where the level of complexity was established according to the expected reader. Specifically, the journalistic genre comprises a corpus of articles published between 2000 and 2005 on the general newspaper *La Repubblica* and a corpus of easy-to-read articles from *Due Parole*, a monthly magazine written in a controlled language for readers with basic literacy skills or mild intellectual disabilities (Piemontese, 1996). The corpus belonging to the Educational genre is articulated into two collections targeting high school (AduEdu) vs. primary school (ChiEdu) students. For the scientific prose,

¹Copyright ©2019 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

the ‘complex’ variety is represented by a corpus of 84 scientific articles on different topics, while the ‘simple’ one by a corpus of 293 Wikipedia articles, extracted from the Italian Portal ‘Ecology and Environment’. For the Narrative genre, we took a dataset specifically developed for research on automatic text simplification. It consists of 56 texts covering short novels for children and pieces of narrative writing for high school L2 students arranged in a parallel fashion, i.e. for each original text a manually simplified version is available. For our study, the original texts and the corresponding simplified versions were chosen as representative of the complex variety and the simple variety, respectively.

All corpora were automatically tagged by the part-of-speech tagger described in Dell’Orletta (2009) and dependency parsed by the DeSR parser (Attardi et al., 2009) to allow the extraction of more than 80 linguistic features, on which we relied to investigate our research questions. These features (detailed in Section 3) capture linguistic phenomena of a different nature, with a focus on morpho-syntactic and syntactic structure, and were selected since they were proven effective for genre classification in previous works, as well as in other scenarios all focused on the analysis of the ‘form’ of the text rather than its content, such as linguistic complexity, readability assessment (Collins-Thompson, 2014), native language identification (Malmasi et al., 2017).

As a preliminary step for the analyses, all documents were split into a fixed number of sections, where each section is composed by a certain number of paragraphs, roughly corresponding to the three main parts of the rhetorical structure of a text (i.e. introductory, body and concluding paragraphs). According to the literature, for some genres, such as academic writing, the distinction into paragraphs is quite rigid and follows the so-called ‘five-paragraphs’ format (Crossley et al., 2011) which adheres to the rhetorical goals of the document, i.e. the first and the last paragraph correspond respectively to the introduction and the conclusion, and the three middle ones to the body part. However, based on a preliminary investigation of our corpora we preferred to define a six-section subdivision in order to avoid flattening too much the distinctions across genres. The corpora under analysis indeed are made by documents which are very different in terms of average

length: for instance, scientific articles are on average longer than others (184 sentences per document) and this reflects the fact that the body part is more dense and possibly articulated into more middle paragraphs. For each document, the six sections are thus composed by an average number of sentences that depends on the document length, ranging from 2 sentences per section, for the shortest documents, to ~ 35 for the longest ones. According to this choice, documents shorter than six sentences were discarded, thus we finally relied on a corpus of 1168 documents (see Table 1 for details). As a result of the stage, we represented each section of a document as a vector of features, whose values correspond to the average value that each feature has in all sentences included in the section.

In order to understand whether and to what extent the different parts of a text represent distinctive varieties with a peculiar linguistic structure, we carried out two statistical analyses. First, we assessed whether the difference of the feature values in each section was statistically significant. Specifically, we performed a pairwise comparison between each section and the following one (i.e. 1/2, 2/3, 3/4 etc), as well as between the first and the last section (i.e. 1/6); the latter was deliberately aimed at verifying whether our set of features alone is able to distinguish between the introductory and the closing part of a document, the two more distant sections of a text which are supposed to have a more codified structure. Secondly, we verified whether there is a correlation between the values of features in the two sections under comparison. For both analyses, all data were calculated across and within genre. The cross-genre analysis was focused on genre only, thus considering the two corpora representative of the complex and simple variety as a unique one for each genre. In the second scenario, the two corpora were kept distinct to investigate if there is an effect of genre that is preserved despite language complexity changes.

3 Linguistic features

The set of features extracted from previously identified sections are distinguished into three different categories, according to the level of annotation from which they derive.

Raw Text Features: they include the average word and sentence length (*char_tok* and *n_tokens*

Genre	Corpus	Initial dataset		Analyzed dataset		
		N° Doc	Tokens	N° Doc	Tokens	Avg sentence/section
Journalism	Repubblica (Rep)	318	232.908	304	230.789	5.1
	DueParole (2Par)	321	73.314	303	71.228	2.1
Educational	High-schools educ. materials (AduEdu)	70	48.103	69	47.854	3.9
	Primary schools educ. materials (ChilEdu)	60	23.192	52	22.382	3.5
Scientific Prose	Scientific articles (ScientArt)	84	471.969	84	471.883	35.9
	Wikipedia articles (WikiArt)	293	205.071	249	200.681	4.9
Narrative	Terence&Teacher-original versions (TT orig)	56	27.833	53	25.931	4.2
	Terence&Teacher-simplified versions (TT simp)	56	25.634	54	23.866	4.3

Table 1: Statistics about the corpora used in the study.

in Table 2), calculated as the number of characters per token and of tokens per sentence, respectively.

Morpho-syntactic Features: i.e. distribution of unigrams of part-of-speech distinct into 14 coarse-grained pos tags (cpos_) and the 37 fine-grained tags (pos_) according to the ISST-TANL annotation.

Syntactic Features: these features model grammatical phenomena of different types, i.e:

- the *probability of syntactic dependency types* e.g. subject (*dep_subj*), direct object (*dep_dobj*), modifiers, calculated as the distribution of each type out of the total dependency types according to the ISST-TANL dependency tagset;
- the *length of dependency links*, i.e. the average length of all dependency links (each one calculated as the number of words occurring between the syntactic head and the dependent) (*avg_links_l*) and of the maximum dependency link (*max_links_l*);
- the *order of constituents* with respect to the syntactic head: as a proxy of canonicity effects, it is calculated the relative position of the subject, object and adverb with respect to the verbal head and the position of the adjective with respect to the nominal head;
- the *parse tree structure*, in terms of features calculating: the depth of the whole parse tree (*sent_depth*) (in terms of the longest path from the root of the dependency tree to some leaf); the width of the parse tree (*sent_width*), measured as the highest number of nodes placed on the same level; the average number of dependents for all verbal and nominal heads (*avg_dependent*);
- *subordination features*: within the group of syntactic features, a in-depth analysis was devoted to model subordination phenomena by measuring: the average distribution of subordinate clauses for sentence (*avg_sub_clause*), the percentage of sub-

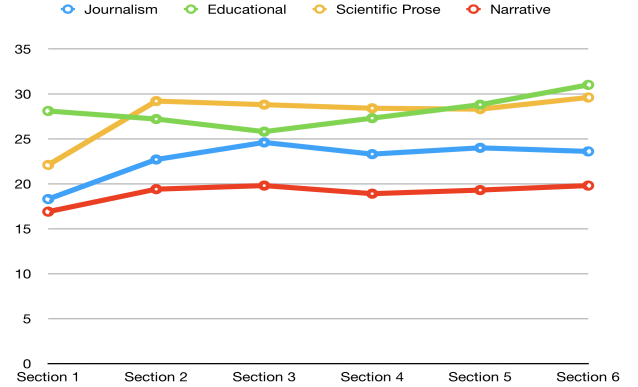


Figure 1: Average sentence length in the 6 sections across genres.

ordinate clauses with respect to the main clause (% *sub_main*) and the percentage of embedded subordinate clauses, i.e. subordinate clauses dependent on other embedded subordinate clauses (% *sub_minor*); for each type, it is also calculated the average depth (*subord_depth*) and weight (*subord_width*) of the parse tree generated by the subordinate clauses and their relative order with respect to the clause on which they depend.

4 Data Analysis

Table 2 illustrates the main findings we obtained. Specifically, it shows all features which turned out to have a statistically significant variation in at least one of the six pairwise comparisons, or a correlation score > 0.3 according to the Spearman's correlation coefficient. A first clear result is that the higher number of features varying in a statistically significant way occurs in the journalistic and scientific genre, both considered as whole (i.e. row *g* for each feature) and with respect to the language complexity variety (rows *s* and *c*). The opposite trend is reported for educational texts, which is probably due to the heterogeneous nature of this

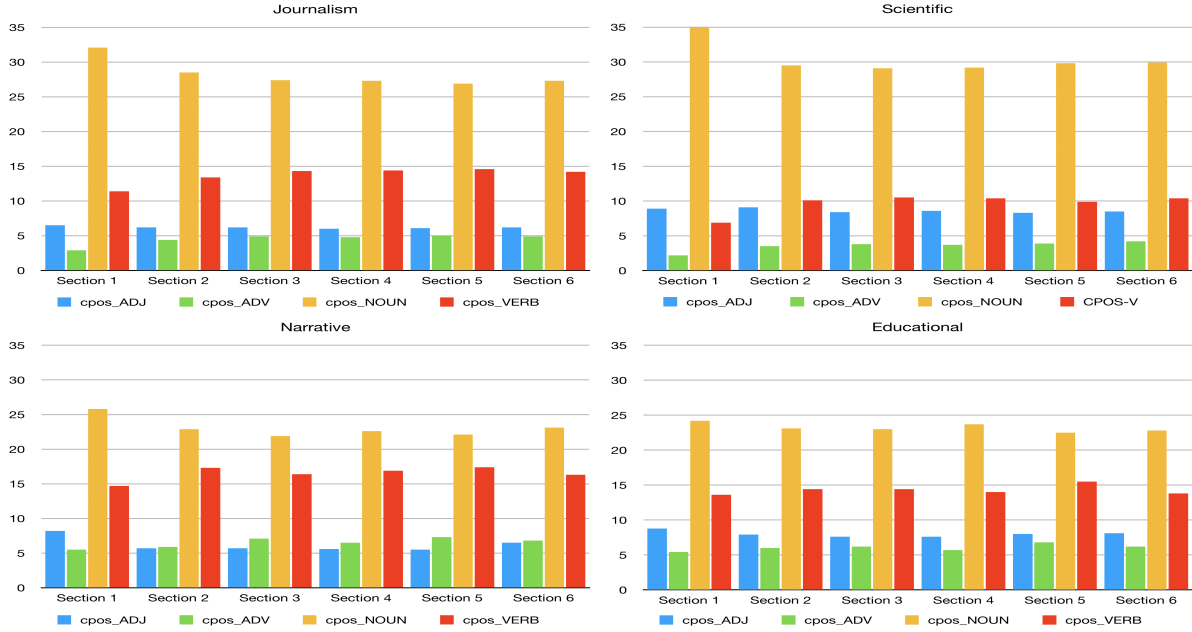


Figure 2: Distribution of lexical parts-of-speech in the four genres.

genre that includes documents of different textual typologies (course books, pieces of literature etc.).

If journalism and scientific prose are the two genres with the highest internal variance, the comparison between sections allows us to get a better understanding of this data. Specifically, for both genres, the majority of significant variations are observed between the first and the second section and between the first and the last one. This suggests that the introduction is a stylistic unit with a peculiar linguistic structure with respect to the body and the conclusion. It is characterized e.g. by shorter sentences (Figure 1), likely due to the presence of the title in both newspaper and scientific articles, and by a distinctive distribution of Parts-of-speech (Figure 2). With this respect, this data are consistent with other studies in the literature, e.g. (Voghera, 2005), and also with previous findings we obtained on the same corpora (Brunato et al., 2016), showing that scientific prose and newswire texts rely more on the nominal style. However, with the proposed approach, we were able to go further in this analysis, highlighting that noun/verb ratio is always higher in the first section than all other ones. Besides, at least for newspaper articles, this feature appears as a genre marker which is not affected by language complexity, since the same tendency is observed when the ‘simple’ and the ‘complex’ corpus are analyzed independently. The same does not hold for other features related to syntax and, in particular,

to the use of subordination. In this case, the ‘shift’ between the introduction and the subsequent part of texts yields significant variations only for articles of *Repubblica*. Specifically, the first section contains less embedded sentences (*sent_depth*: 1st sect: 5.55; 2nd sect: 7.76), and a lower presence of subordinate clauses, which appear as structurally simpler e.g. in terms of depth (*subord_depth*: 1st sect: 1.67; 2nd sect: 3.5) and width (*subord_width*: 1st sect: 0.94; 2nd sect: 1.97). Conversely, for the simple variant of this genre (i.e. the articles of the easy-to-read newspaper *2Parole*), we do not observe significant changes affecting these features; this is not particularly surprising since subordination is always less represented in this corpus with respect to all the other ones.

Leaving aside the similar tendencies characterizing the introduction, Journalistic and Scientific prose show a different behavior when we focus on the internal structure of text. While in this case much fewer features vary in a significant way, the majority occurs in the journalistic genre only, especially between the second and the third section. Again, they concern a different distribution of morpho-syntactic categories but also some syntactic features related to subordination. According to these data, we can conclude that the journalistic genre has a more rigorous structure and that it is possible to capture the boundaries between different parts by using linguistic features that are not related to the content of the article.

features	Journalism						Scientific Prose						Narrative						Educational					
	1/2	2/3	3/4	4/5	5/6	1/6	1/2	2/3	3/4	4/5	5/6	1/6	1/2	2/3	3/4	4/5	5/6	1/6	1/2	2/3	3/4	4/5	5/6	1/6
Raw text features																								
n_tokens	g	✓✓	✓*	-*	-*	✓✓	✓✓	-	-	-*	-*	✓✓	-	-	-	-	-	✓*	-*	-*	-*	-*	-*	-*
	s	✓✓	-	-	-	-	✓✓	-	-	-	-	✓✓	-*	-	-	-	-	✓	-*	-*	-*	-*	✓*	-*
	c	✓✓	-	-*	-*	✓✓	-*	-*	-*	-*	-*	-*	-	-	-	-	-	✓	-*	-	-*	-	-	-
char_tok	g	-*	-*	-*	-*	✓	✓	-	-*	-*	-	✓	-	-*	-*	-	-	-	-*	-*	-*	-*	-*	-*
	s	-*	-*	-*	-*	✓✓	✓	-	-*	-*	-	✓✓	-	-*	-*	-	-	-	✓*	-*	-*	-*	-*	-*
	c	-	-*	-*	-*	-	-*	✓*	-*	-*	-*	-*	-	-*	-*	-	-	-	-*	-*	-*	-*	-*	-*
Morpho-syntactic features																								
cpos_ADJ	g	-	-	-	-	-	✓	✓	-*	-	-	*	-	-	-	-	-	-	-	-*	-*	-	-*	-*
	s	✓	-	-	-	-	✓✓	-	-*	-	-	✓	-	-	-	-	-	-	✓✓	-	-	-	-	-*
	c	✓	-	-	-	✓	-	-*	-*	-*	-*	-*	-	-	-	-*	-*	-*	-	-*	-*	-	-*	-
cpos_ADV	g	✓✓*	✓*	-	-	✓✓	✓✓*	-	-	-*	-	✓✓	-	-	-	-	-	✓	-	-	-*	✓	-	-
	s	✓✓	-	-	-	✓✓	✓✓*	-	-	-	-	✓✓	-	-	-	-	-	✓	✓	-	-*	✓	-	-
	c	✓✓*	-*	-	-	✓✓	-*	-*	-*	-*	-*	-*	-	-	-	-	-	-	-	-	-*	-	-	-
cpos_CONJ	g	✓✓	✓	-	-	✓✓	✓✓	-	-	-	-	✓✓	-	-	-	✓	-	-	-	✓	✓	-	-	-
	s	✓✓	-	-	-	✓✓	✓✓	-	-	-	-	✓✓	-	-	-	-	-	-	-	✓	-	-	-	-
	c	✓✓	-	-	-	✓✓	✓*	-*	-*	-*	-*	✓*	-	-*	-	-*	-	-	-	-	-	-	-	-
cpos_NOUN	g	✓✓*	✓*	-*	-*	✓✓	✓✓	-	-	-*	-	✓✓	✓✓	-	-	-*	-	✓✓*	-	-*	-*	-*	-*	-*
	s	✓✓*	-*	-	-	✓✓	✓✓	-	-	-	-	✓✓	✓	-	-	-*	-	✓✓	✓	-*	-*	-*	-*	-*
	c	✓✓*	✓✓*	-*	-*	✓✓	-*	-*	-*	-*	-*	-*	✓*	-	-	-*	-	✓*	-	-	-*	-*	-*	-
pos_PROP_N	g	✓✓*	-*	-*	-*	✓✓*	✓✓	-*	-*	✓*	-*	✓✓	✓	-*	-*	-*	-*	-*	-*	-*	-*	-*	-*	-*
	s	✓✓*	-*	-*	-*	✓✓*	✓✓	-*	-*	-*	-*	✓✓	-	-*	-*	-*	-*	-	✓*	-*	-*	-*	-*	-*
	c	✓✓	✓✓	-	-*	✓✓	-*	-*	-*	-*	-*	-*	-	-*	-*	-*	-*	-	-*	-*	-*	-*	-	-*
cpos_VERB	g	✓✓	✓	-	-	✓✓	✓✓	-	-	-	-	✓✓	✓	-	-	-	-	-	-	-	-	✓	✓	-
	s	✓✓	-	-	-	✓✓	✓✓	-	-	-	-	✓✓	✓	-	-	-	-	-	-	-	-	✓	✓	-
	c	✓✓	✓✓	-	-*	✓✓	-*	-*	-*	-*	-*	-*	✓	-	-	-	-	-	-	-*	-*	-	-	-
pos_AUX	g	✓*	✓*	-*	-*	✓	✓✓*	-	-	-	-	✓✓	-	-	-	-	-	-	-	-	-	-	✓	-
	s	-*	-*	-*	-*	-	✓✓	-	-	-	-	✓✓	-	-	-*	-	-	-	-	-	-	-	-	-*
	c	✓✓*	✓✓*	-*	-*	✓✓	-*	-	-*	-	-*	-*	-	-	-	-*	-	-	-	-	-	-	-	-
Syntactic features																								
dep_dobj	g	✓	✓	-	-	✓✓	✓✓	-	-	-	-	✓✓	✓	-	-	✓	-	✓	-	✓✓	-	-	-	✓
	s	-	✓	-	-	✓✓	✓✓	-	-	-	-	✓✓	✓	✓✓	-*	-	-	✓	-	✓✓	-	-*	-	✓
	c	✓✓	-	-	-	✓✓	-*	-*	-*	-*	-*	-*	-	-	-	-	-	-	-	-	-	-	-	-
dep_subj	g	-	-	-	-	-	✓✓	-	-	-	-	✓✓	✓	-	-	-	-	-	-	-*	-	-	-	-
	s	-	-	-	-	-	✓✓	-	-	-	-	✓✓	-	-*	-	-	-	-	-*	-*	-	-*	-	-
	c	✓	-	-	-	✓✓	-*	-*	-*	-*	-*	-*	✓	-	-	-	-	-	-	-	-	-	-	-
max_links_1	g	✓✓	-	✓*	-*	✓✓	✓✓	-	-	-*	-	✓✓	-	-	-	-	-*	-	-*	-*	-*	-*	-*	-*
	s	✓✓	-	-	-	✓	✓✓	-	-	-	-	✓✓	-	-	-	-	-	-	-	-*	-*	-*	-*	-*
	c	✓✓	-	✓	-*	✓✓	-*	-*	-*	-*	-*	-*	-	-	-	-	-*	-	-*	✓	-	-	-	-
avg_links_1	g	✓✓	-	✓	-	✓✓	✓✓	-	-	-	-	✓✓	-	-	-	-	-	-	-	-	-	-*	-	-
	s	✓	-	-	-	-	✓✓	-	-	-	-	✓✓	-	-	-	-	-	-	-	-	-*	-*	-	-
	c	✓✓	-	✓	-	✓✓	-*	-*	-*	-*	-*	-	-	-	-*	-	-*	-	-	-*	-	-	-	-
sent_depth	g	✓✓	-*	-*	-*	✓✓	✓✓	-	-	-*	✓✓	✓✓	-*	-	-*	-*	-*	-	-*	-*	-*	-*	-*	-*
	s	-	-	-	✓	✓	✓✓	-	-	✓	-*	✓✓	-*	-	-*	-*	-*	-	-	-*	-*	-*	-*	✓*
	c	✓✓	-	-*	-*	✓✓	-*	-*	-*	-*	-*	-*	-	-	-*	-*	-	-	-*	-*	-	-*	-*	-*
sent_width	g	✓✓	✓	-	-	✓✓	✓✓	-	-	-	-	✓✓	-	-	-	-	-	-	-*	-*	-	-*	-	-
	s	✓✓	-	-	-	-	✓✓	-	-	-	-	✓✓	-	-	-	-	-	-	-*	-*	-	-*	-*	-*
	c	✓✓	-	✓	-*	✓✓	-*	-*	-*	-*	-*	-*	-	-*	-	-	-*	-	-	-*	✓	-*	-	-
avg_dependent	g	✓✓	✓	-*	-*	✓✓	✓✓	-*	-*	-*	✓*	✓✓	✓	-*	-*	-*	-*	✓	-*	-*	-*	-*	-*	-*
	s	✓	-	-	-*	-	✓✓	-	-	-	-	✓✓	-	-	-*	-	-*	✓	-	-*	-*	-*	-*	-*
	c	✓✓	✓	-*	-*	✓✓	-*	-*	-*	-*	-*	-*	-	-*	-*	-	-*	✓	-*	-*	✓	-*	-*	-*
Subordination features																								
avg_sub_clause	g	✓✓	✓*	-*	-	✓✓	✓✓	-	-	-	-	✓✓	✓	-	-	-*	-*	✓✓	-	-	-	-	-*	-*
	s	-	-	-	-	✓✓	✓✓	-	-	-	-	✓✓	-	-	-	-*	-*	✓	-	-	-	-*	-*	-*
	c	✓✓	-	-	-*	✓✓	-*	-*	-	-*	-*	-*	✓✓	✓	-	-	-	-	-*	-	-	-	-	-
subord_depth	g	✓✓	✓	-	-	✓✓	✓✓	-	-	-	✓	✓✓	-*	-	-*	-*	-*	-	-	-*	-*	-*	-	-
	s	-	-	-	-	✓✓	✓✓	-	-	-	-	✓✓	-	-	-	-*	-*	✓✓	-	-	-*	-*	-*	✓
	c	✓✓	-	-	-	✓✓	-*	-*	-*	-*	-*	-*	-	-	-	-*	-	-	✓	-	-	-	-	-
subord_width	g	✓✓	✓	-	-	✓✓	✓✓	-	-	-	✓	✓✓	-	-	-	-	-	-	-	-	-*	-	-	-
	s	-	-	-	-	✓	✓✓	-	-	-	-	✓✓	-	✓	-	-*	-*	✓	-	✓	-*	-*	-*	-
	c	✓✓	-	-	-	✓✓	-*	-*	-	-*	-*	-*	-	-	-	-*	-	-	-	-*	✓	-*	-	-
% sub_main	g	✓✓	✓	-	-	✓✓	✓✓	-	-	-	-	✓✓	-	-	-	-	-*	✓✓	-	-	-	-	-	-
	s	-	-	-	-	✓✓	✓✓	-	-	-	-	✓✓	-	-	-	-*	✓✓	-	✓*	✓✓*	-*	✓*	-	-
	c	✓✓	-	-	-	✓✓	-*	-*	-*	-*	-*	-*	-	-	-	-	-	-	✓	-	-	-	-	-
% sub_minor	g	✓✓	-*	-	-*	✓✓	-	-	-	-	-	✓	-	-	-	-	✓	-	-*	✓	-	-	-	✓
	s	-	-	-	-	✓	-	-	-	-	-	✓	-	-	-	-	-*	-*	-	✓	-	-*	-*	-*
	c	✓✓	-	-	-	✓✓	-*	-*	-	-*	-	-	✓	-	-	-*	-	-	-	-	-	-*	-	✓

Table 2: A set of linguistic features resulting as significant in at least one pairwise comparison. ✓✓ means highly statistically significant ($p < 0.001$), ✓ statistically significant ($p < 0.05$), - no significance; * correlation related to the Spearman's rank correlation coefficient ($\rho > 0.3$), g=global corpus, s=simple variety of the corpus, c=complex variety of the corpus.

5 Conclusion

In this paper we have presented a novel approach to the study of language variation, which relies on the prerequisites of the linguistic profiling methodology but with the specific purpose of modeling the stylistic form of the different parts within a text. A cross-genre investigation on four traditional genres in Italian, and two levels of complexity for each, showed that morpho-syntactic and syntactic features are differently distributed across subsections of texts belonging to a specific genre and language variety. This approach has important implications for research on genre variation since it suggests that the characterization of texts and texts varieties should benefit by inspecting corpora from this fine-grained perspective. A better understanding of linguistic phenomena characterizing the introductory, middle and conclusive parts of a text is also highly relevant not only to enhance automatic genre classification but also for other natural language processing applications devoted to modeling style: e.g. in education, as a component of intelligent tutoring systems able to provide detailed feedback to students in writing courses or for the automatic generation of texts with the stylistic properties of a specific genre and level of complexity.

Acknowledgments

This work was partially supported by the 2-year project ADA, Automatic Data and documents Analysis to enhance human-based processes, funded by Regione Toscana (BANDO POR FESR 2014-2020).

References

- Giuseppe Attardi, Felice Dell’Orletta, Maria Simi, Joseph Turian. 2009. Accurate dependency parsing with a stacked multilayer perceptron. In *Proceedings of EVALITA 2009 - Evaluation of NLP and Speech Tools for Italian 2009*. Reggio Emilia, Italy, December 2009.
- Douglas Biber. 1993. Using register-diversified corpora for general language studies. *Computational Linguistics*, 19(2), 219–242.
- Dominique Brunato and Felice Dell’Orletta. 2017. On the order of words in Italian: a study on genre vs complexity. *International Conference on Dependency Linguistics (Depling 2017)*, 18-20 September 2017, Pisa, Italy.
- Dominique Brunato, Felice Dell’Orletta, Simonetta Montemagni and Giulia Venturi. 2016. Monitoraggio linguistico di Scritture Brevi: aspetti metodologici e primi risultati. A. Manco e A. Mancini (eds.), *Scritture Brevi: segni, testi e contesti. Dalle iscrizioni antiche ai tweet*, Collana di studi Quaderni di AION-Linguistica, Università di Studi di Napoli “L’Orientale”, Napoli, 149–176.
- Andrea Cimino, Martijn Wieling, Felice Dell’Orletta, Simonetta Montemagni, Giulia Venturi. 2017. Identifying Predictive Features for Textual Genre Classification: the Key Role of Syntax. *Proceedings of 4th Italian Conference on Computational Linguistics (CLiC-it)*, 11-13 December, 2017, Rome.
- Kevyn Collins-Thompson. 2014. Computational Assessment of text readability. *Recent Advances in Automatic Readability Assessment and Text Simplification. Special issue of International Journal of Applied Linguistics*, 165:2, John Benjamins Publishing Company, 97-135.
- Crossley, S.A., Dempsey, K., McNamara, D.S. 2011. Classifying paragraph types using linguistic features: Is paragraph positioning important? *Journal of Writing Research*.
- Felice Dell’Orletta. 2009. Ensemble system for part-of-speech tagging. In *Proceedings of EVALITA 2009 - Evaluation of NLP and Speech Tools for Italian 2009*, Reggio Emilia, Italy, December 2009.
- S. Malmasi, E. Keelan, A. Cahill, J. Tetreault, R. Pugh, C. Hamill, D. Napolitano, and Y. Qian 2017. A report on the 2017 native language identification shared task. In *Proceedings of the 12th Workshop on Building Educational Applications Using NLP*.
- Maria Emanuela Piemontese. 1996. *Capire e farsi capire. Teorie e tecniche della scrittura controllata*. Napoli, Tecnodid.
- Efstathios Stamatatos, Nikos Fakotakis and George Kokkinakis. 2001. Automatic text categorization in terms of genre and author. *Computational Linguistics*, (26) 471–495.
- Hans van Halteren. 2004. Linguistic profiling for author recognition and verification. In *Proceedings of the Association for Computational Linguistics (ACL04)*, 200207.
- Miriam Voghera. 2005. La misura delle categorie sintattiche. In Chiari Isabella / De Mauro Tullio (eds.) *Parole e numeri. Analisi quantitative dei fatti di lingua*, Aracne, Roma, 125–138.

Annotating Shakespeare's Sonnets with Appraisal Theory to Detect Irony

Nicolò Busetto

Department of Linguistic Studies
Ca Foscari University
Ca Bembo - Venezia
830070@stud.unive.it

Rodolfo Delmonte

Department of Linguistic Studies
Ca Foscari University
Ca Bembo - Venezia
delmont@unive.it

Abstract

English. In this paper we propose an approach to irony detection based on Appraisal Theory (Martin and White (2005)) in Shakespeare's Sonnets, a well-known data set that is statistically valuable. In order to produce meaningful experiments, we created a gold standard by collecting opinions from famous literary critics on Shakespeare's Sonnets focusing on irony. We started by manually annotating the data using Appraisal Theory as a reference theory. This choice is motivated by the fact that Appraisal annotation schemes allow smooth evaluation of highly elaborated texts like political commentaries. The annotation is then automatically compiled and checked against the gold standard in order to verify the persistence of certain schemes that can be identified as ironic, satiric or sarcastic. Upon observation, irony detection reaches a final match of 80%¹.

Italiano. In questo articolo si propone un approccio basato sulla Appraisal Theory per l'individuazione dell'ironia nei Sonetti di Shakespeare, un dataset che è statisticamente valido. Allo scopo di produrre esperimenti significativi, abbiamo creato un gold standard raccogliendo le opinioni di famosi critici letterari sullo stesso corpus, con l'ironia come tema. Abbiamo poi annotato manualmente i sonetti utilizzando gli strumenti e i tratti della Appraisal Theory che permettono di ottenere una valutazione di testi altamente elaborati come gli articoli di politica. L'annotazione è

stata poi raccolta automaticamente e confrontata con il gold standard per verificare la persistenza di certi schemi che possono essere identificati come ironici, satirici o sarcastici, raggiungendo una corrispondenza finale del 80%.

1 Introduction

Shakespeare's Sonnets are a collection of 154 poems which is renowned for being full of ironic content (Weiser (1983)), (Weiser (1987)) and for its ambiguity thus sometimes reverting the overall interpretation of the sonnet. Lexical ambiguity, i.e. a word with several meanings, emanates from the way in which the author uses words that can be interpreted in more ways not only because inherently polysemous, but because sometimes the additional meaning they evoke can sometimes be derived on the basis of the sound, i.e. homophone (see "eye", "I" in sonnet 152). The sonnets are also full of metaphors which many times requires contextualising the content to the historical Elizabethan life and society. Furthermore, there is an abundance of words related to specific language domains in the sonnets. For instance, there are words related to the language of economy, war, nature and to the discoveries of the modern age, and each of these words may be used as a metaphor of love. Many of the sonnets are organized around a conceptual contrast, an opposition that runs parallel and then diverges, sometimes with the use of the rhetorical figure of the chiasmus. It is just this contrast that generates irony, sometimes satire, sarcasm, and even parody. Irony may be considered in turn as: what one means using language that normally signifies the opposite, typically for humorous or emphatic effect; a state of affairs or an event that seems contrary to what one expects and is amusing as a result. As to sarcasm this may be regarded the

¹Copyright © 2019 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0)

use of irony to mock or convey contempt. Parody is obtained by using the words or thoughts of a person but adapting them to a ridiculously inappropriate subject. There are several types of irony, though we select verbal irony which, in the strict sense, is saying the opposite of what you mean for outcome, and it depends on the extra-linguistics context(Attardo(1994)). As a result, Satire and Irony are slightly overlapping but constitute two separate techniques; eventually Sarcasm can be regarded as a specialization or a subset of Irony. It is important to remark that in many cases, these linguistic structures may require the use of nonliteral or figurative language, i.e. the use of metaphors. This has been carefully taken into account when annotating the sonnets by means of Appraisal Theory Framework (hence ATF). In our approach we will follow the so-called incongruity presumption or incongruity-resolution presumption. Theories connected to the incongruity presumption are mostly cognitive-based and related to concepts highlighted for instance, in (Attardo(2000)). The focus of theorization under this presumption is that in humorous texts, or broadly speaking in any humorous situation, there is an opposition between two alternative dimensions. As a result, we will look for contrast in our study of the sonnets, produced by the contents of manual classification. The purpose of this study is to show how ATF can be useful for detecting irony, considering its ambiguity and its elusive traits.

2 Producing the Gold Standard

In order to produce a gold standard that may encompass strong hints to classification in terms of humour as explained above, we collected literary critics' reviews of the sonnets. We used criticism from a set of authors including (Frye(1957)) (Calimani(2009)) (Melchiori(1971)) (Eagle(1916)) (Marelli(2015)) (Schoenfeldt(2010)) (Weiser(1987)) (Serpieri(2002)) all listed in the reference section. The gold standard classification has been produced by second author and checked by first author. It is organized into a number of separate fields in a sequence to allow the reader to get a better picture of the sonnet in the collection. All classifications are reported in a supplementary file in the Appendix. Here below are the classifications for two sonnets:

- *SONNET 8*

SEQUENCE: 1-17 Procreation *MAIN*

THEME: One against many *ACTION:* Young man urged to reproduce *METAPHOR:* Through progeny the young man will not be alone *NEG.EVAL:* The young man seems to be disinterested *POS.EVAL:* Young man positive aesthetic evaluation *CONTRAST:* Between one and many

- *SONNET 21*

SEQUENCE: 18-86 Time and Immortality *MAIN THEME:* Love *ACTION:* The Young man must understand the sincerity of poet's love *METAPHOR:* True love is sincere *NEG.EVAL:* The young man listens the false praise made by others *POS.EVAL:* Young Man positive aesthetic evaluation *CONTRAST:* Between true and fictitious love

As can be seen, we indicate *SEQUENCE* for the thematic sequence into which the sonnet is included; this is followed by *MAIN THEME* which is the theme the sonnet deals with; *ACTION* reports the possible action proposed by the poet to the protagonist of the poem; *METAPHOR* is the main metaphor introduced in the poem sometimes using words from a specialized domain; *NEG.EVAL* and *POS.EVAL* stand for Negative Evaluation and Positive Evaluation contained in the poem in relation to the theme and the protagonist(s); finally, *CONTRAST* is the key to signal presence of opposing concrete or abstract concepts used by Shakespeare to reinforce the arguments purported in the poem. Many sonnets have received more than one possible pragmatic category. This is due to the difficulty in choosing one category over another. In particular, it has been particularly hard to distinguish Irony from Satire, and Irony from Sarcasm. Overall, we ended up with 54 sonnets receiving a double marking over 98, representing the total number of sonnets with some kind of pragmatic label by the literary critics, with a ratio of 98/154, corresponding to a percentage of 63.64%. We ended up with the count of annotated sonnets reported above in Table 1.

Eventually, as commented in the section below, the introduction of annotations based on Appraisal Theory has helped in choosing best pragmatic classification. In fact, literary critics were simply hinting at "irony" or "satire", but the annotation gave us a precise measure of the level of contrast present in each of the sonnets regarded generically as "ironic".

Table 1: Final distribution of sonnets in the 5 pragmatic categories

Type	Quantity
Blank	57
Irony	73
Satire	20
Parody	4
Sarcasm	47
Duplicated	54

2.1 Appraisal Theory for Poetry and Literary Texts

The experiment we have been working on is an attempt to describe irony, parody and sarcasm in terms of a strict scientifically viable linguistic theory, the Appraisal Framework Theory (Martin and White(2005)), as has already been done in the past by other authors (see (Taboada and Grieve(2004)) (Read and Carrol(2012)) but also (Stingo and Delmonte(2016)) (Delmonte and Marchesini(2017)) . The idea is as follows: produce a complete annotation of the sonnets using the tools made available by the theory and then verify how well it fits into the gold standard produced. The primary purpose of the Appraisal Framework Theory(hence AFT) is to delineate the interpersonal dimension of communication, supplying schemes by which it is possible to recognize evaluative sequences within texts and information about the positioning of the author in relation to evaluated targets.²

The annotation has been organized around only one category, Attitude, and its direct subcategories, in order to keep the annotation at a more workable level, and to optimize time and space in the XML annotation. Attitude includes different options for expressing positive or negative evaluation, and expresses the author's feelings. The main category is divided into three primary fields with their relative positive or negative polarity, namely:

- *Affect* is every emotional evaluation of things, processes or states of affairs, (e.g. like/dislike), it describes proper feelings and any emotional reaction within the text aimed towards human behaviour/process and phenomena.

- *Judgement* is any kind of ethical evaluation of human behaviour, (e.g. good/bad), and considers the ethical evaluation on people and their behaviours.
- *Appreciation* is every aesthetic or functional evaluation of things, processes and state of affairs (e.g. beautiful/ugly; useful/useless), and represent any aesthetic evaluation of things, both man-made and natural phenomena.

Eventually, we end up with six different classes: Affect positive, Affect Negative, Judgement Positive, Judgement Negative, Appreciation Positive, Appreciation Negative. Overall in the annotation there is a total majority of positive polarities with a ratio of 0.511, in comparison to negative annotations with a ratio of 0.488. In short, the whole of the positive poles is 607, and the totality of the negative poles is 579 for a total number of 1186 annotations. Judgement is the more interesting category because it allows social moral sanction, in that it refers to two subfields, Social Esteem and Social Sanction - which however we decided not to mark. In particular, whereas the positive polarity annotation of Judgement extends to Admiration and Praise, the negative polarity annotation deals with Criticism and Condemnation or Social Esteem and Social Sanction (see (Martin and White(2005)), p.52). In particular, Judgement is found mainly in the final couplet of the sonnets.

The annotation work on the texts has been accomplished by first author and checked by second author. Given the level of objective difficulty in understanding the semantic content of the sonnets, we have decided not to resort to additional annotators - second author produced the annotation as part of his Master thesis work. So far, we have not been able to produce a measure for interannotator agreement: however, since I was obliged to correct 35% of all annotations that measure could be approximated by 65% of agreement. The tags we used for the annotation include a tag for <text> contains the whole text of the sonnet; <p> to mark stanzas, and <s> to mark lines. Focusing on the annotation of the evaluative sequences instead, every time we found an evaluative word (or sequence of words), we delimited the item/phrase within the tags <apprsl></apprsl>. Subsequently, following the general indications mentioned above provided by

²Further information can be found on the dedicated website dedicated to the Appraisal Framework Theory: <http://www.languageofevaluation.info/appraisal/>

(Martin and White(2005)), we assigned one of the three subcategories – affect, judgement and appreciation – as an attribute of the tag <apprsl>, also providing the positive/negative sentiment orientation as a value of the attribute. Here below we show the annotation for Sonnet 40 which is highly contrasted:

```
<?xml version="1.0" encoding="ISO-8859-1"?> <text> <p> <s> Take all my loves, my love,  
yea take them all, </s> <s> What hast thou then  
more than thou hadst before? </s> <s> No love,  
my<apprsl affect="positive">love,</apprsl>that  
thou mayst<apprsl appreciation="positive">  
true</apprsl>love call, </s> <s> All mine was  
thine, before thou hadst this more: </s>  
</p> <p> <s> Then if for my<apprsl af-  
fect="positive">love,</apprsl>thou my<apprsl  
affect="positive">love</apprsl>receivest,  
</s> <s> I cannot<apprsl judg-  
ment="negative">blame</apprsl>thee, for  
my<apprsl affect="positive">love</apprsl>thou  
usest, </s> <s> But yet<apprsl judge-  
ment="negative">be blamed,</apprsl>if  
thou thy self<apprsl judge-  
ment="negative">deceivest</apprsl>  
</s> <s> By<apprsl apprecia-  
tion="negative">wilful</apprsl>taste  
of what thy self<apprsl apprecia-  
tion="negative">refusest</apprsl> </s> </p>  
<p> <s> <apprsl judgement="positive">I  
do forgive</apprsl><apprsl judge-  
ment="negative">thy robbery</apprsl> <ap-  
prsl appreciation="positive">gentle</apprsl>  
thief </s> <s> textbfAlthough<apprsl  
judgement="negative">thou steal thee  
all my poverty;</apprsl> </s> <s>  
And yet love knows it is a<apprsl af-  
fect="negative">greater grief</apprsl> </s>  
<s> To<apprsl appreciation="negative">bear  
love's wrong,</apprsl>than<apprsl apprecia-  
tion="negative">hate's known injury</apprsl>.  
</s> </p> <p> <s> <apprsl apprecia-  
tion="negative">Lascivious</apprsl>grace,  
in whom <apprsl apprecia-  
tion="negative">all ill</apprsl> well  
shows, </s> <s> Kill me with<apprsl af-  
fect="negative">spites</apprsl>yet <ap-  
prsl judgement="positive">we must not be  
foes</apprsl>. </s> </p> </text>
```

In the choice of which and how many items to annotate, we adopted the following linguistic criteria to enhance the notational analysis.

- Semantic criteria:

Anytime one or more verb/noun modifiers are found, when they do not represent meaningful evaluation by themselves, they are annotated together with the part of speech that they contribute to modify. Any instance of evaluation of a multiword expression, is annotated as a single appraisal unit. Any instance of evaluation of rhetorical or figurative language, is annotated as a single appraisal unit. When possible, the evaluations are embedded so as to include appraisal units into a bigger evaluative unit, in order to fully capture figures of speech such as oxymora, apagoges, rhetorical questions, interjections and the like.

- Syntactic Criteria:

Without exceeding the length of the proposition, it is allowed to annotate phrases as single appraisal unit up until a clause-level, whenever they express opinions or evaluations. Additionally, for those cases where complex phrasal structures were found, we limited ourselves to the annotation of the most evaluative part within the overall sequence, so as to avoid overproduction of long annotation. Again, when possible, the clauses have been de-structured so that through embedding we were able to capture the evaluation on a clause-level in greater detail. It is allowed to annotate evaluative sequences on a clause level even beyond the punctuation marks limits. However, these annotations are very rare. In case of dyad/triad of items, whenever they share the same attribute and the same polarity orientation, they are annotated as single evaluative units. In case of more than three items in a row that share the same attribute and the same polarity orientation, they were annotated separately.

As to interpretation criteria, we assumed that sonnets with the highest contrast could belong to the category of Sarcasm. The reason for this is justified by the fact that a high level of Negative Judgements accompanied by Positive Appreciations or Affect is by itself interpretable as the intention to provoke a sarcastic mood. As a final

result, there are 44 sonnets that present the highest contrast and are specifically classified according to the six classes above (see Figure 1 in the Appendix). There is also a group that contains ambiguous sonnets which have been classified with a double class, mainly by Irony and Sarcasm. As a first remark, in all these sonnets, negative polarity is higher than positive polarity with the exception of sonnet 106. In other words, if we consider this annotation as the one containing the highest levels of Judgement, we come to the conclusion that possible Sarcasm reading is mostly associated with presence of Judgement Negative and in general with high Negative polarity annotations (see table 2 below). As a first result, we may notice a very high convergence existing between critics' opinions as classified by us with the label highest contrast and the output of manual annotation by Appraisal classes.

Table 2: Quantitative data for six appraisal classes for sonnets with highest contrast

Classes	Sum	Mean	St.Dev.
Appr.Pos	56	2.534	8.199
Appr.Neg	25	1.134	3.691
Affct.Pos	53	2.4	7.733
Affct.Neg	77	3.467	11.202
Judgm.Pos	32	1.445	4.721
Judgm.Neg	122	5.467	17.611

In the group of 50 sonnets classified, mainly or exclusively, with Irony, the presence of Judgement Negative is much lower than in the previous table for Sarcasm (see Figure 2 in the Appendix). In fact only half of them – 25 – has annotation for that class, the remaining half introduces two other negative classes: mainly Affect Negative, but also Appreciation Negative - see table 3 below. As to the main Positive class, we can see that it is no longer Judgement Positive, but Appreciation Positive which is present in 33 sonnets. This is followed by Affect Positive which is better distributed.

In other words we can now consider that Sarcasm is characterized by a majority of negative evaluations 224 over 141; while Irony is characterized by a majority of Positive evaluations 262 over 183 and that the values are sparse and unequally distributed. The final table concerns the number of sonnets with blank evaluation by critics which amount to 60. As a rule, this group of son-

Table 3: Quantitative data for six appraisal classes for sonnets with lowest contrast

Classes	Sum	Mean	St.Dev.
Appr.Pos	139	5.346	18.821
Appr.Neg	65	2.5	8.844
Affct.Pos	64	2.462	8.708
Affct.Neg	81	3.115	11.009
Judgm.Pos	59	2.269	8.029
Judgm.Neg	37	1.423	5.047

Table 4: Quantitative data for six appraisal classes for sonnets with no contrast

Classes	Sum	Mean	St.Dev.
Appr.Pos	88	3.034	1.269
Appr.Neg	59	2.034	7.638
Affct.Pos	89	3.069	11.483
Affct.Neg	109	3.759	14.052
Judgm.Pos	49	1.689	6.367
Judgm.Neg	8	0.276	1.079

nets look different from the two groups we already analysed. The prevailing trait is Affect Negative; Judgement Negative is only occasionally present; the second preminent trait is Affect Positive. In order to know how much the difference is, we can judge from the quantities shown in table 3 above (but see also Figure 3 in the Appendix).

In particular, in this case the ratio Negative/Positive is more balanced 226 over 176 with a majority of Positive annotations as happened with Irony but with a lower gap. The appraisal category with highest number of annotations is now Affect, whereas in the case of Irony it was Appreciation, and in Sarcasm it was Judgement. So eventually we have been able to differentiate the three main and more frequent pragmatic categories by means of Appraisal Framework features: they are characterized by a different distribution of positive vs. negative evaluations and also by a prominent presence of one of the three main subcategories into which Appraisal has been subdivided that is Appreciation for Irony, Judgement for Sarcasm and Affect where no evaluation has been expressed.

3 Conclusion

In this paper we have presented work carried out to annotate and experiment with the theme of irony in Shakespeare's Sonnets. The gold standard for the

experiment has been created by collecting comments produced by literary critics on the presence of some kind of thematic, semantic and syntactic opposition in the sonnets as to produce some sort of irony. At first the sonnets have been annotated using the framework of Appraisal Theory and then we checked the results: we obtained a very high level of matching with the critics' opinions at 80%. Eventually, Appraisal framework has shown its ability to classify and diversify different levels of irony effectively.

References

- Salvatore Attardo. 1994. *Linguistic Theories of Humor*. Mouton de Gruyter, Berlin – New York.
- Salvatore Attardo. 2000. Irony as relevant inappropriateness. *Journal of Pragmatics*, 34(32).
- Dario Calimani. 2009. *William Shakespeare, I sonetti della menzogna*. Carrocci, Roma.
- Rodolfo Delmonte and Giulia Marchesini. 2017. A semantically-based approach to the annotation of narrative style. In *Proceedings of the 13th Joint ISO-ACL Workshop on Interoperable Semantic Annotation (ISA-I3)*, pages 14–25, Stroudsburg, PA, USA. ACL.
- R.L. Eagle. 1916. *New light on the enigmas of Shakespeare's Sonnets*. John Long Limited, London.
- Northrop Frye. 1957. *Anatomy of Criticism: Four Essays*. Princeton University Press.
- Maria Antonietta Marelli. 2015. *William Shakespeare, I Sonetti – con testo a fronte*. Garzanti.
- J. Martin and P.R. White. 2005. *Language of Evaluation, Appraisal in English*. Palgrave Macmillan, London and New York.
- Giorgio Melchiori. 1971. *Shakespeare's Sonnets*. Adriatica Editrice, Bari.
- J. Read and J. Carrol. 2012. Annotating expressions of appraisal in english. *Language Resources and Evaluation*, 46:421–447.
- Michael Schoenfeldt. 2010. *Cambridge introduction to Shakespeare's poetry*. Cambridge University Press, Cambridge.
- Alessandro Serpieri. 2002. *Polifonia Shakespeariana*. Bulzoni, Roma.
- Michele Stingo and Rodolfo Delmonte. 2016. Annotating satire in italian political commentaries with appraisal theory. In *Natural Language Processing meets Journalism - Proceedings of the Workshop, NLP MJ-2016*, pages 74–79, Stroudsburg, PA, USA. ACL.
- M. Taboada and J. Grieve. 2004. Analyzing appraisal automatically. In *Proceedings of the AAAI Spring Symposium on Exploring Attitude and Affect in Text: Theories and Applications*, pages 158–161. AAAI Press.
- David K. Weiser. 1983. <http://www.jstor.org/stable/43343552> Shakespearean irony: The 'sonnets'. *Neuphilologische Mitteilungen*, 84(4):456–469.
- David K. Weiser. 1987. *Mind in Character – Shakespeare's Speaker in the Sonnets*. The University of Missouri Press.

APPENDIX.
Figures Of the Six Pragmatic Categories for Appraisal-Based Classification

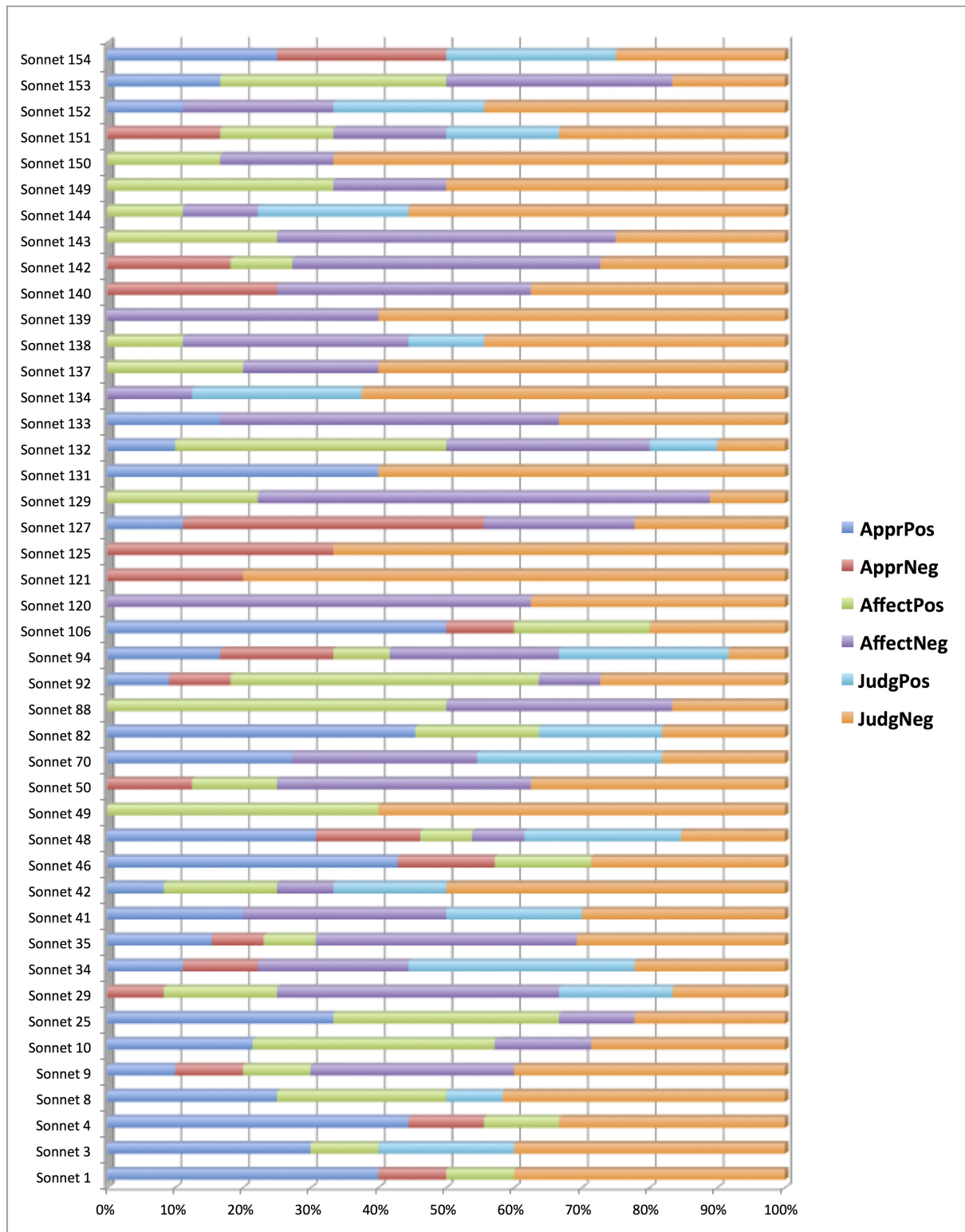


Figure 1: Subdivision into six appraisal classes for sonnets with highest contrast

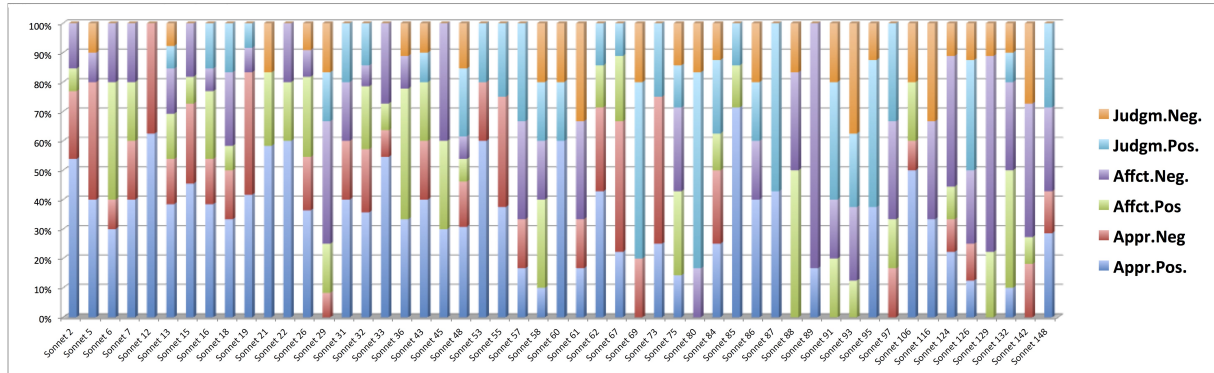


Figure 2: Subdivision into six appraisal classes for sonnets with lowest contrast

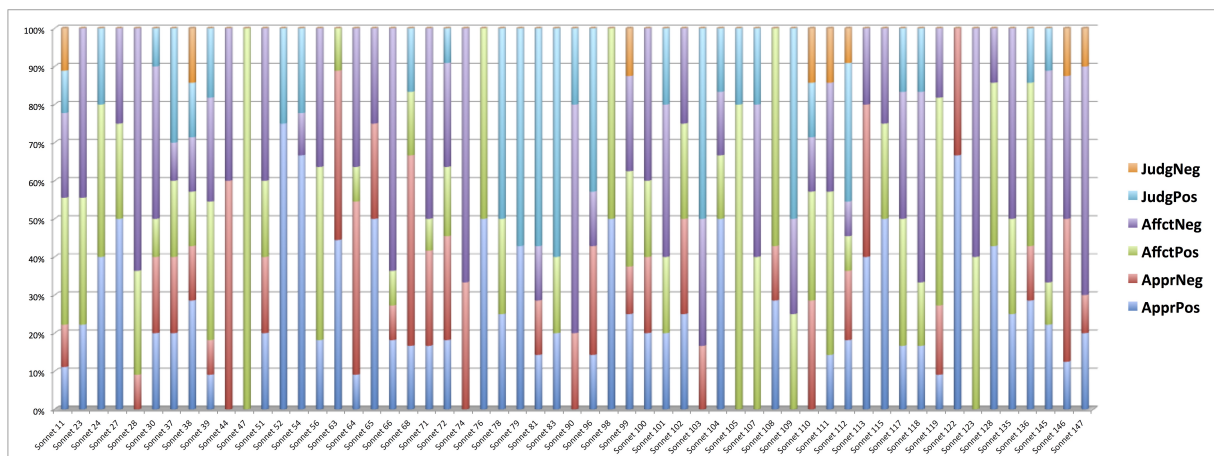


Figure 3: Subdivision into six appraisal classes for sonnets with no contrast

Embeddings Shifts as Proxies for Different Word Use in Italian Newspapers

Michele Cafagna^{1,3}, Lorenzo De Mattei^{1,2,3} and Malvina Nissim³

¹Department of Computer Science, University of Pisa, Italy

²ItaliaNLP Lab, ILC-CNR, Pisa, Italy

³University of Groningen, The Netherlands

{m.cafagna,m.nissim}@rug.nl, {lorenzo.demattei}@di.unipi.it

Abstract

We study how words are used differently in two Italian newspapers at opposite ends of the political spectrum by training embeddings on one newspaper's corpus, updating the weights on the second one, and observing vector shifts. We run two types of analysis, one top-down, based on a pre-selection of frequent words in both newspapers, and one bottom-up, on the basis of a combination of the observed shifts and relative and absolute frequency. The analysis is specific to this data, but the method can serve as a blueprint for similar studies.

1 Introduction and Background

Different newspapers, especially if positioned at opposite ends of the political spectrum, can render the same event in different ways. In Example (1), both headlines are about the leader of the Italian political movement “Cinque Stelle” splitting up with his girlfriend, but the Italian left-oriented newspaper *la Repubblica*¹ (rep in the examples) and right-oriented *Il Giornale*² (gio in the examples) describe the news quite differently. The news in Example (2), which is about a baby-sitter killing a child in Moscow, is also reported by the two newspapers mentioning and stressing different aspects of the same event.

- (1) rep La ex di Di Maio: “E’ stato un amore intenso ma non abbiamo retto allo stress della politica”
[en: *The ex of Di Maio: “It’s been an intense love relationship, but we haven’t survived the stress of politics”*]

gio Luigino single, è finita la Melodia
[en: *Luigino single, the Melody is over*]

- (2) rep Mosca, “la baby sitter omicida non ha agito da sola”
[en: *Moscow, “the killer baby-sitter has not acted alone”*]

gio Mosca, la donna killer: “Ho decapitato la bimba perché me l’ha ordinato Allah”
[en: *Moscow, the killer woman: “I have beheaded the child because Allah has ordered me to do it”*]

Often though, the same words are used, but with distinct nuances, or in combination with other, different words, as in Examples (3)–(4):

- (3) rep Usa: agente uccide un nero disarmato e immobilizzato
[en: *Usa: policeman kills an unarmed and immobilised black guy*]

gio Oklahoma, poliziotto uccide un nero disarmato: “Ho sbagliato pistola”
[en: *Oklahoma: policeman kills an unarmed black guy: “I used the wrong gun”*]

- (4) rep Corte Sudan annulla condanna, Meriam torna libera
[en: *Sudan Court cancels the sentence, Meriam is free again*]

gio Sudan, Meriam è libera: non sarà impiccata perché cristiana
[en: *Sudan: Meriam is free: she won’t be hanged because Christian*]

In this work we discuss a method to study how the same words are used differently in two sources, exploiting vector shifts in embedding spaces.

The two embeddings models built on data coming from *la Repubblica* and *Il Giornale* might contain interesting differences, but since they are separate spaces they are not directly comparable. Previous work has encountered this issue from a diachronic perspective: when studying meaning shift in time, embeddings built on data from different periods would encode different usages, but they need to be comparable. Instead of constructing separate spaces and then aligning them

Copyright © 2019 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0)

¹<https://www.repubblica.it>

²<http://www.ilgiornale.it>

(Hamilton et al., 2016b), we adopt the method used by Kim et al. (2014) and subsequently by Del Tredici et al. (2016) for Italian, whereby embeddings are first trained on a corpus, and then updated with a new one; observing the shifts certain words undergo through the update is a rather successful method to proxy meaning change.

Rather than across time, we update embeddings across sources which are identical in genre (newspapers) but different in political positioning. Specifically, we train embeddings on articles coming from the newspaper *La Repubblica* (leaning left) and update them using articles coming from the newspaper *Il Giornale* (leaning right). We take the observed shift of a given word (or the shift in distance between two words) as a proxy for a difference in usage of that term, running two types of analysis. One is top-down, and focuses on a set of specific words which are frequent in both corpora. The other one is bottom-up, focusing on words that result potentially interesting on the basis of measures that combine the observed shift with both relative and absolute frequency. As a byproduct, we also learn something about the interaction of shifts and frequency.

2 Data

We scraped articles from the online sites of the Italian newspapers *la Repubblica*, and *Il Giornale*. We concatenated each article to its headline, and obtained a total of 276,120 documents (202,419 for *Il Giornale* and 73,701 for *la Repubblica*).

For training the two word embeddings, though, we only used a selection of the data. Since we are interested in studying how the usage of the same words changes across the two newspapers, we wanted to maximise the chance of articles from the two newspapers being on the same topic. Thus, we implemented an automatic alignment, and retained only the aligned news for each of the two corpora. All embeddings are trained on such aligned news.

2.1 Alignment

We align the two datasets using the whole body of the articles. We compute the tf-idf vectors for all the articles of both newspapers and create subsets of relevant news filtering by date, i.e. considering only news that were published in the range of three days before and after of one another. Once this subset is extracted, we compute cosine similarities for all news in one corpus and in the other

corpus using the tf-idf vectors, we rank them and then filter out alignments whose cosine similarity is under a certain threshold. The threshold should be chosen taking into consideration a trade-off between keeping a sufficient number of documents and quality of alignment. In this case, we are relatively happy with a good but not too strict alignment, and after a few tests and manual checks, we found that threshold of 0.185 works well in practice for these datasets, yielding a good balance between correct alignments and news recall. Table 1 shows the size of the aligned corpus in terms of number of documents and tokens.

newspaper	#documents	#tokens
<i>la Repubblica</i>	31,209	23,038,718
<i>Il Giornale</i>	38,984	18,584,121

Table 1: Size of the aligned corpus.

2.2 Shared lexicon

If we look at the most frequent content words in the datasets (Figure 1), we see that they are indeed very similar, most likely due to the datasets being aligned based on lexical overlap.

This selection of frequent words already constitutes a set of interesting tokens to study for their potential usage shift across the two newspapers. In addition, through the updating procedure that we describe in the next section, we will be able to identify which words appear to undergo the heaviest shifts from the original to the updated space, possibly indicating a substantial difference of use across the two newspapers.

2.3 Distinguishability

Seeing that frequent words are shared across the two datasets, we want to ensure that the two datasets are still different enough to make the embeddings update meaningful.

We therefore run a simple classification experiment to assess how distinguishable the two sources are based on lexical features. Using the scikit-learn implementation with default parameters (Pedregosa et al., 2011), we trained a binary linear SVM to predict whether a given document comes from *la Repubblica* or *Il Giornale*. We used ten-fold cross-validation over the aligned dataset with only word n-grams 1-2 as features and obtained an overall accuracy of 0.796, and 0.794 and 0.797 average precision and recall, respectively.

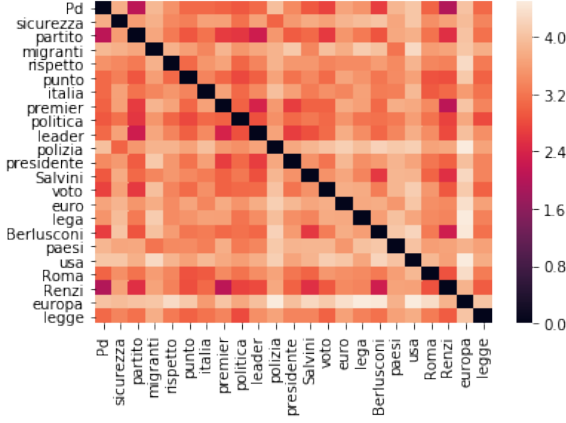


Figure 3: Distance matrix between a small set of high frequency words on *la Repubblica*. The lighter the color the larger the distance.

However, frequency plays an important role, too (Schnabel et al., 2015). To account for this, we explore the impact of both absolute and relative frequency for each word w . We take the overall frequency of a word summing the individual occurrences of w in the two corpora ($total_w$). We also take the difference between the relative frequency of a word in the two corpora, as this might be influencing the shift. We refer to this difference as gap_w , and calculate it as in Equation 1.

$$(1) \quad gap_w = \log\left(\frac{freq_w^r}{|r|}\right) - \log\left(\frac{freq_w^g}{|g|}\right)$$

A negative gap_w indicates that the word is relatively more frequent in *Il Giornale* than in *la Repubblica*, while a positive value indicates the opposite. Words whose relative frequency is similar in both corpora exhibit values around 0.

We observe a tiny but significant negative correlation between $total_w$ and $shift_w$ (-0.093 , $p < 0.0001$), indicating that the more frequent a word, the less it is likely to shift. In Figure 2 we see all the dark dots (most frequent words) concentrated at the bottom of the scatter plot (lower shifts).

However, when we consider gap_w and $shift_w$, we see a more substantial negative correlation (-0.306 , $p < 0.0001$), suggesting that the gap has an influence on the shift: the more negative the gap, the higher the shift. In other words, the shift is larger if a word is relatively more frequent in the corpus used to update the embeddings.

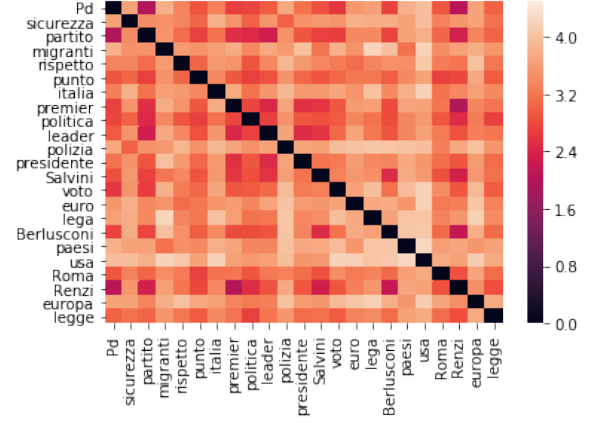


Figure 4: Distance matrix between a small set of high frequency words after updating with *Il Giornale*. The lighter the color the larger the distance.

4 Analysis

We use the information that derives from having the original *spaceR* and the updated *spaceRG* to carry out two types of analysis. The first one is top-down, with a pre-selection of words to study, while the second one is bottom-up, based on measures combining the shift and frequency.

4.1 Top-down

As a first analysis, we look into the most frequent words in both newspapers and study how their relationships change when we move from *spaceR* to *spaceRG*. The words we analyse are the union of those reported in Figure 1. Note that in this analysis we look at pairs of words at once, rather than at the shift of a single word from one space to the next. We build three matrices to visualise the distance between these words.

The first matrix (Figure 3) only considers *SpaceR*, and serves to show how close/distant the words are from one another in *la Repubblica*. For example, we see that “partito” and “Pd”, or “premier” and “Renzi” are close (dark-painted), while “polizia” and “europa” are lighter, thus more distant (probably used in different contexts).

In Figure 4 we show a replica of the first matrix, but now on *SpaceRG*; this matrix now let’s us see how the distance between pairs of words has changed after updating the weights. Some vectors are farther than before and this is visible by the lighter color of the figure, like “usa” and “lega” or “italia” and “usa”, while some words are closer like “Berlusconi” and “europa” or “europa” and “politica” which feature darker colour. Specific

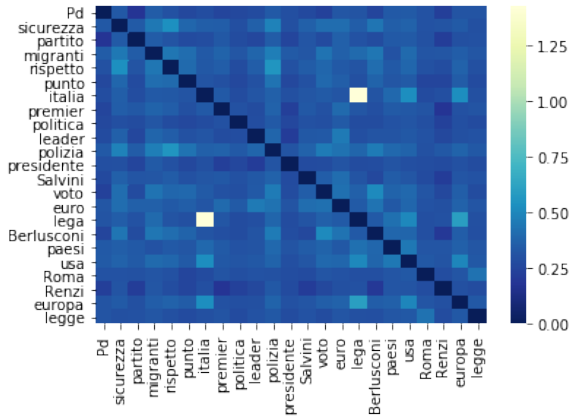


Figure 5: Difference matrix between embeddings from *spaceR* and *spaceRG* normalised with the logarithm of the absolute frequency difference in *spaceRG*. The lighter the colour, the larger the distance between pairs of words.

analysis of the co-occurrences of such words could yield interesting observations on their use in the two newspapers.

In order to better observe the actual difference, the third matrix shows the shift from *spaceR* to *spaceRG*, normalised by the logarithm of the absolute difference between the $total_{w1}$ and $total_{w2}$ (Figure 5).³ Lighter word-pairs shifted more, thus suggesting different contexts and usage, for example “italia” and “lega”. Darker pairs, on the other hand, such as “Pd”-“Partito” are also interesting for deeper analysis, since their joint usage is likely to be quite similar in both newspapers.

4.2 Bottom-up

Differently from what we did in the top-down analysis, here we do not look at how the relationship between pairs of pre-selected words changes, rather at how a single word’s usage varies across the two spaces. These words arise from the interaction of *gap* and *shift*, which yields various scenarios. Words with a large negative gap (relative frequency higher in *Il Giornale*) are likely to shift more, but it’s probably more of an effect due to increased frequency than a genuine shift. Words that have a high gap (occurring relatively less in *Il Giornale*) are likely to shift less, most likely since adding a few contexts might not cause much shift.

The most interesting cases are words whose

³Note that this does not correspond exactly to the *gap* measure in Eq. 1 since we are considering the difference between two words rather than the difference in occurrence of the same word in the two corpora.

relative frequency does not change in the two datasets, but have a high shift. Zooming in on the words that have small gaps ($-0.1 < gap_w < 0.1$), will provide us with a set of potentially interesting words, especially if they have a shift higher than the average shift. We also require that words obeying the previous constraints occur more than the average word frequency over the two corpora. Low frequency words are in general less stable (Schnabel et al., 2015), suggesting that shifts for the latter might not be reliable. High frequency words shift globally less (cf. Figure 2), so a higher than average shift could be meaningful.

Figure 6 shows the plot of words that have more or less the same relative frequency in the two newspapers ($-0.1 < gap > 0.1$ and an absolute cumulative frequency higher than average), and we therefore infer that their higher than average shift is mainly due to usage difference. Some comments are provided next to the plot.

These words can be the focus of a dedicated study, and independently of the specific observations that we can make in this context, this method can serve as a way to highlight the hotspot words that deserve attention in a meaning shift study.

4.3 A closer look at nearest neighbours

As a last, more qualitative, analysis, one can inspect how the nearest neighbours of a given word of interest change from one space to the next. In our specific case, we picked a few words (deriving them from the top-down, thus most frequent, and bottom-up selections), and report in Table 2 their top five nearest neighbours in *SpaceR* and in *SpaceRG*. As in most analyses of this kind, one has to rely quite a bit on background and general knowledge to interpret the changes. If we look at “Renzi”, for example, a past Prime Minister from the party close to the newspaper “la Repubblica”, we see that while in *SpaceR* the top neighbours are all members of his own party, and the party itself (“Pd”), in *SpaceRG* politicians from other parties (closer to “Il Giornale”) get closer to Renzi, such as Berlusconi and Alfano.

5 Conclusions

We experimented with using embeddings shifts as a tool to study how words are used in two different Italian newspapers. We focused on a pre-selection of high frequency words shared by the two newspapers, and on another set of words which were

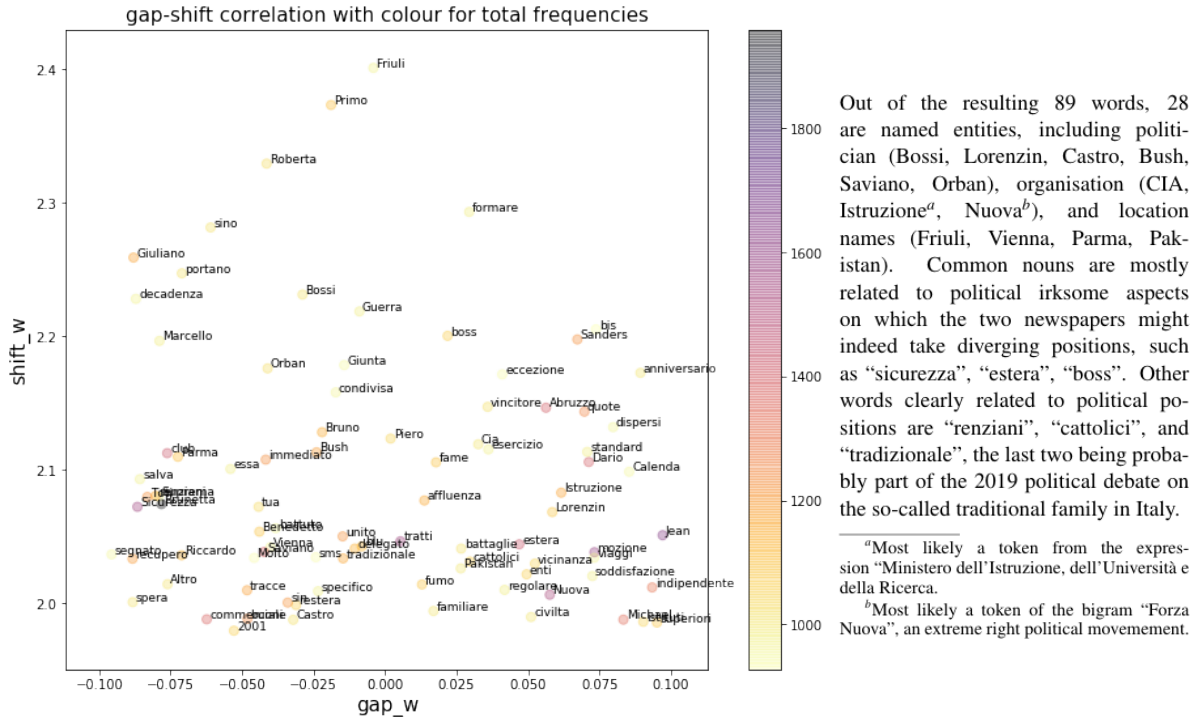


Figure 6: Gap-Shift scatter plot like in Figure 2, zoomed in the gap region -0.1 - 0.1 and shift greater than 1.978 (average shift). Only words with cumulative frequency higher than average frequency are plotted.

Table 2: A few significant words and their top 5 nearest neighbours in *SpaceR* and *SpaceRG*.

<i>SpaceR</i>	<i>SpaceRG</i>
“migranti” [<i>en: migrants</i>]	
barconi [<i>large boats</i>] (0.60)	eritrei [<i>Eritreans</i>] (0.61)
naufraghi [<i>castaways</i>] (0.57)	Lampedusa [] (0.60)
disperati [<i>wretches</i>] (0.56)	accoglienza [<i>hospitality</i>] (0.59)
barcone [<i>large boat</i>] (0.55)	Pozzallo [] (0.58)
carrette [<i>wrecks</i>] (0.53)	extracomunitari [<i>non-European</i>] (0.57)
“Renzi” [past Prime Minister]	
Orfini [] (0.65)	premier [] (0.60)
Letta [] (0.64)	Nazareno [] (0.59)
Cuperlo [] (0.63)	Berlusconi [] (0.58)
Pd [] (0.62)	Cav [] (0.57)
Bersani [] (0.61)	Alfano [] (0.56)
“politica” [<i>en: politics</i>]	
leadership [] (0.65)	tecnocrazia [<i>technocracy</i>] (0.60)
logica [<i>logic</i>] (0.64)	democrazia [<i>democracy</i>] (0.59)
miri [<i>aspire to</i>] (0.63)	partitica [<i>of party</i>] (0.58)
ambizione [<i>ambition</i>] (0.62)	democratica [<i>democratic</i>] (0.57)
potentati [<i>potentates</i>] (0.61)	legalità [<i>legality</i>] (0.56)

highlighted as potentially interesting through a newly proposed methodology which combines observed embeddings shifts and relative and absolute frequency. Most differently used words in the two newspapers are proper nouns of politically active individuals as well as places, and concepts that are highly debated on the political scene.

Beside the present showcase, we believe this methodology can be more in general used to highlight which words might deserve deeper, dedicated analysis when studying meaning change.

One aspect that should be further investigated is the role played by the methodology used for aligning and/or updating the embeddings. As an alternative to what we proposed, one could employ different strategies to manipulate embedding spaces towards highlighting meaning changes. For example, Rodda et al. (2016) exploited Representational Similarity Analysis (Kriegeskorte and Kievit, 2013) to compare embeddings built on different spaces in the context of studying diachronic semantic shifts in ancient Greek. Another interesting approach, still in the context of diachronic meaning change, but applicable to our datasets, was introduced by Hamilton et al. (2016a), who use both a global and a local neighborhood measure of semantic change to disentangle shifts due to cultural changes from purely linguistic ones.

Acknowledgments

We would like to thank the Center for Information Technology of the University of Groningen for providing access to the Peregrine high performance computing cluster.

References

- Marco Del Tredici, Malvina Nissim, and Andrea Zaninello. 2016. Tracing metaphors in time through self-distance in vector spaces. In *Proceedings of the Third Italian Conference on Computational Linguistics (CLiC-it 2016)*.
- William L Hamilton, Jure Leskovec, and Dan Jurafsky. 2016a. Cultural shift or linguistic drift? comparing two computational measures of semantic change. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing*, volume 2016, page 2116. NIH Public Access.
- William L Hamilton, Jure Leskovec, and Dan Jurafsky. 2016b. Diachronic word embeddings reveal statistical laws of semantic change. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1489–1501.
- Yoon Kim, Yi-I Chiu, Kentaro Hanaki, Darshan Hegde, and Slav Petrov. 2014. Temporal analysis of language through neural language models. In *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science*, pages 61–65, Baltimore, MD, USA, June. Association for Computational Linguistics.
- Nikolaus Kriegeskorte and Rogier A Kievit. 2013. Representational geometry: integrating cognition, computation, and the brain. *Trends in cognitive sciences*, 17(8):401–412.
- Tomas Mikolov, Kai Chen, G.s Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *Proceedings of Workshop at ICLR*, 2013, 01.
- F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.
- Radim Řehůřek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta, May. ELRA. <http://is.muni.cz/publication/884893/en>.
- Martina Astrid Rodda, Marco SG Senaldi, and Alessandro Lenci. 2016. Panta rei: Tracking semantic change with distributional semantics in ancient greek. In *CLiC-it/EVALITA*.
- Tobias Schnabel, Igor Labutov, David Mimno, and Thorsten Joachims. 2015. Evaluation methods for unsupervised word embeddings. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 298–307.
- Peter D Turney and Patrick Pantel. 2010. From frequency to meaning: Vector space models of semantics. *Journal of artificial intelligence research*, 37:141–188.

Suitable Doesn't Mean Attractive.

Human-Based Evaluation of Automatically Generated Headlines

Michele Cafagna^{1,3}, Lorenzo De Mattei^{1,2,3}, Davide Bacciu¹ and Malvina Nissim³

¹Department of Computer Science, University of Pisa, Italy

²ItaliaNLP Lab, ILC-CNR, Pisa, Italy

³CLCG, University of Groningen, The Netherlands

{m.cafagna,m.nissim}@rug.nl, {lorenzo.demattei,bacciu}@di.unipi.it

Abstract

We train three different models to generate newspaper headlines from a portion of the corresponding article. The articles are obtained from two mainstream Italian newspapers. In order to assess the models' performance, we set up a human-based evaluation where 30 different native speakers expressed their judgment over a variety of aspects. The outcome shows that (i) pointer networks perform better than standard sequence to sequence models, creating mostly correct and appropriate titles; (ii) the suitability of a headline to its article for pointer networks is on par or better than the gold headline; (iii) gold headlines are still by far more inviting than generated headlines to read the whole article, highlighting the contrast between human creativity and content appropriateness.

1 Introduction and Background

Progress in language generation has made it really hard to tell if a text is written by a human or is machine-generated. The recently developed GPT-2 transformer-based language model (Radford et al., 2019), when prompted with an arbitrary input, is able to generate synthetic texts which are impressively human-like. But what makes generated text *good* text?

We investigate this question in the context of automatically generated news headlines.¹

Copyright © 2019 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0)

¹A growing interest in headline generation is witnessed also in the organisation of a multilingual shared task at RANLP 2019, using Wikipedia data: <http://multiling.iit.demokritos.gr/pages/view/1651/task-headline-generation>

Headlines could be seen as very short summaries, so that one could use evaluation methods typical of summarisation (Gatt and Krahmer, 2018), but they are in fact a very special kind of summaries. In addition to being suitable in terms of content, newspaper titles must also be inviting towards reading the whole article. A model that, given an article, learns how to generate its title must then be able to cover both the summarisation as well as the luring aspect.

We collect articles from Italian newspapers online, and generate their headlines automatically. In contrast to the feature-rich approach of Colmenares et al. (2015), which requires substantial linguistic preprocessing for feature extraction, we rely on recent developments in language modelling, and train three different sequence-to-sequence models that learn to generate a headline given (a portion of) its article. We compare these generated headlines to one another and to the gold headline through a series of human-based evaluations which take several aspects into account, ranging from grammatical correctness to attractiveness towards reading the full article. The factors we measure are in line with the requirements for human-based evaluation mentioned by Gatt and Krahmer (2018), and are useful since it is known that standard metrics based on lexical overlap are not accurate indicators for the goodness of generated text (Liu et al., 2016).

Contributions We offer three main contributions: (i) a model which generates headlines from Italian news articles and which we make publicly available; (ii) a framework for human-based evaluation of generated headlines, which can serve as a blueprint for the evaluation of other types of generated texts; (iii) insights on the performance of different headline generators, and on the distinction between the concepts of suitable and attractive when evaluating headlines.

model	example generated headlines
s2s	Al Qaida : “ L’ Europa non è un pericolo per i nostri fratelli ” la Samp batte la Sampdoria e la Samp non si ferma mai
pn	Teramo , bimbo di sei anni muore sotto gli occhi dei genitori mentre faceva il bagno Brescia , boa constrictor : sequestrati due metri e mezzo in un anno di animali
pnc	Argentina , Obama : “ Paladino dei poveri e dei piu vulnerabili ” . E il Papa si divide Cagliari , cane ha preferito rimandare il cane dal veterinario di Santa Margherita di famiglia

Table 1: Examples of headlines generated by the three models.

2 Task, Data, and Settings

The task is conceptually straightforward: given an article, generate its headline. Luckily, correspondingly straightforward is obtaining training and test data. We scraped the websites of two major Italian newspapers, namely *La Repubblica*² and *Il Giornale*³, collecting a total of approximately 275,000 article-headline pairs. The two newspapers are not equally represented, with *Il Giornale* covering 70% of the data.

After removing some duplicates, and instances featuring headlines shorter than 20 characters (which are typically commercials), we were left with a total of 253,543 pairs, which we split into training (177,480), validation (50,709), and test (25,354) sets, preserving in each the proportion of the two newspapers.

We used the training and validation sets to develop three different models that learn to generate a headline given an article. To keep training computationally manageable, each article was truncated after the first 500 tokens.⁴ As an alternative to keep the text short but maximally informative, we also experimented with selecting relevant portions of the articles using the TextRank algorithm, a graph-model that ranks sentences in a text according to their importance (Mihalcea and Tarau, 2004). However, preliminary experiments on our validation set did not seem to yield better results than just selecting the first N-tokens of an article. Also, using TextRank would make a less natural comparison to the settings used for the human evaluation (see Section 4), so we did not pursue this option further.⁵

²<https://www.repubblica.it>

³<http://www.ilgiornale.it>

⁴We do not control for sentence endings, so the last sentence of each truncated article might get truncated.

⁵Each article is also equipped with a short summary, often complementary to the title in content. We do not use this

3 Models

The models that we trained and evaluated are described below. In Table 1 we show two generated examples for each of the three models to give an idea of their output.

Sequence-to-Sequence with Attention (S2S)

We used a sequence-to-sequence model (Sutskever et al., 2014) with attention (Bahdanau et al., 2014) with the configuration used by See et al. (2017) but we used a bidirectional instead of a unidirectional layer. This choice applies to all the models we used. The final configuration is 1 bidirectional encoder-decoder layer with 256 LSTM cells each, no dropout and shared embeddings with size 128; the model is optimised with Adagrad with learning rate 0.15 and gradient clipped (Mikolov, 2012) to a maximum magnitude of 2. We experimented also with a version using pretrained Italian embeddings, but since some preliminary evaluation didn’t show better results, we eventually decided not to use this other model.

Pointer Generator Network (PN) The hybrid pointer-generator network architecture See et al. (2017) can copy words from the source text via a *pointing mechanism*, and generate words from a fixed vocabulary. This allows for a better handling of out-of-vocabulary words, providing accurate reproduction information, while retaining the ability to reproduce novel words. The base architecture is a sequence-to-sequence model, except for the pointing mechanism and for the fact that the copy attention parameters are shared with the regular attention. An additional layer (so called *bridge* (Klein et al., 2017)) is trained between the encoder and the decoder and is fed with the latest encoder states. Its purpose is to learn to generate

text in the current experiments, but plan to exploit it in future work.

initial states for the decoder instead of initialising them directly with the latest encoder states.

Pointer Generator Network with Coverage (PNC) This model is basically a Pointer Generator Network with an additional coverage attention mechanism that is intended to overcome the copying problem typical of sequence-to-sequence models (See et al., 2017). This is basically a vector, computed by summing up all the attention distributions over all previous decoder timesteps. This unnormalised distribution over the document words is expected to represent the degree of coverage that the words have received from the attention mechanism until then. This vector, called *coverage vector*, is used to penalise the attention over already generated words, to minimise the risk of generating repetitive text.

4 Evaluation

Evaluating automatically generated text is non-trivial. Given that many different generated texts can be correct, existing measures are usually deemed insufficient (Liu et al., 2016). The problem is even more acute for headline generation, since due to their nature and function, simple content evaluation based on word overlap is most likely not exhaustive. Human-based evaluation could provide a richer picture.

When discussing human-based (intrinsic) evaluation of summarisation models, Gatt & Krahmer (2018) mention two core aspects: *linguistic fluency or correctness*, and *adequacy or correctness relative to the input*, in terms of the system’s rendition of the content. These also relate to the aspects examined in the context of evaluating the generation of the final sentence of a story, such as *grammaticality*, *(logical) consistency*, and *context relevance* (Li et al., 2018).

We took these factors into consideration when designing our evaluation settings. Since headlines must also carry some “attraction” factor to read the whole article, we included this aspect as well.

4.1 Settings

We call a case each set of an article and its four corresponding headlines to be evaluated, namely the three automatically generated ones, and the original (gold) title.

We prepared an evaluation form⁶, which in-

⁶An example can be found here: <https://forms.gle/MB31uEGT856af2MP7>

cluded five different questions for each case (see Figure 1). Each subject could see the four headlines and answer questions Q1–Q3. The corresponding article, in the truncated form that was also seen in training by the models, was only shown to the subjects after Q3, and they would then answer Q4–Q5. This choice was made in order to ensure that first questions were answered on the basis of the headlines only, especially for the validity of Q3. The order in which gold and generated titles were shown was randomised, though it was the same for each case for all participants.

Each form comprised 20 cases to evaluate, and was sent to 3 participants. We created 10 different forms, thus obtaining judgements for 200 total cases with 30 different participants (600 separate judgements). The participants are all native speakers of Italian, and balanced for gender (15F/15M). We also aimed at a wide range of ages (17–77) and education levels (middle school diploma to PhD). This variety was sought in order to prevent as much as possible judgements that are based too strongly on personal biases, taste, and familiarity with specific topics over others.

The headlines used for this evaluation exercise were randomly selected from the test set. When extracting them though, we excluded all cases where at least one model produced a headline containing at least an unknown word (represented with the special token $\langle UNK \rangle$), since this would make the headline look too weird and not much comprehensible. This led to excluding approximately 50% of the samples. The model with the highest proportion of headlines with at least one UNK was the S2S (37%), followed by the PNC (31%), and the PN (30.2%). In terms of topics, random picking ensured a variety of topics; manual inspection anyway showed that most news were mainly about chronicle facts, and international politics.

4.2 Analysis

We discuss the results in detail for questions Q1, Q3, Q4, Q5. For Q2, we simply note that the most similar in content are always the two pointer networks, and the most dissimilar are all three pairs that involve the gold headlines. This suggests that human titles focus on aspects of the article that are different from those picked by the generator, most likely as humans can abstract away from the actual text and use much more creativity.

The four titles are shown (repeated for each question below)

- A. Usa , la fabbrica del vetro d' aria per il telefono d' aria in Usa
- B. Se il lavoro va ai robot : un automa vale sei operai
- C. Usa , Trump : " Trump si difende l' occupazione e l' economia nazionale "
- D. Usa , la beffa del condizionatore d' aria " made in Usa " : " Ecco come si difende "

And the following questions are then asked:

[at this stage the subjects only see titles, without the article]	
Q1. Questi titoli sono scritti correttamente?	yes,no for each
Q2. Secondo te, questi titoli parlano dello stesso articolo?	yes,no for pairs of titles
Q3. Quale di questi titoli ti invoglia maggiormente a leggere l'intero articolo?	pick one
[now the subjects also see the (truncated) article]	
<p>New York . Chiamiamola la beffa del condizionatore d' aria " made in Usa " . La marca è Carrier , filiale della multinazionale United Technologies . Un caso ormai celebre , che Donald Trump addita come un esempio della sua azione efficace a tutela della classe operaia . A novembre , appena eletto presidente (ma non ancora in carica) , Trump si occupa dello " scandalo Carrier " : vogliono chiudere una fabbrica di condizionatori a Indianapolis per trasferirla in Messico , delocalizzando a Sud del confine 800 posti di lavoro . Il presidente - eletto fa fuoco e fiamme , chiama il chief executive dell' azienda . Forse interviene la casa madre , United Technologies , che ha grosse commesse per l' esercito e non vuole inimicarsi il neo - presidente . Sta di fatto che Carrier cede alle pressioni , fa dietrofront : la fabbrica resta sul suolo Usa , nello Stato dell' Indiana . Tripudio di Trump che canta vittoria via Twitter : " Ecco come si difende l' occupazione e l' economia nazionale " . Passano i mesi e il caso viene dimenticato . Fino a quando il chief executive Greg Hayes rivela ai sindacati che i 16 milioni di investimento nella sede di Indianapolis vanno tutti in robotica , automazione : " Alla fine ci saranno meno posti di prima . Dobbiamo ridurre i costi , per essere competitivi " . La morale è crudele , la vittoria di Trump si [...]</p>	
Q4. Ritieni che il titolo sia appropriato all'articolo?	yes,no for each
Q5. Quale ti sembra più adatto? Ordinali	rank 1-4

Figure 1: Sample evaluation case. Subjects are presented with the gold and generated headlines in random order, and must answer a progression of questions, without and with seeing the article. Q1 targets correctness, Q2 targets the similarity in topic focus, Q3 targets attractiveness, Q4 and Q5 target appropriateness (absolute, and relative to one another). In this example, A=s2s, B=gold, C=pnc, D=pn.

Grammatical Correctness (Q1) When asked to evaluate whether the headlines were written correctly, the participants assessed all headlines as correct more frequently than not correct, with Gold and PN having the best ratio of yes vs no (Figure 2). What is, however, interesting is that even Gold headlines are frequently judged as not correct, implying that either the participants were very strict, or correctness is not a necessary or particularly typical feature of newspaper headlines. While it is important for us to assess how well the generators perform also in terms of well-formed sequences, if (grammatical) correctness is not strictly a property of newspaper headlines, this

evaluation question might have to be formulated differently. In any case, among the models, for the current question, the PN behaves almost on par with the gold headlines.

Attractiveness (Q3) In the large majority of the cases, the gold headline was chosen as the most inspiring for reading the whole article (Figure 3). Among the models, the headlines generated by the PN is mostly chosen, followed by the PNC, and lastly by the S2S. Such results suggest that there is something in the way experts create headlines, most likely related to human creativity, rhetoric and communication strategies, which systems are

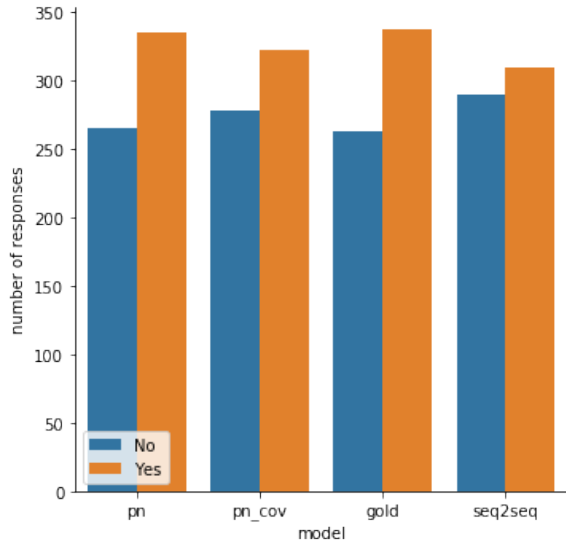


Figure 2: Correctness judgments (Q1)

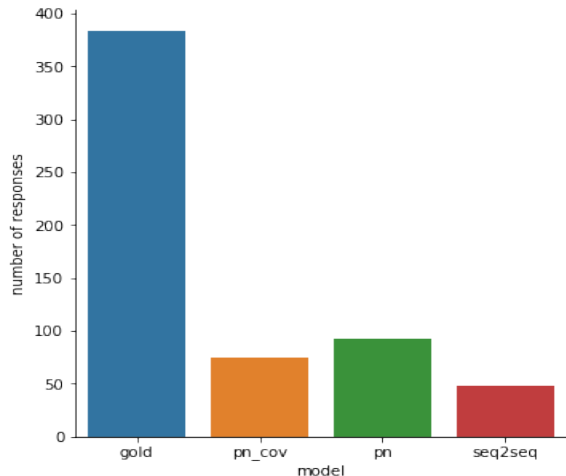


Figure 3: Attractiveness judgements (Q3)

not yet able to reproduce. Additionally, some on-line newspapers’ business models can be heavily clickbait-based, causing headlines to be more sensational than faithful to the article’s actual contents.

Suitability (Q4-Q5) There are two results to be analysed in the context of assessing how appropriate a headline is with respect to its article. In terms of a binary evaluation for each headline (Figure 4, left), in all cases, including gold, the headline is deemed not appropriate more than the times is deemed appropriate. In the case of gold, this could be due to the fact that excessive creativity to make the title attractive can make it less adherent to the actual content. In the case of the generated headlines, they might just not be good enough.

	G	S2S	PN	PNC	tot
correctness	0.439	0.427	0.345	0.337	0.387
attractiveness	–	–	–	–	0.120
suitability	0.349	0.354	0.374	0.313	0.348
suitability-rank	0.444	0.364	0.339	0.398	0.389

Table 2: Krippendorff’s alpha scores for the human annotations. The rightmost column shows the agreement over all systems plus gold headlines.

The rank shows a possibly unexpected trend (Figure 4, right side). The headline chosen as most appropriate (ranked 1st) is most of the times the one produced by the PN model, even more so than the gold. Not only, the gold is also the headline that features last (ranked 4th, thus least suitable) more than any of the other titles. This is reflected in the average rank (see caption of Figure 4), as the gold headline comes in last, and the PN-generated title is comparatively the most preferred.

4.3 Agreement

Given that we obtained three separate judgments per case, in addition to the separate evaluations, we can also assess how much the subjects agree with one another. Table 2 shows the values for Krippendorff’s alpha over all of the annotated aspects. Low scores suggest that the task is highly subjective, and this is especially true for the evaluation of how attractive a headline is towards reading the whole article. Possibly surprising is the score regarding the evaluation of the headline’s correctness, which could be viewed as a more objective feature to assess. Such relatively low score could be due to the vagueness of Q1, in combination with the nature of headlines, which even in their human version might be formulated in ways that do not necessarily abide to grammatical rules.

5 Conclusions

The quality of three different sequence-to-sequence models that generate headlines starting from an article was comparatively assessed through human judgement, which we contextually used to evaluate the original headlines as well. The best system is a pointer network model, with correctness judgements on par with the gold headlines. Evaluating the generated output on different levels, especially attractiveness, which typically characterises news headlines, uncovered an interesting aspect: gold headlines appear to be the most attractive to read the whole article, but are not con-

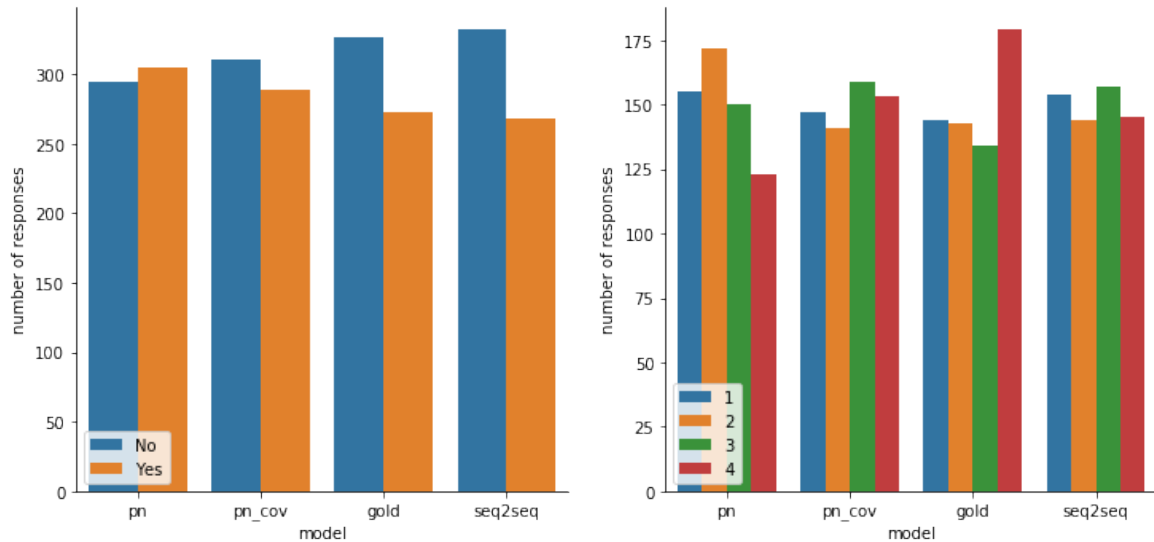


Figure 4: Suitability. Left: suitability judgment for each headline (yes/no). Right: headlines are ranked according to most (1) to least (4) appropriate for each corresponding article. Average ranking: PN=2.401; Seq2Seq=2.488; PN_C=2.530; GOLD=2.580

sidered the most suitable, on the contrary, they are judged as the most unsuitable of all. Therefore, when automatically generating headlines, just relying on content might never lead us to titles that are human-like and attractive enough for people to read the article. This should be considered in any future work on news headline generation. At the evaluation stage, it would also be beneficial to involve professional journalists. A first contact with one of the newspapers at the early stages of our evaluation experiments did not yet yield any concrete collaboration, but expert judgement on the quality of the generated headlines is something we would like to include in the future.

One aspect that we have not explicitly considered in our experiments is that the headlines come from different newspapers (positioned at opposite ends of the political spectrum), and can carry newspaper-specific characteristics. Robust headline generation should consider this, too.

Acknowledgments

We are deeply grateful to all of the participants to our evaluation. We also would like to thank the Center for Information Technology of the University of Groningen for providing access to the Peregrine high performance computing cluster. A heartfelt thank you also to Angelo Basile, with whom we discussed both theoretical and implementation aspects of this work.

References

- Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.
- Carlos A Colmenares, Marina Litvak, Amin Mantrach, and Fabrizio Silvestri. 2015. HEADS: Headline generation as sequence prediction using an abstract feature-rich space. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 133–142.
- Albert Gatt and Emiel Krahmer. 2018. Survey of the state of the art in natural language generation: Core tasks, applications and evaluation. *Journal of Artificial Intelligence Research*, 61:65–170.
- Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander M. Rush. 2017. OpenNMT: Open-source toolkit for neural machine translation. In *Proc. ACL*.
- Zhongyang Li, Xiao Ding, and Ting Liu. 2018. Generating reasonable and diversified story ending using sequence to sequence model with adversarial training. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1033–1043.
- Chia-Wei Liu, Ryan Lowe, Iulian Serban, Mike Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How not to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2122–2132.

- Rada Mihalcea and Paul Tarau. 2004. Texttrank: Bringing order into text. In *Proceedings of the 2004 conference on empirical methods in natural language processing*.
- Tomáš Mikolov. 2012. Statistical language models based on neural networks. *Presentation at Google, Mountain View, 2nd April*, 80.
- Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8).
- Abigail See, Peter J Liu, and Christopher D Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083.
- Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112.

There and Back Again: Cross-Lingual Transfer Learning for Event Detection

Tommaso Caselli, Ahmet Üstün

Rikjuniversiteit Groningen, Groningen, The Netherlands

{t.caselli|a.ustun}@rug.nl

Abstract

English. In this contribution we investigate the generalisation abilities of a pre-trained multilingual Language Model, namely Multilingual BERT, in different transfer learning scenarios for event detection and classification for Italian and English. Our results show that zero-shot models have satisfying, although not optimal, performances in both languages (average F1 higher than 60 for event detection *vs.* average F1 ranging between 40 and 50 for event classification). We also show that adding extra fine-tuning data of the evaluation language is not simply beneficial but results in better models when compared to the corresponding non zero-shot transfer ones, achieving highly competitive results when compared to state-of-the-art systems.

1 Introduction

Recently pre-trained word representations encoded in Language Models (LM) have gained lot of popularity in Natural Language Processing (NLP) thanks to their ability to encode high level syntactic-semantic language features and produce state-of-the-art results in various tasks, such as Named Entity Recognition (Peters et al., 2018), Machine Translation (Johnson et al., 2017; Ramachandran et al., 2017), Text Classification (Eriguchi et al., 2018; Chronopoulou et al., 2019), among others. These models are pre-trained on large amounts of unannotated text and then fine-tuned using the induced LM structure to generalise over specific training data. Given their success in monolingual environments, espe-

cially for English, there has been a growing interest in the development of *cross-lingual* as well as *multilingual* representations (Vulić and Moens, 2015; Ammar et al., 2016; Conneau et al., 2018; Artetxe et al., 2018) to investigate different cross-lingual transfer learning scenarios, including zero-shot transfer, i.e. the direct application of a model fine-tuned using data in one language to a different test language.

Following the approach in Pires et al. (2019), in this paper we investigate the generalisation abilities of Multilingual BERT (Devlin et al., 2019)¹ on English (EN) and Italian (IT). Multilingual BERT is particularly well suited for this task because it easily allows the implementation of cross-lingual transfer learning, including zero-shot transfer.

We use event detection as our downstream task, a highly complex semantic task with a well established tradition in NLP (Ahn, 2006; Ji and Grishman, 2008; Ritter et al., 2012; Nguyen and Grishman, 2015; Huang et al., 2018). The goal of the task is to identify event mentions, i.e. linguistic expressions describing “things” that happen or hold as true in the world, and subsequently classify them according to a (pre-defined) taxonomy. The complexity of the task relies in its high dependence on the context of occurrence of the expressions that may trigger an event mention. Indeed, the *eventiveness* of an expression is prone to ambiguity because there exists a continuum between eventive and non-eventive readings in the space of event semantics (Araki et al., 2018). Such intrinsic ambiguity of event expressions challenges the generalisation abilities of stochastic models and allows to investigate advantages and limits of transfer learning approaches when semantics has a pivotal role in the resolution of a problem/task.

We explore different multi-lingual and cross-

Copyright © 2019 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

¹<https://github.com/google-research/bert>

lingual aspects of transfer learning with respect to event detection through a series of experiments, focusing on the following research questions:

RQ1 How well do Multilingual BERT fine-tuned models generalise in zero-shot transfer learning scenarios on both languages?

RQ2 Do we obtain more robust models by fine-tuning zero-shot models with additional (training) data of the evaluation language?

Our results show that Multilingual BERT obtains satisfying performances in zero-shot scenarios for the identification of event triggers (average F1 63.53 on Italian and 66.79 on English), while this is not the case for event classification (average F1 42.86 on Italian and 51.26 on English). We also show that extra fine-tuning the zero-shot models with data of the evaluation language is not just beneficial, but it actually gives better results than models fine-tuned on the corresponding test language only (i.e. fine-tuning and test in the same language), and achieves competitive results with state-of-the-art systems developed using dedicated architectures. Our code is available (<https://github.com/ahmetustun/BertForEvent>).

2 Data

We have used two corpora annotated with event information: the TempEval-3 corpus (TE3) for English (UzZaman et al., 2013) and the EVENTI corpus for Italian (Caselli et al., 2014). The corpora have been independently annotated with language specific annotation schemes, grounded on a shared metadata markup language for temporal information processing, ISO-TimeML (ISO, 2008), thus sharing definitions and tags’ names for the markable expressions. The corpora are composed by contemporary news articles² and have been developed in the context of two evaluation campaigns for temporal processing, namely TempEval-3 and EVENTI@EVALITA 2014.

Events are defined as anything that can be said to happen, or occur, or hold true, with no restriction to parts-of-speech (POS), including verbs, nouns, adjectives, and also

²We have excluded the extra test set on historical news from the Italian data set, and the automatically annotated training set from the English one.

prepositional phrases (PP). Every event mention is further assigned to one of 7 possible classes: OCCURRENCE, ASPECTUAL, PERCEPTION, REPORTING, I(NTENSIONAL) STATE, I(NTENSIONAL) ACTION, and STATE, capturing the relationship the event participates (such as factual, evidential, reported, intensional). Although semantically interoperable, one of the most relevant annotation differences that may impact the evaluation of the zero-shot models concerns the marking of modal verbs and copulas introducing event nouns, adjectives or PPs. While in English these elements are never annotated as event triggers, this is done in Italian. A detailed description of additional language specific adaptations and differences between English and Italian is reported in Caselli and Sprugnoli (2017).

Tables 1 and 2 illustrate the distribution of the annotation of events for POS (token based) and classes (event based), respectively. Both corpora, when released, did not explicitly have a development section. Following previous work (Caselli, 2018), we generated development sets by excluding from the training data all the documents that composed the test data for Italian and English in the SemEval 2010 TempEval-2 campaign (Verhagen et al., 2010).

The Italian corpus is larger than the corresponding English version, although the distribution of events, both per POS and per class, is comparable. The different distribution of the REPORTING, I_STATE, I_ACTION, and STATE classes reflects differences in annotation instructions rather than language specific characteristics. For instance, in Italian, the class REPORTING is assigned only if the event mention is an instance of a speech verb/noun (*verba/nomina dicendi*), while in English this constraint is less strict.

3 Model

Multilingual BERT (Bidirectional Encoder Representations from Transformers) shares the same framework of the monolingual English BERT_{BASE} (Devlin et al., 2019). BERT is a pre-trained LM that improves over existing fine-tuning approaches by jointly conditioning on both left and right contexts in all layers to generate pre-trained deep bidirectional representations. Multilingual BERT’s architecture contains an encoder consisting of 12 Transformer blocks with 12 self-attention heads (Vaswani et al., 2017), and

POS	TE3			EVENTI			Examples
	Train	Dev	Test	Train	Dev	Test	
Verb	8,141	393	542	11,269	193	2,426	en:run; it:correre
Noun	2,268	124	175	6,710	111	1,499	en:attack; it:attacco
Adjectives	165	8	21	610	9	118	en:(is) dormat; it:(è) dormiente
Other/PP	29	1	8	146	1	25	en:on board; it:a bordo
Total	10,603	526	746	18,735	314	4,068	

Table 1: Distribution of events per POS in each corpus per Training, Development, and Test data.

Classes	TE3			EVENTI			Examples
	Train	Dev	Test	Train	Dev	Test	
OCCURRENCE	6,530	302	466	9,041	162	1,949	en:run; it:correre
ASPECTUAL	264	33	35	446	14	107	en:start; it:inizio
PERCEPTION	79	4	2	162	2	37	en:see; it:vedere
REPORTING	1,544	67	92	714	8	149	en:say; it:dire
I.STATE	651	29	36	1,599	29	355	en:like; it:piacere
L.ACTION	827	57	47	1,476	25	357	en:attempt; it:tentare
STATE	708	34	68	4,090	61	843	en:keep; it:tenersi
Total	10,603	526	746	17,528	301	3,798	

Table 2: Distribution of event classes in each corpus per Training, Development, and Test data.

hidden size of 768.

Unlike the original BERT, Multilingual BERT is pre-trained on the concatenation of monolingual Wikipedia pages of 104 languages with a shared word piece vocabulary. One of the peculiar characteristics of this multilingual model is that it does not make use of any special marker to signal the input language, nor has any mechanism that explicitly indicates that translation equivalent pairs should have similar representations.

For the fine-tuning, we use a standard sequence tagging model. We apply a softmax classifier over each token by passing the token’s last layer of activation to the softmax layer to make a tag prediction. Since BERT’s wordpiece tokenizer can split words into multiple tokens, we take the prediction for the first token (piece) per word, ignoring the rest. No parameter tuning was performed, learning rate was set to $1e-4$, and batch size to 8.

4 Experiments

Event detection is best described as composed by two sub-tasks: first, identify if a word, w , in a given sentence S is an instance of an event mention, ev_w ; and subsequently, assign it to a class C , $ev_w \in C$. We break the experiments in two blocks: in the first block, we investigate the quality of the fine-tuned Multilingual BERT models on the identification of the event mentions only. This is an easier task with respect to classification, as it can be framed as a binary classification task. In this way, we can actually have a sort of maximal threshold of the performance of the zero-

shot cross-lingual transfer learning models. In the second block of experiments, we investigate the ability of the models in performing the two sub-tasks “at once”, i.e. identifying and classifying an event mention. This is a more complex task, especially in zero-shot transfer learning scenarios, because the ISO-TimeML classes are assigned following syntactic-semantic criteria: the same word can be assigned to different classes according to the specific syntactic context in which it occurs. For each language pair and direction of the transfer (i.e. $EN_{train}-IT_{test}$ vs. $IT_{train}-EN_{test}$), we also benchmark the performance in monolingual fine-tuned transfer scenarios (i.e. $IT_{train}-IT_{test}$ vs. $EN_{train}-EN_{test}$), to have an upper-bound limit of Multilingual BERT and an indirect evidence of the intrinsic quality of the proposed multilingual model. For the English data, we also test the performance using English BERT_{BASE}, so to better understand limits of the multilingual model.

Finally, we compare our results to the best systems that participated in the corresponding evaluation campaigns in each language, as well as to state-of-the-art systems. In particular, we selected:

- HLT-FBK (Mirza and Minard, 2014), a feature-based SVM model for Italian (best system at EVENTI@EVALITA);
- ATT1 (Jung and Stent, 2013), a feature-based MaxEnt model for English (best system for event detection and classification at TempEval-3);
- CRF4TimeML (Caselli and Morante, 2018),

a feature-based CRF model for English that has obtained state-of-the-art results on event classification;

- Bi-LSTM-CRF (Reimers and Gurevych, 2017; Caselli, 2018), a neural network model based on a Bi-LSTM using a CRF classifier as final layer. The architecture has been originally developed and tested on English (Reimers and Gurevych, 2017), and subsequently adapted to Italian (Caselli, 2018). The English version of the system reports state-of-the-art scores for the event detection task only, while the Italian version obtained state-of-the-art results for detection and classification.

5 Results

All scores for the Multilingual BERT models have been averaged against 5 runs (Reimers and Gurevych, 2017). Subscript numbers correspond to standard deviation scores. Tables 3 and 4 illustrate the results on the Italian test data for the event detection and the event detection and classification sub-tasks, respectively. Results on the English test are illustrated in Table 5 for event detection and in Table 6 for event detection and classification. For each experiment, we also report the number of fine-tuning epochs.

The main take-away is that the portability of the zero-shot models is not the same for the two sub-tasks: for the event detection sub-task, both models obtain close results (average F1 63.53 on Italian *vs.* average F1 66.79 on English), while this is not the case for the event detection and classification sub-task (average F1 42.86 on Italian *vs.* average F1 51.26 on English), suggesting this sub-task as being intrinsically more difficult. We also observe that the zero-shot models have different behaviors with respect to Precision and Recall: the zero-shot transfer on Italian has a high Precision and a low Recall, while the opposite happens on English.⁴ The stability of the zero-shot models seems to be influenced by the size of the fine-tuning training data. In particular, zero-shot transfer learning on English consistently results in more stable models, as the lower scores

⁴For instance, average Precision for event detection is 93.11 on Italian *vs.* 53.19 on English, while average Recall is 51.71 on Italian and 89.92 on English, respectively. A similar pattern is observed for the detection and classification sub-task.

for the standard deviation show when compared to the Italian counterpart (+/- 2.04 for $EVENTI_{train}$ on the TE3 test data *vs.* +/- 7.45 for $TE3_{train}$ on the EVENTI test data for the event detection sub-task; +/- 2.67 for $EVENTI_{train}$ on the TE3 test data *vs.* +/- 3.15 for $TE3_{train}$ on the EVENTI test data for the event detection and classification sub-task).

Annotation differences in the two languages have an impact in the evaluation of the zero-shot models. To measure this, we excluded all modal and copula verbs both as predictions on the English test by the zero-shot Italian model, and as gold labels from the Italian test, when applying the zero-shot English model. In both cases we observe an improvement, with an increase of the average F1 to 72.26 on English and 66.01 on Italian. Although other language specific annotations may be at play, the Italian zero-shot model appears to be more powerful than the English one.

The addition of extra fine-tuning with data from the evaluation language results in a positive outcome, improving performances in both sub-tasks. In three out of the four cases (event detection on English, and event detection and classification on English and Italian) the extra-fine tuning with the full training set of the evaluation language results in better models than the corresponding non zero-shot ones. Adding training material targeting the evaluation test is a well know technique in domain adaptation (Daumé III, 2007). Quite surprisingly with respect to previous work that used this approach, we observe an improvement also with respect to fine-tuned transfer scenarios, i.e. models tuned and tested on the same language, suggesting that the multilingual model is actually learning from both languages.

In terms of absolute scores, our results for the zero-shot scenarios are in line with the findings reported in Pires et al. (2019) for typologically related languages, such as English and Italian. However, limits of zero-shot transfer scenarios seem more evident in semantic tasks when compared to morpho-syntactic ones. For instance, Pires et al. (2019) reports absolute F1 scores comparable to ours on Named Entity Recognition on 4 language pairs, while results on POS tagging achieve an accuracy above 80% on all language pairs. More recently, Wu and Dredze (2019) have shown a similar behavior to our zero-shot scenarios of Multilingual BERT in a text classification task.

Fine Tuning	Epochs	EVENTI F1
TE3 _{train} - zero-shot	1	63.53 _{7.45}
TE3 _{train} + EVENTI _{dev}	1 + 2	77.57 _{1.73}
TE3 _{train} + EVENTI _{train}	1 + 1	87.17 _{0.56}
EVENTI _{train}	1	87.36 _{1.16}
(Caselli, 2018)	n/a	87.79
HLT-FBK	n/a	86.68

Table 3: Event mention detection - test on Italian. Best scores in bold.

Fine Tuning	Epochs	TE3 F1
EVENTI _{train} - zero-shot	1	66.79 _{2.04}
EVENTI _{train} + TE3 _{dev}	1 + 2	80.67 _{1.11}
EVENTI _{train} + TE3 _{train}	1 + 1	81.87 _{0.13}
TE3 _{train}	1	81.39 _{1.23}
(Reimers and Gurevych, 2017) ³	n/a	83.45
ATT1	n/a	81.05

Table 5: Event mention detection - test on English. Best scores in bold.

Fine Tuning	Epochs	EVENTI F1
TE3 _{train} - zero-shot	2	42.86 _{3.15}
TE3 _{train} + EVENTI _{dev}	1 + 2	55.38 _{1.34}
TE3 _{train} + EVENTI _{train}	1 + 3	73.90 _{0.45}
EVENTI _{train}	2	73.69 _{0.80}
(Caselli, 2018)	n/a	72.97
HLT-FBK	n/a	67.14

Table 4: Event detection and classification - test on Italian. Best scores in bold.

Fine Tuning	Epochs	TE3 F1
EVENTI _{train} - zero-shot	2	51.26 _{2.67}
EVENTI _{train} + TE3 _{dev}	1 + 2	64.16 _{2.82}
EVENTI _{train} + TE3 _{train}	1 + 3	68.97 _{0.94}
TE3 _{train}	2	63.36 _{1.47}
CRF4TimeML	n/a	72.24
ATT1	n/a	71.88

Table 6: Event detection and classification - test on English. Best scores in bold.

6 Discussion

Extra fine-tuning Extra fine-tuning, even with a minimal amount of data as shown by the results using the development sets, shifts the model’s predictions to be more in-line with the corresponding language specific annotations. Furthermore, it reduces the effects of cross-lingual transfer based on the presence of the same word pieces between the fine-tuned and the evaluation languages due to the single multilingual vocabulary of Multilingual BERT (Pires et al., 2019). This also results in an increasing stability of the models and a reduction of the differences in the average scores for Precision and Recall with respect to the zero-shot models.

Comparison to other systems Zero-shot models obtain satisfying, though not optimal, results as they fall far from both the state-of-the-art models and the best performing systems in the corresponding evaluation exercises (i.e. HLT-FBK for Italian and ATT1 for English). Extra fine-tuning with the development data provides competitive models against the best systems in the evaluation exercises only. When the full training data is used for extra fine-tuning in the target evaluation language, results are very close to the state of the art, although only in one case the Multilingual BERT model is actually outperforming it (namely, on event detection and classification for Italian). These models also obtain very competitive results with respect to state-of-the-art systems, indicating that multilinguality does not seem to negatively

affect the quality of the pre-trained LM. However, results on English using English BERT_{BASE} appears to be partially in line with this observation. By applying the same settings, we obtain an average F1 on event detection of 82.85,⁵ and an average F1 for event detection and classification of 71.09. Although results of the monolingual model are expected to be higher in general, in this case, we observe that the differences in performance between the two tasks are not in the same range. BERT_{BASE} obtains an increase of 2% on event detection but it reaches almost 11% on event detection and classification. Differences in class labelling between English and Italian (see Section 2) can partially explain this behaviour. However, given the sensitivity of event classification to the syntactic context, these results call for further investigation on the encoding of syntactic information between the monolingual and the multilingual BERT models.

Errors Comparing the errors of the zero-shot models is not an easy task mainly because of the language specific annotations in the two corpora. However, focusing on the three major POS, i.e. nouns, verbs, and adjectives, and on the False Negatives only, both models present a similar proportions of errors, with nouns representing the hardest case (53.84% on Italian vs. 54.90% on English), followed by verbs (30.29% on Italian vs. 17.64% on English), and by adjectives (7.51% on Italian vs. 5.88% on English). When observing the classification mismatches (i.e. correct event mention but

⁵Precision: 81.26; Recall: 84.70

wrong class), both models overgeneralise the OCCURRENCE class in the majority of cases. However, zero-shot transfer on English actually extends mis-classification errors mirroring the distribution of the classes of the Italian training data. In particular, it wrongly classifies English REPORTING events as LACTION (33.33%), and OCCURRENCE as STATE (15.51%) or LACTION (34.48%). Although the syntactic context may have influenced the classification errors, these patterns further highlight the differences in annotations between the two languages.

7 Conclusion

In this contribution we investigated the generalisation abilities of Multilingual BERT on Italian and English using event detection as a downstream task. The results show that Multilingual BERT seems to handle cross-lingual generalisation between Italian and English in a satisfying way, although with some limitations. Limitations in this case come from two sources: annotation differences in the two languages and, partially, the shared multilingual vocabulary. Zero-shot systems appears to be particularly sensitive to the fine-tuning data, and, in these experiments, they provide empirical evidence of the impact of different annotation decisions for events in English and Italian.

We have shown that extra fine-tuning with data of the evaluation language not only is beneficial but it may lead to better systems, suggesting that the multilingual model may be combining information from the two languages, and thus obtaining competitive results with respect to task-specific architectures. This opens up to new strategies for the development of systems by using interoperable annotated data in different languages to improve performances and possibly obtain more robust and portable models across different data distributions.

References

- David Ahn. 2006. The stages of event extraction. In *Proceedings of the Workshop on Annotating and Reasoning about Time and Events*, pages 1–8. Association for Computational Linguistics.
- Waleed Ammar, George Mulcaire, Yulia Tsvetkov, Guillaume Lample, Chris Dyer, and Noah A. Smith. 2016. Massively multilingual word embeddings.
- Jun Araki, Lamana Mulaffer, Arun Pandian, Yukari Yamakawa, Kemal Oflazer, and Teruko Mitamura. 2018. Interoperable annotation of events and event relations across domains. In *Proceedings 14th Joint ACL - ISO Workshop on Interoperable Semantic Annotation*, pages 10–20. Association for Computational Linguistics.
- Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018. A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 789–798, Melbourne, Australia, July. Association for Computational Linguistics.
- Tommaso Caselli and Roser Morante. 2018. Systems Agreements and Disagreements in Temporal Processing: An Extensive Error Analysis of the TempEval-3 Task. In Nicoletta Calzolari (Conference chair), Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Koiti Hasida, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hlne Mazo, Asuncion Moreno, Jan Odijk, Stelios Piperidis, and Takenobu Tokunaga, editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Paris, France, may. European Language Resources Association (ELRA).
- Tommaso Caselli and Rachele Sprugnoli. 2017. It-TimeML and the Ita-TimeBank: Language Specific Adaptations for Temporal Annotation. In Nancy Ide and James Pustejovsky, editors, *Handbook of Linguistic Annotation - Volume II*, pages 969–988. Springer.
- Tommaso Caselli, Rachele Sprugnoli, Manuela Speranza, and Monica Monachini. 2014. EVENTI: Evaluation of Events and Temporal Information at Evalita 2014. In *Proceedings of the First Italian Conference on Computational Linguistics CLiC-it 2014 & and of the Fourth International Workshop EVALITA 2014*, pages 27–34. Pisa University Press.
- Tommaso Caselli. 2018. Italian Event Detection Goes Deep Learning. In *Proceedings of the 5th Italian Conference on Computational Linguistics (CLiC-it 2018)*, Turin, Italy.
- Alexandra Chronopoulou, Christos Baziotis, and Alexandros Potamianos. 2019. An embarrassingly simple approach for transfer learning from pre-trained language models. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2089–2095, Minneapolis, Minnesota, June. Association for Computational Linguistics.
- Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods*

- in *Natural Language Processing*, pages 2475–2485, Brussels, Belgium, October–November. Association for Computational Linguistics.
- Hal Daumé III. 2007. Frustratingly easy domain adaptation. *ACL 2007*, page 256.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June. Association for Computational Linguistics.
- Akiko Eriguchi, Melvin Johnson, Orhan Firat, Hideto Kazawa, and Wolfgang Macherey. 2018. Zero-shot cross-lingual classification using multilingual neural machine translation.
- Lifu Huang, Heng Ji, Kyunghyun Cho, Ido Dagan, Sebastian Riedel, and Clare Voss. 2018. Zero-shot transfer learning for event extraction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2160–2170, Melbourne, Australia, July. Association for Computational Linguistics.
- SemAf/Time Working Group ISO, 2008. *ISO DIS 24617-1: 2008 Language resource management - Semantic annotation framework - Part 1: Time and events*. ISO Central Secretariat, Geneva.
- Heng Ji and Ralph Grishman. 2008. Refining event extraction through cross-document inference. *Proceedings of ACL-08: HLT*, pages 254–262.
- Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2017. Google’s multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.
- Hyuckchul Jung and Amanda Stent. 2013. Att1: Temporal annotation using big windows and rich syntactic and semantic features. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 20–24, Atlanta, Georgia, USA, June. Association for Computational Linguistics.
- Paramita Mirza and Anne-Lyse Minard. 2014. FBK-HLT-time: a complete Italian Temporal Processing system for EVENTI-EVALITA 2014. In *Fourth International Workshop EVALITA 2014*, pages 44–49.
- Thien Huu Nguyen and Ralph Grishman. 2015. Event detection and domain adaptation with convolutional neural networks. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, volume 2, pages 365–371.
- Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*.
- Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy, July. Association for Computational Linguistics.
- Prajit Ramachandran, Peter Liu, and Quoc Le. 2017. Unsupervised pretraining for sequence to sequence learning. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 383–391, Copenhagen, Denmark, September. Association for Computational Linguistics.
- Nils Reimers and Iryna Gurevych. 2017. Reporting score distributions makes a difference: Performance study of lstm-networks for sequence tagging. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 338–348, Copenhagen, Denmark, September. Association for Computational Linguistics.
- Alan Ritter, Oren Etzioni, Sam Clark, et al. 2012. Open domain event extraction from twitter. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1104–1112. ACM.
- Naushad UzZaman, Hector Llorens, Leon Derczynski, James Allen, Marc Verhagen, and James Pustejovsky. 2013. Semeval-2013 task 1: Tempeval-3: Evaluating time expressions, events, and temporal relations. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 1–9, Atlanta, Georgia, USA, June. Association for Computational Linguistics.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.
- Marc Verhagen, Roser Sauri, Tommaso Caselli, and James Pustejovsky. 2010. Semeval-2010 task 13: Tempeval-2. In *Proceedings of the 5th international workshop on semantic evaluation*, pages 57–62. Association for Computational Linguistics.
- Ivan Vulić and Marie-Francine Moens. 2015. Bilingual word embeddings from non-parallel document-aligned data applied to bilingual lexicon induction. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the*

7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers), volume 2, pages 719–725.

Shijie Wu and Mark Dredze. 2019. Beto, bentz, becas: The surprising cross-lingual effectiveness of bert. *arXiv preprint arXiv:1904.09077*.

PESInet: Automatic Recognition of Italian Statements, Questions, and Exclamations With Neural Networks

Sonia Cenceschi

SUPSI University

Via Pobietto 11, Manno, Switzerland

sonia.cenceschi@supsi.ch

Roberto Tedesco

Politecnico di Milano

P.za L. da Vinci 32, Milano, Italy

roberto.tedesco@polimi.it

Licia Sbattella

Politecnico di Milano

P.za L. da Vinci 32, Milano, Italy

licia.sbattella@polimi.it

Davide Losio

Politecnico di Milano

P.za L. da Vinci 32, Milano

davide.losio@mail.polimi.it

Mauro Luchetti

Politecnico di Milano

P.za L. da Vinci 32, Milano, Italy

mauro.luchetti@mail.polimi.it

Abstract

PESInet is an Automatic Prosody Recognition system aiming at classifying Information Units as *Statement*, *Question* or *Exclamation*. PESInet adopts a modular architecture, with a master NN evaluating the results of two independent BLSTM NNs that work on audio and its transcription. PESInet has been trained with our own three-class, balanced corpus composed of about 1.5 million text phrases and 60 000 utterances of recited and spontaneous speech. PESInet reached an accuracy of 80% on three classes, and 91% on two classes (*Question* vs *Non-question*). Finally PESInet, compared against human listeners on a two-class test based on a different corpus, reached a better Accuracy (89% for PESInet, against 80% for human listeners).

1 Credits

The Prosody Extraction by Sound Interpreting network (PESInet) is part of the Lend Your Voice (LYV) project, which has been funded by the Polisocial Award¹ 2016, in collaboration with Fondazione Sequeri Esagramma².

2 Introduction

The goal of PESInet was to investigate whether clues derived from text could improve the recognition of simple prosodic forms in Information Units

(IUs). In particular, we focused on *Statement*, *Question*, and *Exclamation* which are proposition's structures and are independent of the pragmatic function of the corresponding IU: each one can assume a large set of illocutionary acts, as explained into the Language into Act theory (L-Act) described in Cresti (2014). An IU is composed of a textual realisation (i.e., a written phrase) and an acoustic realisation (an audio recording of a speaker uttering such a phrase), and conveys a specific informative intention (Austin, 1975; Cresti, 2000). We designed a modular model based on Neural Networks (NNs), able to highlight how much audio and text affected recognition accuracy. Moreover, to validate our results, we compared our NN model against human listeners, on a set of IUs that did not overlap with the corpus we used to train the model.

3 Background

The majority of studies on prosody regards the automatic recognition or detection of single prosodic clues (Ren et al., 2004; Jeon and Liu, 2009; Tamburini and Wagner, 2007; Taylor, 1993). Others, deal with the detection of phrase boundaries or prosodic phrases (Liu et al., 2006; Wightman and Ostendorf, 1991; Rosenberg, 2009). Just a few works, however, focus on *modality* detection. In the following we briefly introduce some of them. *Question* detection is investigated in Tang et al. (2016) using Recurrent Neural Networks (RNN), in the Mandarin language. Authors propose sev-

¹<http://www.polisocial.polimi.it>

²<https://www.esagramma.net>

Copyright 2019 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

eral RNN and Bidirectional RNN (BRNN) models, trained on a simulated call-centre recordings consisting of just 2850 Question and 3142 Non-question IUs. The best result is an F_1 score of 85.5%.

The work described in Yuan and Jurafsky (2005) focuses on *Question* and *Statement* detection, from text and audio, for Chinese; authors investigate the influence of text in prosody comprehension, on a telephone corpus (with transcriptions). Their classifier achieves an error rate of 14.9% with respect to a 50% chance-level rate. Quang et al. (2007) use decision trees to automatically detect *Questions* in a small elicited French and Vietnamese corpora, leveraging both acoustic and lexical features (unigrams, bigrams, and presence of so-called “interrogative terms”). The best result is an F_1 of 80% for the Vietnamese language.

Finally, the work described in Li et al. (2016) combines Convolutional NNs (CNNs) and Bidirectional Long Short-Term Memory NNs (BLSTM) to extract textual and acoustic features for recognising stances (Affirmative, Neutral, Negative opinions) in the Mandarin language. It exploits a small, manually-tagged corpus of four debate videos (1254 IUs). Combining both audio and text this system reaches an Accuracy of 90.3%.

None of the works mentioned above is perfectly comparable with ours and, on the other hand, all of them are based on ad-hoc corpora (as we did). This makes impossible to compare the results we obtained against other approaches. We, however, validated our results comparing our model against human listeners.

4 The corpus

Our own corpus is composed of eBooks, EPUB3 audio-books (an EPUB3 audio-book contains both text and audio recording, time-aligned at the level of sentence), and the LIT/DIA-LIT corpus (Biffi, 1976; Buroni, 2009), which contains audio recordings of Italian TV shows, with transcriptions.

From eBooks, the textual part of EPUB3 audio-books, and transcriptions of LIT/DIA-LIT we extracted about 1.5 million sentences, balanced on the three target classes: *Statement*, *Question*, and *Exclamation*.

From LIT/DIA-LIT audio recordings and the audio part of EPUB3 audio-books, we collected

about 60 000 utterances (again, balanced on the three target classes). Both sentences and utterances were tagged with the correct class, leveraging the punctuation marks we found in text/transcriptions. Of course, we removed such punctuation marks from the textual part of the corpus. Moreover, we discarded all the sentences containing a sub-phrase or other complex syntactic structures. In doing so we aimed at retaining plain simple examples of statements, questions, and exclamations.

We are aware that leveraging punctuation marks for tagging sentences can lead to confounds, as exclamation marks is also used for Vocatives and Orders, while the full stop is also used for Orders. Anyway, it was simply not possible to manually review the text collection and manually solve the problem. Thus, we assume our corpus is affected by a small amount of noise (in other words, we assume Exclamations and Statements are way more frequent than Vocatives and Orders).

Notice that the question marks might be used for different question typologies (rhetorical, information-seeking, confirmation-seeking, biased), and that question could be further partitioned into open questions, polar questions, etc. Thus, the question mark is used to tag sentences with wildly divergent phonetic forms. This is not, however, a blocking issue: it only makes harder for the classifier to learn the input/output correlation. In particular, this is one of the reasons that lead us to the idea of leveraging text to improve the classification of IUs.

Summing up, we built three corpora:

- **ACorpus**: audio corpus composed of about 60 000 .wav labelled samples.
- **TCorpus**: textual corpus composed of about 1.5 million .txt labelled samples.
- **MCorpus**: mixed corpus composed of all the ACorpus files, with their transcriptions (from the TCorpus); about 60 000 labelled samples.

5 Features extraction

From acoustic and textual samples we derived a set of features that our NNs leveraged for training and recognition.

5.1 Acoustic features

With a sample rate of 44.1 kHz, we adopted a window of 2048 samples with a hop-size of 1024 sam-

ples (i.e., every 23 ms a new vector of acoustic features is produced). Notice that our window is larger than the one usually adopted by ASRs; in fact, we are not interested in phone recognition and, on the other hand, prosody phenomena appear in larger temporal scale than the one involving individual phones.

We tried several window sizes, and several acoustic features; in particular we experimented with different combinations of Cepstrum coefficients. At the end, we come up with the following 129 acoustic features, normalised (to minimise dependency on speakers and recording settings) and calculated by means of Praat (Boersma and others, 2001), as they provided the best results:

- pitch value, with its delta and delta-delta
- energy, with its delta and delta-delta
- the first 40 Cepstrum coefficients, with their deltas and delta-deltas
- energy of such 40 Cepstrum coefficients (as MFCC defines), with its delta and delta-delta

Notice that we did not adopt a true “deep” architecture, as features were not “discovered” by the network. The field of audio analysis already provides a huge set of well-known, informative features; thus, in our opinion, there is no point in let the network approximating them. Moreover, pre-calculated features permit to simplify the network. Summing up, each utterance was transformed into an array that contains a column of 129 real numbers every 23ms.

5.2 Textual features

To feed the model with textual samples we used the usual word embedding technique, which represents the vocabulary in a continuous vector space of 300 dimensions (Sahlgren, 2008). In particular we adopted Italian Word Embeddings, a pre-trained model of 700 000 words based on GloVe (Pennington et al., 2014).

Summing up, each sentence was transformed into an array that contains a column of 300 real numbers for each token. Notice that punctuation marks were discarded and no lemmatisation was applied.

Available at: <http://hlt.isti.cnr.it/wordembeddings/>

6 Architecture

PESInet is composed of three different NNs:

1. Audio-based NN
2. Text-based NN
3. Master NN combining the prediction of the two preceding NNs

We developed two NN architectures: for Audio-based and Text-based NNs, and for Master NN.

6.1 The convolutional block

Acoustic and textual features defined in Section 5 generated low-level pieces of information, looking at very local phenomena. For considering higher-level phenomena, both the Audio-based and the Text-based NNs relied on the same architecture, leveraging an initial multi-layer convolutional block.

A convolutional layer is composed of several kernels with a predefined width, which “scan” the input array. Each kernel, after the training phase, specialises in finding certain *patterns* in the input sequence. The network learns “high level” features (i.e., common prosody contours, for the Audio-based NN, or particular word sequences for the Text-based NN) from our low-level feature set.

Features related to prosody unfold along different time extents (Cutugno et al., 2005): we found dependencies both in short and long time periods. So the idea was to use different kernel widths, in order to allow the network to consider different pattern lengths. The hint to adopt this technique come from various papers (Sbattella et al., 2014; Gussenhoven, 2008; Büring and others, 2009), which thoroughly analysed the idea of simultaneously analysing the input at different temporal granularities with the use of differently-sized kernels.

In particular, our convolutional block is composed of three layers, which “scan” at three different temporal granularity levels. In general, if s is the stride adopted for kernels at any temporal granularity level and d_i is the kernel height at the i -th temporal granularity level, the kernel height at the $(i + 1)$ -th temporal granularity level is $d_{i+1} = d_i + s$; see Figure 1, for a simplified example with two levels. Stride is chosen so that, after each shift of the filter, the kernel will include a small subset of the previously analysed input.

Finally, padding is applied to the input sequence, so that the shorter kernels (and, by construction, all the other, longer kernels) fit the sequence length.

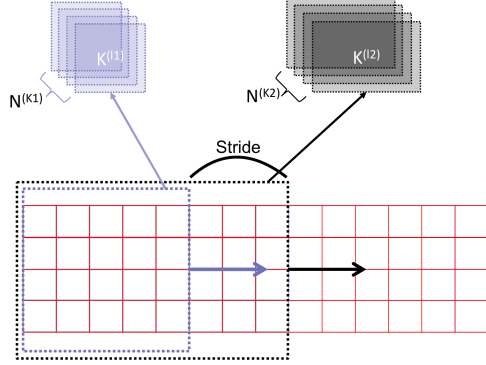


Figure 1: Kernels $K^{(l1)}$ and $K^{(l2)}$ at two different temporal granularity levels.

Being the kernels of different heights, they will cause the outputs to have different dimension as well, relatively to the layers they come from. These dimensions are adjusted in the following layer of the network. Figure 2 shows a simplified schema with two differently-sized kernel groups.

6.2 Audio-based and Text-based NNs

Both the Audio-based and the Text-based NNs relied on a multi-layer network. The general architecture is composed of three BLSTM layers on top of the convolutional block. We connected the first convolutional layer to the first BLSTM layer; then, the second convolutional layer is connected, together with the output from the first BLSTM, to the second BLSTM layer; finally, the third convolutional layer is connected, together with the output of the second BLSTM layer, to the third BLSTM layer. Figure 3 shows the way in which the convolutional block is used.

The Softmax layer shown in the Figure 3 is used during the training phase and then removed, as the Text-based and Audio-based NNs are combined together with the Master NN.

6.3 Master NN and PESInet

The Master NN is composed of a fully-connected layer, and a Softmax layer. PESInet, the resulting network, is shown in Figure 4. Notice that PESInet is supposed to work on utterances, while the text is generated by means of an ASR; in fact, this is the setting we expect to be adopted during

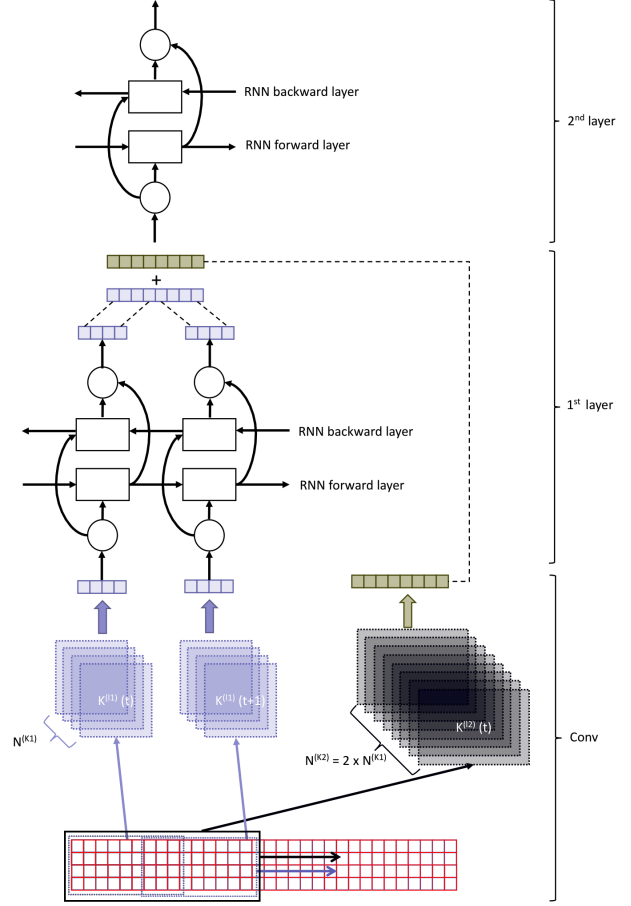


Figure 2: Convolution with two kernel sizes (i.e., two temporal granularity levels).

actual usage of PESInet. Our corpus, conversely, was based on human-generated text; we are aware that in doing so we did not consider the errors due to the ASR and, as a consequence, overestimated the figures obtained during the training/validation procedure. The rationale was highlighting the contribution of text-related features to the recognition of prosodic forms, and thus we decided to avoid the “noise” introduced by ASR-related errors.

As a final remark on the ASR, notice that it is supposed to not add any punctuation mark to the transcription it generates.

7 Training

The architecture was implemented, trained, and tested using the TensorFlow library along with Python 3.6. The code itself was run on a machine equipped with 32GB of RAM, a Xeon Intel processor and a Nvidia Titan X (Pascal) GPU. During training, we adopted the early stopping (using Accuracy as reference index); moreover, to improve

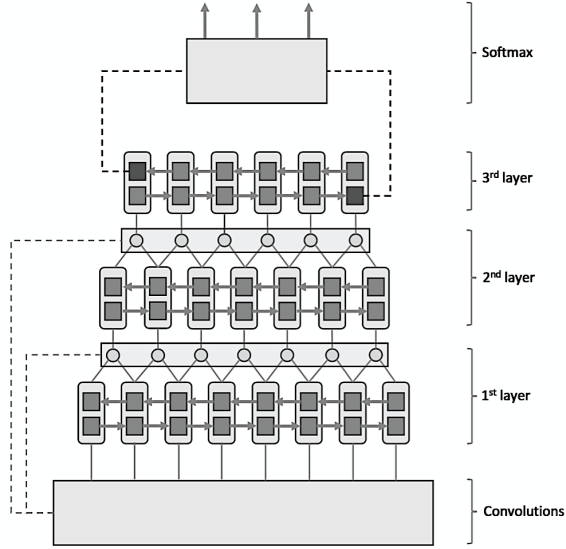


Figure 3: Structure of the Text-based and Audio-based NNs.

the learning effectiveness, we used the variational drop-out on recurrent layers. We started training, independently, the Audio-based and the Text-based NNs, on 80% of their respective corpora: ACorpus and TCorpus. Then, once removed the final Softmax layer from them, these NNs were attached to the Master NN, and a further training –involving 80% of the MCorpus– was performed on PESInet. In particular, we investigated three approaches:

1. Allowing PESInet to train only the Master NN weights (all the others remain fixed).
2. Allowing PESInet to change all its internal weights (also those already trained).
3. Training PESInet from scratch, skipping training of Audio-based and Text-based NNs.

8 Evaluation

Validation was performed using 20% of the corpus. We experimented with several feature combinations, hyperparameter values, and network structures, before reaching the final models.

The Audio-based and Text-based NNs gave the following Accuracies: 0.68 and 0.79. It's interesting that the Text-based NN gave a better Accuracy than the Audio-based NN. This was surprising, as, after all, prosody is an acoustic phenomenon. Nevertheless, data seem to show that

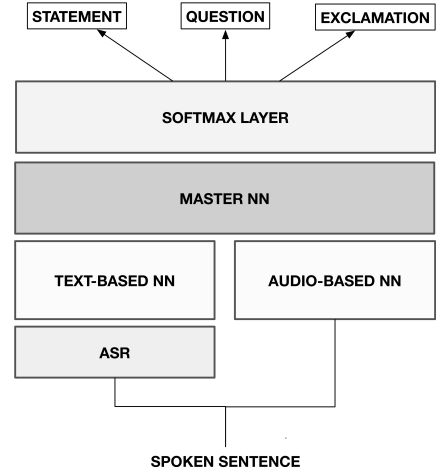


Figure 4: PESInet structure.

		Predicted		
		Stat.	Excl.	Quest.
True	Stat.	1366	234	155
	Excl.	285	1068	316
	Quest.	216	484	1130

Table 1: Confusion matrix for Audio-based NN.

the words composing the utterance are indeed a good predictor of prosody. Moreover, considering that ACorpus was much smaller than TCorpus, the surprisingly low results of Audio-based NN can be explained.

Table 1 and Table 2 show the confusion matrices for the two NNs. It's interesting to notice that Audio-based NN predicted Statements much better than the other two classes, while Text-based NN was also very good in recognising Questions.

About PESInet, Table 3 shows that the approach 2 obtained, as expected, the best results. As the confusion matrix of Table 4 shows, audio and text cooperated to improve recognition of all the three classes.

As a further experiment, we trained and tested PESInet on two classes: *Question* vs *Non-question*, adapting the same PESInet architecture to handle 2 classes. The corpus tags

		Predicted		
		Stat.	Excl.	Quest.
True	Stat.	48 478	7233	3358
	Excl.	8786	43 887	6064
	Quest.	4494	5905	48 495

Table 2: Confusion matrix for Text-based NN.

Trained NN	PT	F ₁	Loss	Acc.
1. Master NN	yes	0.79	0.55	0.77
2. PESInet	yes	0.80	0.49	0.80
3. PESInet	no	0.80	0.55	0.78

Table 3: Results for PESInet. PT: Pre-training Text-based and Audio-based NNs.

		Predicted		
		Stat.	Excl.	Quest.
True	Stat.	1444	205	106
	Excl.	222	1242	205
	Quest.	92	215	1523

Table 4: Confusion matrix for PESInet.

Trained NN	PT	F ₁	Loss	Acc.
2. PESInet	yes	0.91	0.39	0.91

Table 5: Results for PESInet, two classes.

{*Exclamation, Statement*} were rewritten as *Non-question*, and we randomly extracted a number of *Non-Question* samples equals to the *Question* samples. Then, we used 90% of such dataset for training and 10% for testing. Accuracy reached 91% (Table 5).

8.1 PESInet against human listeners

Finally, to validate the results we obtained, we conducted a perceptive experiments with 302 Italian speakers (Cenceschi et al., 2018b; Cenceschi et al., 2018a). The aim of the experiment was to understand the role of acoustic clues and textual clues in the perception of various prosodic forms.

The experiment was divided into several tests; each test was about a specific prosodic form: users were asked to listen a set of IUs and select which of them carried the expected prosodic form. In that experiment we used an ad-hoc audio/textual corpus called SI-CALLIOPE, where 14 professional actors spoke a set of 139 sentences, for a total of 1946 IUs. Notice that SI-CALLIOPE did not share anything, in terms of sentences and speakers, with corpora we used to train PESInet.

In particular, for the *Question/Non-question* test, each user listened to a set of audios randomly extracted from 714 question IUs and 1232 non-question IUs. The average accuracy was 80% (std. dev.: 7.24%).

Running the two-class version of PESInet on the same test, we got an Accuracy of 89%.

We argue that this surprisingly good Accu-

racy for our NN (or surprisingly bad Accuracy for human listeners) could be caused by *de-contextualisation*: in the experiment each IU was given in isolation, without any dialogue context; probably, listeners were more affected by that lacking of context than our NN. Anyway, this is just a hypothesis that should be investigated and deepened with further experiments, as the comparison could be tainted by a large number of other confounds, such as the non ecological nature of the task and the stratification of the repertoire of Italian speakers.

9 Conclusions and discussion

PESInet got an Accuracy of 80% on three classes and 91% on two classes. Moreover, PESInet reached very good results when compared to human listeners on a totally different corpus. Although this human/NN comparison should be taken with a grain of salt, we believe that it is a hint that the network works well and the results are truly promising. As a future work, more recordings should be added to ACorpus and MCorpus to improve the performance of the Audio-based NN and, as a consequence, of the whole PESInet.

Currently, we are working for cleaning the code and streamlining the training procedure, as we plan to release the code.

References

- John Langshaw Austin. 1975. *How to do things with words*, volume 88. Oxford university press.
- Marco Biffi. 1976. Il lit-lessico italiano televisivo: l'italiano televisivo in rete. *L'italiano televisivo: 1976-2006. Atti del convegno-Milano, 15-16 giugno 2009*, pages 35–69.
- Paul Boersma et al. 2001. Praat, a system for doing phonetics by computer. *Glott international*, 5(9/10):341–345.
- Daniel Buring et al. 2009. Towards a typology of focus realization. *Information structure*, pages 177–205.
- Edoardo Buroni. 2009. La voce del telegiornale. aspetti prosodici del parlato telegiornalistico italiano in chiave diacronica. l'italiano televisivo 1976–2006. *Atti del Convegno, Milano*, pages 15–16.
- Sonia Cenceschi, Licia Sbattella, and Roberto Tedesco. 2018a. Towards automatic recognition of prosody. In *Proc. 9th International Conference on Speech Prosody 2018*, pages 319–323.

- Sonia Cenceschi, Licia Sbattella, and Roberto Tedesco. 2018b. Verso il riconoscimento automatico della prosodia. *STUDIA ISV*, pages 433–440.
- Emanuela Cresti. 2000. *Corpus di italiano parlato: Introduzione*, volume 1. Accademia della Crusca.
- F. Cutugno, G. Coro, and M. Petrillo. 2005. Multigranular scale speech recognizers: Technological and cognitive view. In Springer, editor, *Congress of the Italian Association for Artificial Intelligence*, 227–330, Berlin.
- Carlos Gussenhoven. 2008. Types of focus in english. In *Topic and focus*, pages 83–100. Springer.
- Je Hun Jeon and Yang Liu. 2009. Automatic prosodic events detection using syllable-based acoustic and syntactic features. In *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 4565–4568. IEEE.
- Linchuan Li, Zhiyong Wu, Mingxing Xu, Helen M Meng, and Lianhong Cai. 2016. Combining cnn and blstm to extract textual and acoustic features for recognizing stances in mandarin ideological debate competition. In *INTERSPEECH*, pages 1392–1396.
- Yang Liu, Elizabeth Shriberg, Andreas Stolcke, Dustin Hillard, Mari Ostendorf, and Mary Harper. 2006. Enriching speech recognition with automatic detection of sentence boundaries and disfluencies. *IEEE Transactions on audio, speech, and language processing*, 14(5):1526–1540.
- Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.
- Vũ Minh Quang, Laurent Besacier, and Eric Castelli. 2007. Automatic question detection: prosodic-lexical features and crosslingual experiments. In *Eighth Annual Conference of the International Speech Communication Association*.
- Yuxi Ren, Sung-Suk Kim, Mark Hasegawa-Johnson, and Jennifer Cole. 2004. Speaker-independent automatic detection of pitch accent. In *Speech Prosody 2004, International Conference*.
- Andrew Rosenberg. 2009. *Automatic detection and classification of prosodic events*. Columbia University.
- Magnus Sahlgren. 2008. The distributional hypothesis. *Italian Journal of Disability Studies*, 20:33–53.
- Licia Sbattella, Roberto Tedesco, and Alessandro Trivilini. 2014. Forensic examinations: Computational analysis and information extraction. In *International Conference on Forensic Science-Criminalistics Research (FSCR)*, pages 1–10.
- Fabio Tamburini and Petra Wagner. 2007. On automatic prominence detection for german. In *Eighth Annual Conference of the International Speech Communication Association*.
- Yaodong Tang, Yuchen Huang, Zhiyong Wu, Helen Meng, Mingxing Xu, and Lianhong Cai. 2016. Question detection from acoustic features using recurrent neural network with gated recurrent unit. In *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*, pages 6125–6129. IEEE.
- Paul A Taylor. 1993. Automatic recognition of intonation from f0 contours using the rise/fall/connection model.
- CW Wightman and Mari Ostendorf. 1991. Automatic recognition of prosodic phrases. In *Acoustics, Speech, and Signal Processing, 1991. ICASSP-91., 1991 International Conference on*, pages 321–324. IEEE.
- Jiahong Yuan and Dan Jurafsky. 2005. Detection of questions in chinese conversational speech. In *Automatic Speech Recognition and Understanding, 2005 IEEE Workshop on*, pages 47–52. IEEE.

What Makes a Review Helpful?

Predicting the Helpfulness of Italian TripAdvisor Reviews

Giulia Chiriatti[◊], Dominique Brunato[◊], Felice Dell’Orletta[◊], Giulia Venturi[◊]

• University of Pisa

chiriattigiulia@gmail.com

[◊]Istituto di Linguistica Computazionale “Antonio Zampolli” (ILC-CNR)

ItaliaNLP Lab - www.italianlp.it

chiriattigiulia@gmail.com

{dominique.brunato, felice.dellorletta, giulia.venturi}@ilc.cnr.it

Abstract

In this paper we introduce a classification system devoted to predict the helpfulness of Italian online reviews. It is based on a wide set of features reflecting the different factors involved and tested on different categories of TripAdvisor reviews. For this purpose, we collected the first Italian corpus of online reviews enriched with metadata related to their helpfulness and we carried out an in-depth analysis of the most predictive features.¹

1 Introduction

Predicting and modeling relevant factors that determine the helpfulness of online reviews have been attracting a growing attention in the Natural Language Processing (NLP) community. Both practical applications and the interest to study human variables underlying the assignment of helpful/unhelpful votes are mainly involved. The identification of product reviews which are useful to customers can be important for several e-business purposes (e.g. the development of product recommendation systems) as well as to investigate persuasive elements that make a review helpful for a review reader (Hong et al., 2012; Park, 2018). Several approaches have been devised, differing at the level of predicting methods (mainly regression or classification algorithms) and of typologies of factors considered, including content elements found within the review and contextual ones referring to user profiles. Although various strategies have already been followed, according to the recent survey by Diaz and Ng (2018), a number of issues are still open and deserve to be explored. Among others, they include *i*) the need for “more

sophisticated textual features” that can be useful to model a writing style typical of helpful reviews, and *ii*) the lack of studies focused on languages other than English.

In this paper, we address these open issues and we present a study devoted to predict Italian review helpfulness with a specific focus on the role played by linguistic features in modelling the style of helpful reviews. Similarly to previous studies, we tackled the task as a text classification problem but with two main novelties. Firstly, we relied on different sets of predictors, considering both lexical (content) and structural features (i.e. morpho-syntactic and syntactic) aimed at reconstructing the style of a text (the linguistic “form”). Secondly, we investigated which typology of features are the most effective to predict the helpfulness of online reviews and whether they remain the same across different review categories.

Our contribution. *i*) We collected a corpus of Italian online reviews enriched with metadata related to their helpfulness². *ii*) We developed the first classification system devoted to predict the helpfulness of Italian online reviews, based on features modelling both lexical and linguistic factors involved, and tested it in two experimental scenarios, i.e. in- and out-domain with respect to the training category of reviews. *iii*) We identified and ranked the most predictive features, showing the key role played by linguistic features, especially to predict the helpfulness of reviews belonging to a category very different from the training one.

2 Corpus

We collected a sample of almost 1 million user-generated reviews from the Italian section of TripAdvisor, focusing on two travel-related categories, restaurants and attractions (e.g. parks, historical sites), and two geographical areas, Rome

¹Copyright ©2019 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

²The corpus is available for research purposes at <http://www.italianlp.it/resources/>

and Milan. We also gathered two types of metadata associated with each review: review rating and number of helpful votes. Firstly, we filtered our data according to language (Italian) and length (> 7 tokens), discarding 52.29% of the total reviews. Then we empirically³ set a threshold at a minimum of 3 votes in order to distinguish helpful reviews (3+ votes) from unhelpful ones (0 votes). Some examples of reviews that belong to the two classes are reported in Table 2. In line with studies carried out for the English language (Park, 2018), also in our case review votes tend to be sparse across all categories: in particular reviews with 3+ votes constitute only 5.10% of the unfiltered dataset. For this reason we balanced the data by selecting a comparable number of helpful and unhelpful reviews per restaurant or attraction. As shown in Table 1, our final corpus consists of 42,107 reviews from 1,218 restaurants and 383 attractions for a total of 4,133,312 tokens.

Category	#Helpful	#Unhelpful	#Reviews
Rome rest.	12,635	12,404	25,039
Milan rest.	6,105	5,991	12,096
Attractions	2,564	2,408	4,972
TOTAL	21,304	20,803	42,107

Table 1: Corpus of helpful and unhelpful TripAdvisor reviews.

3 Helpfulness Predictors

According to our research purposes, we considered various categories of features aimed at modeling both the content and the linguistic “form” of online reviews. They can be grouped into three main classes: *lexical*, *linguistic* and *metadata* features. The first typology has already been tested in the literature (Diaz and Ng, 2018) in order to predict review helpfulness on the basis of meaningful words. On the contrary, the use of linguistic features extracted from sentence structure is introduced for the first time in this paper. Differently from previous studies (Kim et al., 2006; Hong et al., 2012) where the distribution of some Parts-Of-Speech was exploited as helpfulness predictor, we rely here on a wide set of linguistic features automatically extracted from the corpus of reviews linguistically annotated. Since they have been shown to have a high discriminative power in different

³In order to choose the threshold value, we considered the mean and the standard deviation of the number of votes in the initial dataset (2.21 ± 0.59).

tasks, e.g. assessment of text readability (Collins, 2014), identification of textual genre of a document (Cimino et al., 2017), we investigated in this study whether they are able to model the linguistic “form” (the style) of helpful reviews. In addition, we explored the contribution of a kind of metadata feature (i.e. the star rating given by the reviewer) that has also been widely tested in studies on helpfulness prediction, as reported in Diaz and Ng (2018).

In order to extract lexical and linguistic predictors of helpfulness, the corpus was linguistically annotated at different levels of analysis. In particular, it was tagged by the PoS tagger described in Dell’Orletta (2009) and dependency-parsed by the DeSR parser (Attardi et al., 2009).

Lexical features. They include two types of features: (i) the distribution of unigrams and bigrams of characters, words and lemmas (hereafter *NGR*); (ii) word embedding combinations (*WE*) obtained by separately computing the average of the vector representations of nouns, verbs and adjectives in the review. The word embeddings were trained on the ItWaC corpus (Baroni et al., 2009) and a collection of Italian tweets⁴ using the *word2vec* toolkit (Mikolov et al., 2013).

Linguistic features. They refer to four main types, modelling diverse aspects of writing style: *raw text features*, i.e. review, sentence and word length, calculated in terms of sentences, tokens and characters, respectively; *features related to lexical richness*, which is captured considering i) the internal composition of the vocabulary of review with respect to the *Basic Italian Vocabulary* and its usage repertoires (De Mauro, 2000), and ii) Type/Token Ratio; *morpho-syntactic features*, i.e. the distribution of unigrams of Parts-of-Speech, and verb moods, tenses and persons; *syntactic features*, which refer to diverse characteristics of sentence structure: i) the depth of the whole parse tree (calculated in terms of the longest path from the root of the dependency tree to some leaf); ii) the length of dependency links (i.e. the tokens occurring between the head and the dependent); iii) the distribution of dependency types, iv) the average depth and the distribution of embed-

⁴<http://www.italianlp.it/resources/italian-word-embeddings/>

Label	Category	Example (<i>Italian</i>)	Example (<i>English</i>)
Helpful	Rome restaurants	La prima regola di un buon ristorante che fa pizza no stop è: Scegliere la pizza che preferisco. Qui non solo non si può scegliere la pizza ma capita spesso che escano le stesse pizze più volte così uno è costretto a mangiare sempre la stessa!! Per non parlare dell'ambiente poi, un vero casino, capisco che l'area bambini è la principale attrazione del ristorante, rivolto soprattutto alle famiglie, ma il casino che si crea non è cmq giustificabile. La pizza è di una qualità davvero scadente, praticamente era cruda!!! La pizza con la Lonza....una semplice focaccia con un pezzo di prosciutto preso molto probabilmente al discount! Ragazzi, carina l'idea di prendersi cura dei pargoli, ma non prendiamoci in giro però.	The first rule of a good restaurant that makes pizza no stop is: Choose the pizza I prefer. Here you can not only choose the pizza but it often happens that the same pizzas come out more times so one is forced to always eat the same one!!! Not to mention the environment then, a real mess, I understand that the children's area is the main attraction of the restaurant, aimed above all at families, but the mess that is created is not justifiable anyway. The pizza is of a really poor quality, practically it was raw!!! Pizza with Lonza....a simple focaccia with a piece of ham most probably taken at the discount store! Guys, nice idea to take care of the little ones, but let's not fool around.
Unhelpful	Milan restaurants	Devo dire che trovandomi per caso in quella zona con i miei amici abbiamo provato il posto è devo dire che è molto accogliente e che la zona per mangiare nel cortile è proprio intima e carina...Per quanto riguarda il mangiare posso dire di essere soddisfatto perché le portate erano nelle mie corde ed avendo preso il pesce ero soddisfatto di quanto cucinato dal cuoco. Bravi mica male.	I must say that finding myself by chance in that area with my friends we tried the place and I must say that it is very welcoming and that the area to eat in the courtyard is really intimate and pretty... As for eating I can say I'm satisfied because the courses were on my ropes and having caught the fish I was satisfied with what the cook had cooked.

Table 2: Examples of helpful vs unhelpful reviews.

ded prepositional chains modifying a noun; *v*) a set of features aimed at modeling the behaviour of verbal predicates, i.e. the number of verbal roots, the average verbal arity and the distribution of verbs by arity, the distribution of verbal predicates with elliptical subject; *vi*) the usage of subordination, calculated considering the ratio between principal and subordinate clauses, and the average depth and the distribution of embedded chains of subordinate clauses; *vii*) a last set of features related to the canonical construction of a sentence in Italian, i.e. the relative ordering of subordinates with respect to the main clause and of subject and object with respect to their verbal head.

The effectiveness of these features to predict helpful online reviews is confirmed by the fact that according to the Wilcoxon rank sum test, 75% of the considered features (i.e. 160 out of 212) turned out to vary in a statistically significant way between helpful and unhelpful reviews. As shown in Table 3, helpful reviews are on average 1-sentence longer than unhelpful ones and they also contain much longer sentences. The correlation between length and helpfulness is not surprising since longer sentences are likely to be more informative, thus offering more contents that might

influence the voting process outcome. The higher sentence length also has an expected effect on some syntactic features correlated to complexity. Sentences occurring in helpful texts have deeper syntactic trees (*Avg. max depth*) and contain more subordinate clauses and embedded prepositional chains. However, they appear as simpler with respect to other features related for instance to canonicity effects. They show a more standard syntactic structure, with a higher distribution of objects in post verbal position and subjects preceding the main verb. Interestingly, helpfulness is also positively correlated with a reader-focused style, as shown by the greater use of pronouns and verbs in the first and second person.

Metadata feature. *Review star rating (STR)* is the rating score assigned by the reviewer, ranging from 1 to 5. Previous research reported in Diaz and Ng (2018) has shown that a connection exists between the rating of the review and its helpfulness. In our dataset rating scores are unequally distributed across the different review categories. Restaurant reviews are more likely to have an extreme rating, either low or high, rather than a neutral one, and helpful reviews follow the same pattern: e.g., in the Rome restaurant cate-

Feature	Help	UnHelp	Diff.
N. sent	4,61	3,46	1,15
Avg. sent length	36.79	26.22	10.57
Avg. clause length	10	11.65	-1.65
% Nouns	23.5	24.5	-1
% Verbs	14.28	12.79	1.49
% Adj	8.32	10.37	-2.41
% Negative adv	1.33	0.97	0.36
% Pronouns	4.99	4.14	0.85
% 1st sing p.	9.23	8.15	1.08
% 2nd pl p.	1.34	1.08	0.26
Avg. prep chains length	11,4	6,3	5,1
Avg. max depth	7,64	6,28	1,36
% Subord clause	62,09	43,89	18,2
% Post obj	78,84	68,66	10,18
% Pre subj	73,13	65,03	8,01

Table 3: A subset of linguistic features whose values vary in a statistically significant way between helpful and unhelpful reviews.

gory 37.05% of the helpful reviews have a rating of 1 and 25.76% a rating of 5. On the contrary, attractions reviews tend to have higher ratings, with 56.12% of the helpful ones belonging to the highest-rated class. Only the attractions category seems to confirm the presence of the *positivity bias* that is discussed in Diaz and Ng (2018), according to which reviews with positive ratings are seen as more helpful.

4 Experiments and Results

We addressed the helpfulness prediction task as a binary classification problem. In order to assess the contribution of each set of features illustrated in Section 3, we defined two experimental scenarios differing at the level of review categories chosen as test data and set-up (in terms of feature configurations). We built a classifier based on the LIBLINEAR implementation of Support Vector Machines with a linear kernel (Fan et al., 2008) and trained on a set of 12,516 reviews written for 411 Rome restaurants. All the features were previously scaled in the same range $[0, 1]$. We evaluated our system by computing the accuracy score for each feature configuration. As baseline for each review category we implemented the score of a classifier which always outputs the most probable class according to the class distribution of the dataset (in this case the *helpful* class).

In the first experimental scenario we tested the feature models generated by the SVM classifier on a test set of 12,523 reviews that belong to the same domain of the training data (i.e. the Rome restaurants category) but were written for restau-

rants different from the ones in the training set. As shown in Table 4, we obtained a general improvement over the baseline with all feature configurations apart from the one that exploits only the metadata feature (*STR*, the star rating of the reviews). Nevertheless, this feature does improve the accuracy score of all models by at least one point, thus confirming its usefulness for helpfulness prediction (Diaz and Ng, 2018). The results also highlight the prominent role of lexical information (*NGR+WE*) in assessing helpfulness, although this is primarily explained by the in-domain scenario. Even if the accuracy of the linguistic model (*LING*) is lower with respect to the one obtained by the other feature models, we found out that linguistic information plays a main role in the helpfulness prediction. It allows achieving an accuracy score of 66% and of 70.81% by also adding review ratings, a value that is in line with that of the lexical model.

Model	Accuracy
STR	49.6%
NGR	69.9%
NGR+STR	71.13%
WE	68.54%
WE+STR	69.96%
NGR+WE	70.17%
NGR+WE+STR	71.14%
LING	66%
LING+STR	70.81%
ALL	70.04%
ALL+STR	71.05%
Baseline	50.46%

Table 4: In-domain classification of helpful vs. unhelpful reviews using different feature models.

In the out-domain scenario we tested the considered feature models on reviews that belong to the other two categories (Milan restaurants and attractions). As reported in Table 5, we observed that the performances of the classifier tested on the reviews of Milan restaurants, even if slightly worse, are very similar to the ones obtained on the test set of Rome restaurants. This result suggests that the system may perform consistently across different geographical areas, although further experiments should be carried out. For example, we might test our models on a greater number of cities or other types of geographical areas. As we expected, the accuracy decreases mainly in the domain more distant from the training one (i.e. the attractions category). This is especially the case of the lexical classification model, that has a drop of 10.5 points.

The star rating feature is also shown to worsen the accuracy scores, probably because of the way the ratings are distributed in the attractions category with respect to the restaurant ones. It is interesting to note that the best performing model resulted to be the one exploiting the linguistic features (with a lower drop of 5.24%), thus showing the predictive power of sentence structure information in predicting review helpfulness.

Model	Milan	Attractions
NGR+WE	69.38%	59.67%
NGR+WE+STR	70.92%	58.02%
LING	65.82%	60.76%
LING+STR	70.92%	60.28%
ALL	69.2%	59.9%
ALL+STR	70.78%	58.49%
Baseline	50.47%	51.56%

Table 5: Out-domain classification of helpful vs. unhelpful reviews in terms of accuracy using different feature models.

5 Discussion

As discussed in the previous section, we found out that linguistic features allow achieving an accuracy almost in line with the one obtained using only lexical information. Interestingly enough, they are the most predictive ones in the out-of-domain scenario. In order to gain insight into which of these features are the most effective in the task of automatic classification, we ranked them according to the absolute value of their weight in the linear SVM model generated with the linguistic feature configuration. Among the 50 top-ranked ones, besides the raw text features (whose role in predicting helpfulness has already been proven in the literature), we found morpho-syntactic and syntactic features. They are typically related to a rich and articulated writing style. This is the case for example of features concerning nominal modification, in particular the number of prepositional chains (holding the 1st position in the ranking) and their average length but also the distribution of adjectives and determiners. Others involve verbal structures, e.g. the number of dependents instantiated by the verbal heads and the frequency of adverbs (especially negation ones). Features related to the usage of subordination, such as the number of subordinate structures and the average depth of parse trees, also appear among the top-ranked. Finally, another group of high-ranked features concerns a subjective writing

style, as shown by the distribution of verbs in the first and second person. These types of features resulted to be discriminant in the comparison between helpful and unhelpful reviews (Section 3). This shows that the writing style of helpful reviews, informative but also personal and reader-focused, has an high predictive power.

The importance of the linguistic features is further confirmed by a second inspection in which the same ranking method was applied to the all-feature model. Also in this case, we found out that 59.6% of the whole set of 212 linguistic features we considered is in the 90th percentile of the ranking of the total 741,339 features.

6 Conclusion

In this paper, we have presented the first approach to the task of review helpfulness prediction for the Italian language. Two experimental scenarios have been tested in a corpus of TripAdvisor reviews belonging to different categories (restaurants and attractions). In line with previous findings obtained for the English language, we confirmed that lexical information plays a significant role in classifying helpful reviews. In addition, we proved for the first time the highly predictive power of linguistic features modeling the writing style independently from the content. This is particularly true in the two out-domain experiments: in the first case (same category, different geographical area), the classifier based on the linguistic features achieves the same accuracy of the model using lexical features and it even outperforms all the other configuration models when tested on the most distant review category (restaurants vs attractions).

Among the possible future issues that we would like to investigate, an interesting one concerns the role played by metadata features. In the reported results, we showed that star ratings are not relevant when considered alone, but they give a plus when combined with both lexical and linguistic features. Beyond this metadata, we would like to extend the analysis to further user information possibly related to review helpfulness.

Acknowledgments

This work was partially supported by the 2-year project PERFORMA (Personalizzazione di PERcorsi FORMativi Avanzati), co-funded by Regione Toscana (POR FSE 2014-2020).

References

- G. Attardi, F. Dell’Orletta, M. Simi and J. Turian. 2009. Accurate dependency parsing with a stacked multilayer perceptron. *Proceedings of Evalita’09, Evaluation of NLP and Speech Tools for Italian*, Reggio Emilia, December.
- M. Baroni, S. Bernardini, A. Ferraresi and E. Zanchetta. 2009. The WaCky Wide Web: A Collection of Very Large Linguistically Processed Web-Crawled Corpora. *Language Resources and Evaluation*, 43(3), pp. 209-226.
- A. Cimino, M. Wieling, F. Dell’Orletta, S. Montemagni S. and G. Venturi. 2017. Identifying Predictive Features for Textual Genre Classification: the Key Role of Syntax. *Proceedings of 4th Italian Conference on Computational Linguistics (CLiC-it)*, 11-13 December, 2017, Rome.
- K. Collins-Thompson. 2014. Computational assessment of text readability: a survey of current and future research. *Recent Advances in Automatic Readability Assessment and Text Simplification*, Special issue of International Journal of Applied Linguistics, (165-2), 97-135.
- F. Dell’Orletta. 2009. Ensemble system for Part-of-Speech tagging. *Proceedings of Evalita’09, Evaluation of NLP and Speech Tools for Italian*, Reggio Emilia, December.
- T. De Mauro. 2000. Grande dizionario italiano dell’uso (GRADIT). Torino, UTET.
- G. O. Diaz and V. Ng. 2018. Modeling and Prediction of Online Product Review Helpfulness: A Survey. *ACL*, 2018.
- R. E. Fan, K.-W. Chang, C.-J. Hsieh, X. Wang and C.-J. Lin. 2008. LIBLINEAR: A Library for Large Linear Classification. *Journal of Machine Learning Research*, 9:1871–1874.
- Y. Hong, J. Lu, J. Yao, Q. Zhu and G. Zhou. 2012. What Reviews are Satisfactory: Novel Features for Automatic Helpfulness Voting. *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 495-504.
- S. M. Kim, P. Pantel, T. Chklovski and M. Pennacchiotti. 2006. Automatically assessing review helpfulness. *Proceedings of the the 2006 Conference on Empirical Methods in Natural Language Processing*, pp. 423-430.
- T. Mikolov, K. Chen, G. Corrado and J. Dean. 2013. Efficient Estimation of Word Representations in Vector Space. *CoRR* abs/1301.3781
- Y.-J. Park. 2018. Predicting the helpfulness of on-line customer reviews across different product types. *Sustainability*, 10(1735), pp. 1-20.

Is This an Effective Way to Annotate Irony Activators?

Alessandra Teresa Cignarella^{1,2}, Manuela Sanguinetti¹, Cristina Bosco¹, Paolo Rosso²

1. Dipartimento di Informatica, Università degli Studi di Torino, Italy

2. PRHLT Research Center, Universitat Politècnica de València, Spain

{cigna|msanguin|bosco}@di.unito.it, proso@dsic.upv.es

Abstract

In this article we describe the first steps of the annotation process of specific irony activators in TWITTIRÒ-UD, a treebank of Italian tweets annotated with fine-grained labels for irony on one hand, and according to the *Universal Dependencies* scheme on the other. We discuss in particular the annotation scheme adopted to identify irony activators and some of the issues emerged during the first annotation phase. This helped us in the design of the guidelines and allowed us to draw future research directions.

1 Introduction

In the last decade, several efforts have been devoted to address the challenges of sentiment analysis and related tasks, working mainly in English and other languages such as Italian, Spanish or French. Provided that most of the existing approaches in NLP are based on supervised semantic shallow analysis and machine learning techniques, there has been a strong push towards the development of resources from where related knowledge can be learned.

In particular the detection of irony is among the tasks currently considered as especially challenging since its presence in a text can reverse the polarity of the opinion expressed, that is using positive words for intending a negative meaning or – less often – the other way around. This can significantly undermine systems' accuracy and makes it crucial to develop irony-aware systems (Bosco et al., 2013; Reyes et al., 2013; Riloff et al., 2013; Wang, 2013; Barbieri et al., 2014; Joshi et al., 2015; Hernández Farías et al.,

2015; Hernández Farías et al., 2016). Additionally, the challenge is further complicated when there is a co-occurrence with sarcasm or satire (Hernández Farías and Rosso, 2016; Joshi et al., 2017; Ravi and Ravi, 2017).

The growing interest in irony detection is also attested by the proposal of shared tasks focusing on this topic within NLP evaluation campaigns. For instance, the pilot task on irony detection proposed for Italian in SENTIPOLC at EVALITA¹, in 2014 and 2016 (Barbieri et al., 2016; Basile et al., 2014), and the related task proposed for French at DEFT at TALN 2017 (Benamara et al., 2017). For what concerns English, after a first task at *SemEval-2015* focusing on figurative language in Twitter (Ghosh et al., 2015), a shared task on irony detection in tweets has been proposed in 2018 (Van Hee et al., 2018). Concerning Spanish, the most recent shared task about irony in social media has been organized at IberLEF 2019 *Irony Detection in Spanish Variants (IroSvA 2019)*, exploring the differences among varieties of Spanish from Spain, Cuba and Mexico (Ortega et al., 2019) in which the organizers also proposed a focus on context, stressing the importance of contextual semantics in ironic productions.

While the majority of the participating systems in the above-mentioned shared-tasks are based on classical machine learning techniques (Cignarella and Bosco, 2019; Frenda and Patti, 2019), researchers have recently started to exploit approaches based on neural networks. Among these, Huang et al. (2017) applied attentive recurrent neural networks (RNNs) that capture specific words which are helpful in detecting the presence of irony in a tweet, while Wu et al. (2018) exploited densely connected LSTMs in a multi-task learning strategy, adding PoS tag features, and Zhang et al. (2019) took advantage of recent advancements in transfer learning techniques.

¹<http://www.evalita.it/>

These settings are a clear indication of the growing interest for a deeper analysis of the linguistic phenomena underlying ironic expressions. Such kind of analysis naturally calls for the exploitation of finer-grained features and resources in order to improve the performance of automatic systems. For instance, an especially fine-grained annotation format for irony is the one proposed in Karoui et al. (2017), concerning French, Italian and English. The same scheme has later been applied on a new Italian corpus: TWITTIRÒ (Cignarella et al., 2018a). The resulting annotated corpus was used as reference dataset in the *IronITA 2018* shared task² on *Irony and Sarcasm Detection in Italian Tweets* (Cignarella et al., 2018b).

1.1 Motivation and Research Questions

The present work is, indeed, part of a wider joint project with other research groups working on English and French (Karoui et al., 2015). As mentioned above, in Cignarella et al. (2018a), we created an Italian corpus of tweets, i.e. TWITTIRÒ, annotated with a fine-grained tagset for irony, and later on, we extended the same resource applying the *Universal Dependencies* (UD) scheme (Nivre et al., 2016), thus creating TWITTIRÒ-UD (Cignarella et al., 2019).

This new corpus collocates in the panorama of treebanks with data extracted from social media, such as those recently developed for Italian and released in the UD repository³, and to the best of our knowledge it is one of the few linguistic resources where sentiment analysis and syntactic annotation are applied within the same framework. The main research question that we want to address is:

RQ1. Is there any syntactic pattern that can help us to automatically detect irony?

The intuition that we follow in this work is that if such “syntactic patterns” which activate irony do actually exist, therefore, they should be particularly evident in the syntactic context of certain lexical elements that create a semantic clash in a text.

For this reason, in the present article, we describe the first steps of the annotation process

of specific irony activators in the TWITTIRÒ-UD corpus, taking advantage of the fact that the annotation format we adopted for the syntactic annotation allows us also to label specific activators at token level and retrieve dependency relations connected to them. In doing so, we are led to the following research questions, anticipated by the title of the paper:

RQ2. Is there an effective way to annotate irony activators?

RQ3. If so, is the one we propose valid?

The paper is organized as follows. In Section 2 the novel dataset TWITTIRÒ-UD and its annotation layers are presented. In Section 3 we describe the annotation process concerning irony activators, and we comment the inter-annotator agreement showing some examples. Finally, in Section 4 and Section 5 we discuss some difficult cases and we conclude the paper.

2 Corpus Description

The current version of TWITTIRÒ-UD comprises 1,424 tweets, annotated at multiple levels: a pragmatic level that attempts to model irony (see Section 2.1) and a syntactic level based on the UD scheme that represents the underlying syntactic structure of the tweets in the corpus (Section 2.2). In addition, we have recently introduced a further level that tries to act as an interface between the previous two (Section 3).

2.1 Annotating Irony

As far as the annotation for irony is concerned, the data of this corpus were manually annotated according to a multi-layered annotation scheme described in Karoui et al. (2017), which in turn includes 4 different levels.⁴ Beyond the annotation of irony vs non-irony (henceforth level 1), the multifaceted annotation scheme is organized in three further layers, namely the *activation type* (level 2), the *categories* (level 3) and the *clues* (level 4).

Irony is often activated by the presence of a clash or a contradiction between two elements (also called P1 and P2). This motivates the annotation of the two different *activation types* at level 2: explicit when both these elements are lexicalized in the message, implicit otherwise.

²<http://di.unito.it/ironita18>.

³[https://github.com/UniversalDependencies/UD_Italian-PoS](https://github.com/UniversalDependencies/UD_Italian-PoSTWTITA)TWITA.

⁴See annotation guidelines at <https://github.com/IronyAndTweets/Scheme>.


```

# sent.id = -----
# twittiro = EXPLICIT EX:OXYMORON PARADOX
# activators = 3 12
# text = Il Pd diviso in due. Non è mai stato così unito. [@user]

1 Il il DET RD Definite=Def|Gender=Masc|Number=Sing|PronType=Art 2 det _ _
2 Pd Pd PROP N SP _ 3 nsubj _ _
3 diviso diviso ADV A Gender=Masc|Number=Sing 0 root _ _
4 in in ADP E _ 5 case _ _
5 due due NUM N NumType=Card 3 obl _ SpaceAfter=No
6 . . PUNCT FS _ 3 punct _ _
7 Non non ADV BN PronType=Neg 12 advmod _ _
8 è essere AUX VA Mood=Ind|Number=Sing|Person=3|Tense=Pres|VerbForm=Fin 12 cop _ _
9 mai mai ADV B _ 12 advmod _ _
10 stato essere AUX VA Gender=Masc|Number=Sing|Tense=Past|VerbForm=Part 12 aux _ _
11 così così ADV B _ 12 advmod _ _
12 unito unito ADV A Gender=Masc|Number=Sing 3 parataxis _ SpaceAfter=No
13 . . PUNCT FS _ 12 punct _ _
14 [ [ PUNCT FB _ 15 punct _ SpaceAfter=No
15 @user @user SYM SYM _ 12 vocative:mention _ SpaceAfter=No
16 ] ] PUNCT FB _ 15 punct _ SpaceAfter=\n

```

Figure 1: Example of tweet in CoNLL-U format.

The main linguistic devices reported in literature as irony triggers are described instead at level 3 by the *categories* of the scheme (i.e. analogy, euphemism, false assertion, oxymoron/paradox, context shift, hyperbole, rhetorical question and other). Table 1 shows the distribution of ironic categories throughout the corpus.

	n#	%
ANALOGY	261	18%
EUPHEMISM	84	6%
EX:CONTEXT SHIFT	185	13%
EX:OXYMORON PARADOX	277	19%
HYPERBOLE	81	6%
IM:FALSE ASSERTION	117	8%
OTHER	198	14%
RHETORICAL QUESTION	221	16%
TOTAL	1,424	

Table 1: Ironic categories in TWITTIRÒ-UD.

Finally the *clues* of level 4 are lexical or morpho-syntactic signals of the *activation types* and *categories* that can be found in a given ironic tweet, such as the preposition “like” or the presence of comparative structures in the *analogy* type, or the adverb “very” for *hyperbole*. For more details about this annotation scheme, see Karoui et al. (2017).

2.2 Annotating Universal Dependencies

The availability of social media data annotated also at syntactic level is a prerequisite for our study and for the kind of annotation we intend to perform; as a dependency-based representation was deemed to be more suitable for our purposes, Universal Dependencies became our natural choice.

To obtain the data thus annotated, we ran UD-Pipe (Straka and Straková, 2017) for tokenization, PoS tagging, lemmatization and dependency parsing, using a model trained on two Italian resources available in the UD repository, the ISDT (Simi et al., 2014) and PoSTWITA-UD (Sanguinetti et al., 2018) treebanks⁵. The former includes multiple text genres (legal texts, news, Wikipedia articles, among others), but it mostly deals with well-edited texts and a standard language. The latter is made up of so-called user-generated contents, an in particular of Twitter posts in Italian. As using both resources for training proved to give better results when analyzing Italian tweets (Sanguinetti et al., 2018), we used the same approach in this work.

Figure 1 shows an example from the TWITTIRÒ-UD corpus⁶ in CoNLL-U format: along with the typical fields indicating the sentence id and the raw text, two resource-specific fields have been introduced, to encode the information on irony categories (described in Section 2.1) and irony activators (see Section 3).

As also described in Cignarella et al. (2019), and as expected, the main critical issues in applying the UD scheme to our corpus namely consisted in finding the proper tags and coding conventions for those linguistic phenomena typically occurring in Italian tweets. The guidelines provided in Sanguinetti et al. (2018) represented a helpful ground-

⁵More details in Cignarella et al. (2019).

⁶The id of the tweet and the user mention are encrypted due to privacy regulations. – Translation: The Democratic Party is split in two. It has never been so united. [@user].

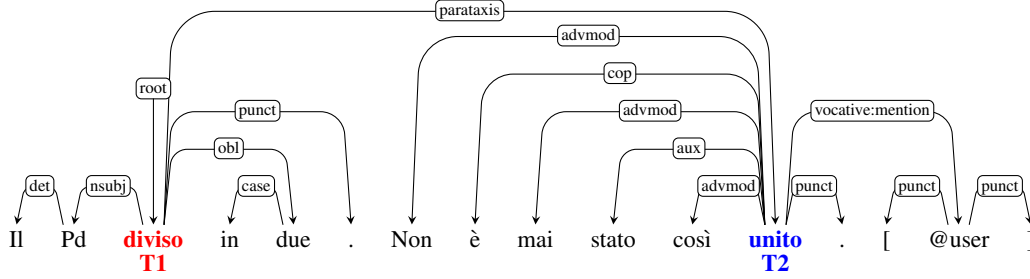


Figure 2: Dependency graph of the tweet in Figure 1 with irony activators T1 and T2 highlighted in red and blue, respectively.

work in this respect.

The fully-annotated treebank, including the annotation of irony categories, is going to be made available with the release of UD version 2.5. Due to its preliminary nature, however, the annotation of irony activators will be included in the resource at a later stage.

3 Annotating Irony Activators

As previously mentioned, irony is activated by the presence of a clash or a contradiction between two elements or two propositions (P1 and P2), which are indeed the triggers of the activation of irony. According to the scheme proposed by Karoui et al. (2017) there are two kinds of *activation types*: EXPLICIT when both these elements are lexicalized in the message, IMPLICIT otherwise.

In this step of our work, we focused our attention on the manual annotation of irony activators and on providing annotation guidelines that could be useful also for other datasets in different languages, within the same multilingual project. Indeed, the starting point of the present work is connected to the work of Karoui (2017), on a French dataset, in which the author tried to annotate at tweet level some elements that are responsible for the activation of irony. In that approach, each tweet had to be annotated using the Glozz tool (Widlöcher and Mathet, 2009), in terms of units and relationships between units (if the relationship existed). Three types of relationship were taken into account: 1) relation of comparison, 2) relation of explicit contradiction, and 3) relation of cause/consequence.

With respect to this work we opted for a finer-grained annotation also taking advantage from the availability of tokenized data and a full syntactic analysis in UD format.

3.1 Our approach

Our aim is to annotate irony activators in the whole TWITTIRÒ-UD corpus. Differently from what proposed in Karoui (2017), in which the elements creating an ironic contrast (P1 and P2) could be words, phrases or even full sentences; in this work, since we want to highlight the interaction between the pragmatic phenomenon of irony and its syntactic representation, we define as irony activators a pair of words T1 and T2 that must correspond to nodes of the syntactic dependency tree.

Given an ironical utterance (in our case a tweet) and its dependency-based syntactic representation, where each node in the tree structure represents a word, T1 and T2 is thus a pair of words – regardless of their grammatical category – such that:

- either they are both lexicalized (in explicit irony) or one of them is left unspecified (implicit irony);
- they act as triggers by signaling the presence of an ironic device.

The intuition behind this choice is inspired by the work of Saif et al. (2016), in which the authors underline the importance of contextual and conceptual semantics of words when calculating their sentiment, which in turn comes from the popular dictum “You shall know a word by the company it keeps!” (Firth, 1957). Our idea is, in fact, to proceed in two steps: firstly, to annotate irony triggers at token level, and subsequently to retrieve the other tokens that “keep company” to them by means of the dependency relations available from the UD annotation.

Therefore, as we have already highlighted in Section 1.1, if any kind of “syntactic pattern” that can help us to automatically detect irony does exist, we assume this will be particularly evident in

the “syntactic circle” around the lexical elements that create a contradiction and are the lexical activators of the ironic realization, namely T1 and T2.

In the present research, being a preliminary study, and in order to validate the strengths and weaknesses of annotation guidelines for irony activators, two skilled annotators (A1 and A2) annotated a first sample of 277 tweets, focusing on the most frequent category: EX:OXYMORON PARADOX, which covers almost 20% of the whole corpus, as it is shown in Table 1 in Section 2.1. In the following sections we will describe the guidelines that emerged throughout the discussion between A1 and A2, we will discuss the most relevant comments reported by the annotators and we will comment on some examples, thus providing an evaluation and the measures of inter-annotator agreement.

3.2 Annotation process

A sample of 277 tweets, from the ironic category EX:OXYMORON PARADOX, was annotated in parallel by two skilled annotators (A1 and A2), experts both in sentiment analysis annotations and also familiar with the CoNLL-U format.

Both of them were asked, given a tweet, to annotate two words T1 and T2 that are responsible for the activation of irony, bearing in mind these basic guiding principles:

- T1 and T2 can be nodes of any type: no specific constraints are given on the morpho-syntactic category;
- the identification of the proper T1 and T2 is guided by the irony category: for example, if the ironic tweet fits the category *oxymoron/paradox*, select the activators so that the type of relation triggered will be a contrast or a contradiction:


 la cosa bella del governo Monti è che ha acceso^{T1} le speranze di tutti e le spegnerà^{T2} pure ...
→ *the good thing about the Monti government is that it has kindled everyone's hopes and it will stifle them as well*

Figure 2 provides an example of annotated tweet, where the words *diviso* (divided) and *unito* (united) have been annotated as T1 and T2, respectively. From a procedural perspective, since the

tokens “diviso” and “unito” are respectively at position 3 and 12 in the CoNLL-U format (cfr. Figure 1), annotators were asked to add a line in the header of the annotation file, such as this one:

activators = 3 12

Furthermore, the annotators were asked to annotate any kind of doubt it might occur to them in order to provide material to a discussion about the efficacy of the guidelines.

3.3 Evaluation and Agreement

In a first phase, the annotators sketched a draft of the guidelines for the annotation of ironic activators T1 and T2, and, as a pilot experiment, they tested their efficacy on a sample of 50 tweets. Discussing the uncertain cases and the instances in disagreement helped to significantly improve the quality of the annotation choices between A1 and A2. In fact, after the first “training phase”, the guidelines were cleared up, and the annotators could proceed to annotate all the 277 OXYMORON PARADOX tweets. The inter-annotator agreement (IAA) on the 277 tweets was later calculated by means of simple observed agreement (expressed in percentage).

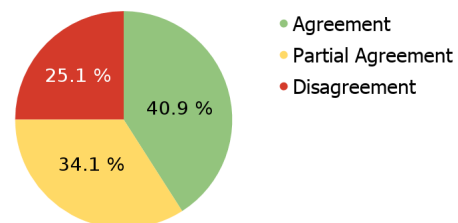


Figure 3: Observed IAA on 277 tweets.

As we can see from Figure 3 a complete agreement was immediately reached on 113 tweets (40.9%), other 94 tweets (34.1%) were in partial agreement (meaning that the annotators agreed only on T1 or T2), while 69 (25%) presented a complete disagreement.


After the first annotation step was completed and the agreement was calculated, the annotators tried to solve the partial disagreement. As a result, the percentage of T1-T2 pairs where agreement has been reached went up to approximately 69.2% (191 tweets), while the proportion of complete disagreement rose to approximately 30.8% (85 tweets).

4 Discussion


Overall, the outcome of the experimental annotation of irony activators is rather encouraging. Not only from a quantitative perspective (see Section 3.3), but also from a qualitative point of view. In fact, annotators pointed out several difficult cases, but in general they were able to find an agreement discussing the possibilities within the few restrictions posed by the guidelines.

Among the unresolved cases of disagreement (difficult cases) we were able to find recurring patterns, that need to be addressed adding new specific rules before continuing with the annotation on the rest of the dataset. Below we provide a short description.

More than two irony activators For instance, in the following tweet a list of names is presented. The contrast is created with *migliori* (best) and all three entities, but it is difficult to only choose one.


 Fantagoverno. **Fabio Volo^{T1}**, **Giovanni Sartori^{T1}**, **Roberto Saviano^{T1}**: ecco il governo dei **Migliori^{T2}** Mario Monti ... URL
→ *Fantagovernment. Fabio Volo, Giovanni Sartori, Roberto Saviano: here is the government of the best Mario Monti... URL*

Multiple categories There is more than one ironic category (e.g. overlap between an ANALOGY and a PARADOX). Such as in the tweet below, in which there is a clear analogy between Superman and Mario Monti; but also the paradoxical sentence “if you didn’t exist you should be invented!” referred to a country (Italy), which, of course already exists.


 E vai adesso con **Mario Monti^{T1}**/**Superman^{T2}**, crisi finita, stipendi in aumento, e riforme. Grazie **Stato^{T1}**! **Se non ci fossi bisognerebbe inventarti!^{T2}**
→ *And now let's go with Mario Monti/Superman, the crisis is over, the salaries are raising, and there are reforms. Thank you country! If you didn't exist you should be invented!*

Paraprosdokian There is a peculiar kind of ironic production, known in literature as “*paraprosdokian*”, in which the latter part of a sentence is surprising or unexpected in a way that causes the reader or listener to reinterpret the first part. This kind of ironic production is not specif-

ically taken into account in the annotation scheme.

 I Soliti Idiotti in scena a **Sanremo^{T1}**. Ieri erano alla **Camera^{T2}**. [@user] #dopofestival
→ *The Usual Idiots on Sanremo's stage. Yesterday there were at the Chamber of Deputies. [@user] #afterfestival*

Different activation type The tweet has been annotated as EXPLICIT, but the elements that create the ironic clash are to be found in the outer world (world knowledge is needed).

 #labuonascuola è avere una scuola.
→ *#thegoodschool is to have a school.*

5 Conclusion

In this article we described the preliminary steps of the annotation process of irony activators in the TWITTIRÒ-UD corpus, a novel Italian treebank of ironic tweets. In particular, we described the problems that emerged during the first annotation phase, the strengths and weaknesses of the scheme itself, in order to highlight future research directions. Being a preliminary study, and having no benchmark to compare with, the results obtained in the observed agreement are rather promising; moreover, the tweets included in TWITTIRÒ were retrieved from different pre-existing Italian corpora (as described in Cignarella et al. (2017)): the heterogeneous sources the data were gathered from thus represents a signal of the potential portability of the scheme and paves the way for a more systematic annotation process of the whole dataset. The next steps will then consist in the guidelines improvement and the annotation of the remaining part of TWITTIRÒ-UD accordingly.

Furthermore, the availability of English and French datasets annotated with the same scheme described in Section 2.1 (see Karoui et al. (2017)) allows the direct applicability of the annotation of irony activators in other languages than Italian. While this can be considered a further validation step to test the overall validity and portability of the scheme, it may also provide useful insights into the linguistic mechanisms underlying verbal irony in different languages.

The actual usability of this kind of resources will be finally tested when training NLP tools for irony detection, in both mono- and multi-lingual settings.

Acknowledgments

The work of C. Bosco and M. Sanguinetti was partially funded by Progetto di Ateneo/CSP 2016 (*Immigrants, Hate and Prejudice in Social Media*, S1618L2BOSC01). The work of P. Rosso was partially funded by the Spanish MICINN under the research project MISMI-S-FAKEHATE on MIS-information and MIScommunication in social media: FAKE news and HATE speech (PGC2018-096212-B-C31).

References

- Francesco Barbieri, Horacio Saggion, and Francesco Ronzano. 2014. Modelling sarcasm in Twitter, a novel approach. In *Proceedings of the 5th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 50–58. ACL.
- Francesco Barbieri, Valerio Basile, Danilo Croce, Malvina Nissim, Nicole Novielli, and Viviana Patti. 2016. Overview of the EVALITA 2016 SENTiment POLarity Classification Task. In *Proceedings of Third Italian Conference on Computational Linguistics (CLiC-it 2016) & Fifth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian (EVALITA 2016)*, volume 1749. CEUR-WS.org.
- Valerio Basile, Andrea Bolioli, Malvina Nissim, Viviana Patti, and Paolo Rosso. 2014. Overview of the EVALITA 2014 SENTiment POLarity Classification task. In *Proceedings of the 4th Evaluation Campaign of Natural Language Processing and Speech tools for Italian (EVALITA 2014)*. ELRA.
- Farah Benamara, Cyril Grouin, Jihen Karoui, Véronique Moriceau, and Isabelle Robba. 2017. Analyse d’opinion et langage figuratif dans des tweets: présentation et résultats du Défi Fouille de Textes DEFT2017. In *Actes de l’atelier DEFT2017 associé à la conférence TALN*. Association pour le Traitement Automatique des Langues (ATALA).
- Cristina Bosco, Viviana Patti, and Andrea Bolioli. 2013. Developing Corpora for Sentiment Analysis: The Case of Irony and Senti-TUT. *IEEE Intelligent Systems*, 28(2):55–63.
- Alessandra Teresa Cignarella and Cristina Bosco. 2019. ATC at IroSvA 2019: Shallow Syntactic Dependency-based Features for Irony Detection in Spanish Variants. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2019)*, co-located with 34th Conference of the Spanish Society for Natural Language Processing (SEPLN 2019). CEUR-WS.org.
- Alessandra Teresa Cignarella, Cristina Bosco, and Viviana Patti. 2017. TWITTIRÒ: a Social Media Corpus with a Multi-layered Annotation for Irony. In *Proceedings of the Fourth Italian Conference on Computational Linguistics (CLiC-it 2017)*, volume 2006, pages 101–106. CEUR-WS.org.
- Alessandra Teresa Cignarella, Cristina Bosco, Viviana Patti, and Mirko Lai. 2018a. Application and Analysis of a Multi-layered Scheme for Irony on the Italian Twitter Corpus TWITTIRÒ. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*, pages 4204–4211. ELRA.
- Alessandra Teresa Cignarella, Simona Frenda, Valerio Basile, Cristina Bosco, Viviana Patti, Paolo Rosso, et al. 2018b. Overview of the EVALITA 2018 task on Irony Detection in Italian Tweets (IronITA). In *Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian (EVALITA 2018)*, volume 2263, pages 1–6. CEUR-WS.org.
- Alessandra Teresa Cignarella, Cristina Bosco, and Paolo Rosso. 2019. Presenting TWITTIRÒ-UD: An Italian Twitter Treebank in Universal Dependencies. In *Proceedings of SyntaxFest 2019*.
- John R Firth. 1957. A synopsis of linguistic theory, 1930–1955. *Studies in linguistic analysis*.
- Simona Frenda and Viviana Patti. 2019. Computational models for irony detection in three spanish variants. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2019)*, co-located with 34th Conference of the Spanish Society for Natural Language Processing (SEPLN 2019). CEUR-WS.org.
- Aniruddha Ghosh, Guofu Li, Tony Veale, Paolo Rosso, Ekaterina Shutova, John Barnden, and Antonio Reyes. 2015. Semeval-2015 Task 11: Sentiment Analysis of Figurative Language in Twitter. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval-2015)*, pages 470–478. ACL.
- Delia Irazú Hernández Farías, José-Miguel Benedí, and Paolo Rosso. 2015. Applying basic features from sentiment analysis for automatic irony detection. In *Pattern Recognition and Image Analysis*, volume 9117, pages 337–344. Springer.
- Delia Irazú Hernández Farías, Viviana Patti, and Paolo Rosso. 2016. Irony detection in Twitter: The role of affective content. *ACM Transactions on Internet Technology (TOIT)*, 16(3):19.
- Delia Irazú Hernández Farías and Paolo Rosso. 2016. Irony, Sarcasm, and Sentiment Analysis. In *Sentiment Analysis in Social Networks*, pages 113–128. Elsevier Science and Technology.
- Yu-Hsiang Huang, Hen-Hsen Huang, and Hsin-Hsi Chen. 2017. Irony detection with attentive Recurrent Neural Networks. In *European Conference on Information Retrieval*, pages 534–540. Springer.

- Aditya Joshi, Vinita Sharma, and Pushpak Bhattacharyya. 2015. Harnessing context incongruity for sarcasm detection. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 757–762. ACL, July.
- Aditya Joshi, Pushpak Bhattacharyya, and Mark J. Carman. 2017. Automatic sarcasm detection: A survey. *ACM Computing Surveys (CSUR)*, 50(5):73.
- Jihen Karoui, Farah Benamara, Véronique Moriceau, Nathalie Aussenac-Gilles, and Lamia Hadrich Belguith. 2015. Towards a contextual pragmatic model to detect irony in tweets. In *53rd Annual Meeting of the Association for Computational Linguistics (ACL 2015)*.
- Jihen Karoui, Farah Benamara, Véronique Moriceau, Viviana Patti, and Cristina Bosco. 2017. Exploring the Impact of Pragmatic Phenomena on Irony Detection in Tweets: A Multilingual Corpus Study. In *Proceedings of the European Chapter of the Association for Computational Linguistics*.
- Jihen Karoui. 2017. *Détection automatique de l’ironie dans les contenus générés par les utilisateurs*. Ph.D. thesis, Université Toulouse 3 Paul Sabatier; Faculté des Sciences Economiques et de Gestion, Université de Sfax (Tunisie).
- Joakim Nivre, Marie-Catherine De Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic, Christopher D. Manning, Ryan T. McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, et al. 2016. Universal Dependencies v1: A Multilingual Treebank Collection. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC-2016)*.
- Reynier Ortega, Francisco Rangel, Irazú Hernández, Paolo Rosso, Manuel Montes, and José E. Medina. 2019. Overview of the Task on Irony Detection in Spanish Variants. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2019), co-located with 34th Conference of the Spanish Society for Natural Language Processing (SEPLN 2019)*. CEUR-WS.org.
- Kumar Ravi and Vadlamani Ravi. 2017. A novel automatic satire and irony detection using ensemble feature selection and data mining. *Knowledge-Based Systems*, 120:15–33.
- Antonio Reyes, Paolo Rosso, and Tony Veale. 2013. A Multidimensional Approach for Detecting Irony in Twitter. *Language Resources and Evaluation*, 47(1):239–268.
- Ellen Riloff, Ashequl Qadir, Prafulla Surve, Lalin-dra De Silva, Nathan Gilbert, and Ruihong Huang. 2013. Sarcasm as Contrast between a Positive Sentiment and Negative Situation. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, (EMNLP 2013)*, pages 704–714. ACL.
- Hassan Saif, Yulan He, Miriam Fernandez, and Harith Alani. 2016. Contextual Semantics for Sentiment Analysis of Twitter. *Information Processing & Management*, 52(1):5–19.
- Manuela Sanguinetti, Cristina Bosco, Alberto Lavelli, Alessandro Mazzei, Oronzo Antonelli, and Fabio Tamburini. 2018. PoSTWITA-UD: an Italian Twitter Treebank in Universal Dependencies. In *Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC-2018)*.
- Maria Simi, Cristina Bosco, and Simonetta Montemagni. 2014. Less is more? Towards a reduced inventory of categories for training a parser for the Italian Stanford Dependencies. In *Language Resources and Evaluation 2014*, pages 83–90. European Language Resources Association (ELRA).
- Milan Straka and Jana Straková. 2017. Tokenizing, POS Tagging, Lemmatizing and Parsing UD 2.0 with UDPipe. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 88–99. ACL.
- Cynthia Van Hee, Els Lefever, and Véronique Hoste. 2018. SemEval-2018 Task 3: Irony detection in English Tweets. In *In Proceedings of the International Workshop on Semantic Evaluation (SemEval-2018)*. ACL.
- Angela P. Wang. 2013. #Irony or #Sarcasm — A Quantitative and Qualitative Study Based on Twitter. In *Proceedings of the PACLIC: the 27th Pacific Asia Conference on Language, Information, and Computation*, pages 349–356. ACL.
- Antoine Widlöcher and Yann Mathet. 2009. La plate-forme glozz: environnement d’annotation et d’exploration de corpus. In *Actes de TALN 2009*.
- Chuhan Wu, Fangzhao Wu, Sixing Wu, Junxin Liu, Zhigang Yuan, and Yongfeng Huang. 2018. Thun_gn at semeval-2018 task 3: Tweet irony detection with densely connected LSTM and multi-task learning. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 51–56.
- Shiwei Zhang, Xiuzhen Zhang, Jeffrey Chan, and Paolo Rosso. 2019. Irony detection via sentiment-based transfer learning. *Information Processing & Management*, 56(5):1633–1644.

Robospierre, an Artificial Intelligence to Solve “La Ghigliottina”

Nicola Cirillo

University of Salerno
Salerno, Italy

n.cirillo9@studenti
.unisa.it

Chiara Pericolo

University of Salerno
Salerno, Italy

c.pericolo@studenti
.unisa.it

Pasquale Tufano

University of Salerno
Salerno, Italy

p.tufano@studenti.u
nisa.it

Abstract

This paper describes Robospierre a system developed to solve the language game “La Ghigliottina” (the guillotine). To find the solution of a game instance, it relies on MWEs automatically extracted through a lexicalized association rules algorithm; on a list of proverbs; and on some lists of titles.

1 Introduction

“La Ghigliottina” is the final game of “L’Eredità”, an Italian quiz show. In this game, the player should find a word linked to a set of five clue words. For example, if these words are *table*, *works*, *watch*, *Premier League* and *police*, the player should give as solution the word *calendar*. The link between a clue and the solution is usually the fact that both these words are part of an MWE (Multi-Word Expression) e.g. *table* and *calendar* are linked because they are part of the MWE *table calendar*. However, there can be also other kind of links. For example, the two words can be both part of a proverb (e.g. *bird* and *world* in the proverb “early bird catches the world”), of a film title (e.g. *river* and *return* in “River of No Return”) or they can be linked semantically (e.g. *Suarez* and *bite* because of the Suarez’s bite to Chiellini during the 2014 World Cup). The task of solving this game was presented as the NLP4FUN task of Evalita 2018 (Basile et al., 2018).

To build our system, first, we collected and analyzed a corpus of 296 game instances: 146 from the tv show and 150 from the board game. Second, we built an association matrix launching a lexicalized association rules algorithm, developed by us, on Paisà (Lyding et al., 2014). Then, we collected from the web a list of titles of books, films, plays and songs; and a list of proverbs. Fi-

nally, we tested the system on the game instances collected and we compared it with other artificial players of “La Ghigliottina”, especially UN-IOR4NLP (Sangati, Pascucci and Monti, 2018), that obtained the best performance on this task at Evalita 2018 (Basile et al., 2018).

2 Related Works

In the field of AI (Artificial Intelligence), games have ever provided challenging tasks that encouraged researchers to develop better and better systems (Yannakakis and Togelius, 2018). In regard to language games, worth citing is the IBM Watson system designed to play Jeopardy!™ (Ferrucci et al., 2013). However, only recently, the task of solving “La Ghigliottina” has attracted the attention of researchers. Besides a first attempt in 2009 (Semeraro et al., 2009), the research on this topic began in 2018 when this task was proposed at the Evalita evaluation campaign (Basile et al., 2018).

2.1 Game Analysis

Sangati, Pascucci and Monti (2018) showed that “the words in the clues are typically nouns, verbs or adjectives, while the ones in the solutions are typically nouns or adjectives (never verbs)”. They also stated that “in most cases each clue word is connected with the solution because they form an MWE”. However, MWEs are not the only possible associations, some game instances require difficult inferences in order to be solved. (Basile et al., 2018).

2.2 Artificial Players

The first artificial player of “La Ghigliottina” is OTTHO (Semeraro et al., 2009; Basile et al., 2016) which employs an association matrix that uses a spreading activation model on a knowledge repository to compute the degree of correlation between two terms (the repository was built using web sources like Wikipedia). During Evalita 2018

(Basile et al., 2018) two artificial players were presented: UNIOR4NLP (Sangati, Pascucci and Monti, 2018) and the system developed by Squadrone (2018). The first is based on MWEs. It employs an association-score matrix that was populated computing the PMI (Pointwise Mutual Information) measure for each pair of words. In computing this measure, only co-occurrences in specific patterns (that represents MWEs) were considered. The second system is based on an algorithm that works in two steps. First, the system extracts a set of possible solutions from a knowledge base using the five clue words. Then, the algorithm verifies the existence of proverbs, aphorisms, and titles in which the possible solutions and the clues co-occur.

3 Our Approach

Our approach is quite similar to the approach of Sangati, Pascucci and Monti (2018) since it also relies on MWEs and makes use of an association matrix to find the solution of the game. However, there are some differences between our approach and theirs.

First, we used MWEs only to find links between two words in Italian corpora while UNIOR4NLP used them also to find associations in other resources like titles and proverbs (Sangati, Pascucci and Monti, 2018). We decided that, in a title and in a proverb, a simple co-occurrence is a valid link. In fact, there are game instances in which a clue is linked to the solution because both appear in the same title or proverb, even if they do not form an MWE. For example, in a game instance, the clue *occasione* (opportunity) is linked to the solution *ladro* (thief) because both appear in the famous Italian proverb “l’occasione fa l’uomo ladro” (opportunity makes a thief) even if they do not form any MWE.

In regard to the links extracted from Italian corpora, we used association rules (Agrawal and Srikant, 1994) instead of PMI. We decided to use this measure because, in MWEs, there is a head and the rest of the expression depends on it. For example, in the MWE *pesca con la mosca* (fly fishing), the word sequence *con la mosca* (with the fly) rarely appear without the noun *pesca* (fishing | peach). However, the noun *pesca* will appear a lot of times without being followed by the word sequence *con la mosca*. The PMI between the terms *pesca* and *mosca* will be low because the noun *pesca* has a relatively high fre-

quency. Conversely, with association rules, this same link will be considered much stronger.

Another difference is that we produced a rule for every MWE and then the link between two words is defined as the score of the rule that has the highest score among all the rules in which one word appear in the consequent and the other in the antecedent (see Subsection 4.1). On the other hand, Sangati, Pascucci and Monti (2018) computed a single PMI value between two words considering all the MWEs in which these words occur. If the two systems compute the link between the words *dare* (to give) and *mano* (hand) and, in the corpus, these two words occur in the MWEs *dare una mano* (give a hand | to help) and *dare la mano* (hold hands). UNIOR4NLP will consider both these MWEs in computing the PMI between *dare* (to give) and *mano* (hand) while our system will generate two different rules: (*una mano* → *dare*) and (*la mano* → *dare*), then it will assign at the link between *dare* and *mano* the highest score between the scores of the two rules. This means that probably UNIOR4NLP will give at this link a higher score than our system.

The last difference is that Sangati, Pascucci and Monti (2018) prioritized the strength of the links over their number while we did the opposite. In fact, they considered all the words linked to each other with at least a minimum score. In this way, it is impossible to determine the number of clues to which a word is linked because every word is always linked with all the five clues. Conversely, in our system, a word is usually linked with only a subset of words. Given a game instance, our system tends to answer with a word that is linked to as many clues as possible.

4 System Description

Robospierre is composed of a scoring system and 7 linguistic resources: an association matrix, a list of proverbs, 5 lists of titles and a list of compound words. This system takes in input a set of five clues that represents a game instance. For each clue, it extracts from the resources all the words that are linked to that clue. Then, a score value is assigned to each word (it represents the strength of that link). The words extracted in this way form the set of candidate solutions. This set is then processed by the scorer that ranks each candidate solution according to the strength of the links between it and the five clues. Finally, the answer produced by the system is the candidate solution that has the highest rank.

4.1 Association Matrix

The association matrix is an $S-C$ matrix where S is the set of candidate solutions and C is the set of possible clues. To list the possible clues, we took the words whose lemma occurs in Paisà (Lyding et al., 2014) at least 10 times. Then we performed the POS tagging on these lemmas with Nooj (Silberstein, 2018) using as lexical resources `_Sdic_it.nod`, `Dnum.nom`, `tronche.nod`, `toponimi.nod`, `ElisioniContrazioni.nod` and as syntactic resources `DNUM.nog` (Vietri, 2014). From the list obtained, we extracted only nouns, adjectives, verbs, and prepositions and then we inflected them (with Nooj). On the other hand, the set of candidate solutions is a subset of the set of possible clues containing only nouns and adjectives.

To populate the matrix, we developed a lexicalized association rules algorithm based on Apriori (Agrawal and Srikant, 1994). In our algorithm, a rule is an implication $A \rightarrow B$ where A and B are sequences of words. To generate the possible rules, our algorithm uses a function written by us: *genMWE*. This function takes five arguments: D , *antecedent*, *consequent*, *position* and *lemmatize*. D is a text; *antecedent* and *consequent* are sequences of POS tags that represent respectively the possible antecedents and the possible consequents of the rules. The argument *position* tells the function where it must search for the consequent in relation to the position of the antecedent. It can take the values *forward*, *backward* and *both*. The value *forward* means that the consequent directly follows the antecedent in the text, the value *backward* means that the consequent directly precedes the antecedent and the value *both* means that the consequent can either follow or precede the antecedent. The argument *lemmatize* can take a Boolean value. If it takes *true*, the antecedents of all the rules will be lemmatized. For example, if we run the function on a text with parameters *antecedent* = *PREP N*, *consequent* = *N*, *position* = *backward* and *lemmatize* = *false*; it will generate rules such as (*di credito* \rightarrow *carta*) (credit card), (*di credito* \rightarrow *carte*) (credit cards), (*da guardia* \rightarrow *cane*) (watchdog), etc. Table 1 shows the parameters used in our experiment. While the algorithm is generating the candidate rules, it counts the occurrences of every rule ($ws_j \rightarrow ws_i$) and the occurrences of the word sequences ws_j that match the pattern of POS tags given as consequent. Finally, the algorithm computes, for every rule, the confi-

Rules	Position	Lemmatize	Example
$N \rightarrow N$	both	False	lupo \rightarrow cane
$A \rightarrow N$	both	False	intenzioni \rightarrow buone
$PREP N \rightarrow N$	backward	False	di vista \rightarrow punto
$PREP DET N \rightarrow N$	backward	False	con la mosca \rightarrow pesca
$CONG N \rightarrow N$	backward	False	e gatti \rightarrow cani
$N \rightarrow PREP$	backward	False	permesso \rightarrow con
$N \rightarrow V$	backward	True	via \rightarrow andare
$DET N \rightarrow V$	backward	True	la spugna \rightarrow gettare
$PREP N \rightarrow V$	backward	True	con mano \rightarrow toccare
$PREP DET N \rightarrow V$	backward	True	per i fondelli \rightarrow prendere

Table 1: Parameters given to the *genMWE* function

dence (1), the lift (2) and a score value (3) used to solve the game instances.

$$conf_r = \frac{\text{Count}(ws_i, ws_j)}{\text{Count}(ws_j)} \quad (1)$$

$$lift_r = \frac{conf_r}{P(ws_i)} \quad (2)$$

$$score_r = \text{Count}(ws_i, ws_j) \cdot conf_r \times 100 \quad (3)$$

We pruned the rules that disrespect one or more of the following constraints:

- $\text{Count}(ws_i, ws_j) > 1$
- $conf_r > 0.001$
- $lift_r > 1$
- $score_r > 2$

Once generated the rules, the score of a link in the association matrix between a pair of words w_i, w_j is defined in the following equation (4).

$$score_{w_i, w_j} = \max_{r \in R_i \subseteq R} (score_r) \quad (4)$$

Where R_i is a subset of R containing all the rules in which the word sequence ws_i includes the word w_i or the word w_j and the word sequence ws_j includes the other word of the pair. If there are no rules with this feature, the two words w_i, w_j are not linked to each other.

To populate the association matrix, we ran this algorithm on the Paisà corpus (Lyding et al., 2014).

4.2 Lists

To handle the links where the two words are part of a proverb or of a title, we collected from the web the following lists:

- Proverbs: A list of 2048 Italian proverbs collected from Wikiquote.¹
- Films: A list of 13098 film titles collected from Film.it.²
- Books: A list of 1633 book titles collected from Cultura&Svago.³
- Songs: A list of 984 Italian song titles collected from various web sources.⁴
- Plays: A list of 739 play titles collected from Wikipedia.⁵

We consider linked two words that appear in the same element of one of these lists. We assigned at these links a fixed score value (see Subsection 5.1).

4.3 Compound Words

The link between a clue and the solution can be also the fact that both the words appear in a compound word. For example, the words *police* and *man* are linked because they appear in the compound word *policeman*. However, there are game instances where the two words appear concatenated in a word that is not a compound. For example, *franco* (frank) and *forte* (strong) can be linked because of the word *Francoforte* (Frankfurt) although this word is not a compound.

¹ Wikiquote. Proverbi italiani.

https://it.wikiquote.org/wiki/proverbi_italiani

² Film.it, Film A-Z.

<https://www.film.it/film/film-a-z/>

³ Cultura&Svago, Mille titoli letteratura mondiale.

<https://www.culturaesvago.com/mille-titoli-letteratura-mondiale/>

⁴ Il blog di Alessandro Paldo, Le 1000 canzoni italiane più belle di sempre.

<http://alessandro-paldo.blogspot.com/2013/10/1-10-1.html?m=1>

Panorama, Le 100 canzoni italiane più belle del ventesimo secolo (fino ad ora...).

<https://www.panorama.it/musica/le-100-canzone-italiane-piu-belle-del-ventunesimo-secolo/>

Le Canzoni d'Amore, Canzoni d'amore Italiane: una lista di brani tra i più belli di sempre.

<http://www.lecanzonidamore.it/canzoni-d-amore-italiane/classifiche-italiane/250-canzone-d-amore-italiane-una-lista-di-brani-tra-i-piu-belli-di-sempre.html>

⁵ Wikipedia, Elenco di opere teatrali.

https://it.wikipedia.org/wiki/Progetto:Teatro/Elenco_di_opere_teatrali

To handle these links, we consider linked two words that appear compounded in a noun listed in the set of possible clues used in the association matrix (see Subsection 4.1). We assigned at this links a fixed score value (see Subsection 5.1).

4.4 Scoring System

Given five clues (a game instance), our system uses the resources presented above to rank the possible solutions and give an answer. This occurs in six steps:

1. For every clue $c \in C$, it generates a set of candidate solutions S finding all the words linked to c in the matrix, in the lists, and in the compound words.
2. It generates, for every candidate solution $s \in S$ a set of scores $V_{s,c}$ that contains a score for every resource in which the clue c and the candidate solution s are linked (5).

$$V_{s,c} = \{score_{s,c,1}, score_{s,c,2}, \dots, score_{s,c,n}\} \quad (5)$$

3. From the set of scores of every candidate solution, the system keeps only the highest (6).

$$v_{s,c} = \max_{i=1,n}(score_{s,c,i}) \quad (6)$$

4. Then, it standardizes every score in an interval (between 0 and 100) and adds to the value obtained a bonus of 100 that represents the existence of a link between that candidate solution and the clue (7)(8)(9).

$$max = \max_{s \in S}(v_{s,c}) \quad (7)$$

$$min = \min_{s \in S}(v_{s,c}) \quad (8)$$

$$std_{s,c} = \left(\frac{v_{s,c} - min}{max - min} \times 100 \right) + 100 \quad (9)$$

5. Once completed the steps 1-4 for all the clues in the game instance, the system sums all the scores of that candidate solution to produce its final score f_s (10).

$$f_s = \sum_{c \in C} std_{s,c} \quad (10)$$

6. The answer given by the system is the candidate solution that obtains the highest final score value (11).

$$\hat{s} = \underset{s \in S}{\operatorname{argmax}}(f_s) \quad (11)$$

5 System Evaluation

To evaluate the artificial players of “La Ghigliottina” Basile et al. (2018) made use of the MRR (Mean Reciprocal Rank) measure weighted by a function that lower the score according to the time taken by the system to provide the answer (12).

$$MRR = \frac{1}{|G|} \sum_{g \in G} \frac{1}{r_g} \max\left(\frac{1}{r_g}, \frac{1}{10}\right) \quad (12)$$

In this equation, G is the set of game instances, r_g is the rank that the solution of the game g has in the set of answers produced by the system, and t_g is the time (in minutes) that the system takes to provide the set of answers (Basile et al., 2018).

The first 100 answers that the system provides are considered in computing the MRR and a game instance is considered solved when the solution is among these 100 answers. According to this evaluation, UNIOR4NLP (Sangati, Pascucci and Monti, 2018) obtained an MRR of 0.6428 and solved the 81.90% of the game instances while Squadrone (2018) obtained an MRR of 0.0134 and solved the 25.71% of the game instances.

Basile et al. (2016) evaluated OTTHO using the precision-k measure. A game is considered k-solved if the solution has rank k or higher in the set of answers provided by the system (13).

$$\text{precision} - k = \frac{k\text{-solved game instances}}{\text{total game instances}} \quad (13)$$

With $k = 1$, the best model of OTTHO obtained a precision of about 0.25 on tv games and about 0.30 on board games. With $k = 100$, it obtained a precision of about 0.50 on tv games and about 0.70 on board games (Basile et al., 2016).

In order to evaluate our system, we collected 294 game instances where the solution was provided: 146 from the tv show and 150 from the board game. Then, we submitted them to the system and computed the MRR (12) considering only the first 100 candidates solutions ranked according to their final scores (10).

To see how the different linguistic resources af-

	All	Tv	Board game
MRR	0.4140	0.4794	0.3660
Correct Answers	72.30%	80.82%	64.00%

Table 2: Result of first test

fect the performance, we tested different version of our system: one with only the association matrix; one with the association matrix and the compound words; and one with the matrix, the compound words and the lists of titles that represents the full system.

Finally, in order to compare our system to UNIOR4NLP (Sangati, Pascucci and Monti, 2018), we submitted the same game instances to the Telegram bot version of UNIOR4NLP and then we computed the precision-k (13) of the two systems for $k = 1$ (since the UNIOR4NLP bot provides only one answer).

5.1 Parameters Used in the Tests

We assigned to the links in the compound words (see Subsection 4.3) a score of 100 since these links seemed very reliable associations.

To the links in the lists of titles (see Subsection 4.2), we assigned a score of 5 because higher values seemed to worsen the performance of the system and, with lower values, the full model (matrix + compound + titles) gives the same answers of the previous one (matrix + compound).

5.2 Analysis of the Results

The result of the first test are displayed in Table 2. Our system obtained a quite good result if compared to the other systems. It was also able to provide the answer always in the first minute as UNIOR4NLP did (Basile et al., 2018). It performed better on the tv games than on the board games. Maybe because in the tv games, the links are more often based on MWEs while in the board game, there are more links based on titles, proverbs and semantic associations and our system does not treat these links as good as it treats the links based on MWEs (the links based on semantic associations are not even treated). This hypothesis is confirmed by the fact that the list of proverbs and the lists of titles worsen the performance of the system (see Table 3).

We suppose that this problem is caused by the

Models	Precision-1		
	All	Tv	Board game
Matrix	0.3480	0.4014	0.2933
Matrix + compounds	0.3514	0.4178	0.3067
Matrix + compounds + titles	0.3446	0.4178	0.3000
UNIOR4NLP	0.5608	0.6643	0.4600
	Tot (296)	Tot (146)	Tot (150)

Table 3: Result of second and third tests

fact that we assigned at every link in the lists the same score. However, there are titles and proverbs that are more likely to produce reliable links and some others that are not. The more an element is known, the more the links in it must be reliable. Maybe, assigning at every element in the lists a score that represents how much that element is known, might lead to an improvement of system performance. This score might be based on the number of results retrieved when that element is searched with a search engine like Google.

The result of the third test are displayed in Table 3. As the result show, our system was not able to reach the performance of UNIOR4NLP. However, we found among the game instances 20 games to which our system answered correctly while UNIOR4NLP did not. We will analyze some of these instances that are of particular interest.

The first is the following:

```
CLUES: cravatta; neve; S.
Martino; pizza; altare
ANSWER: pala
```

Our system gave to this game instance the correct answer *pala* (shovel | blade | altarpiece) while UNIOR4NLP gave the answer *bianca* (white). We suppose that UNIOR4NLP gave this answer because, sometimes, it overestimates the strength of a link and ignores the other links. We believe that the answer *bianca* is mainly due to the clue *neve* (snow) since UNIOR4NLP considered both the compound noun *Biancaneve* (Snow-white) and the frequent co-occurrence between the adjective *bianca* and the noun *neve* to compute the PMI between these two terms. On the other hand, our system found three weak links: between *pala* and *neve*; between *pala* and *pizza* and between *pala* and *altare* (altar). These links were sufficient to assign to this word the highest rank among the candidate answers produced.

Another interesting game instance is the following:

```
CLUES: introduzione; cowboy;
fungo; 23; fare tanto
ANSWER: cappello
```

UNIOR4NLP gave to this game instance, the answer *proiettili* (bullets). Our system gave the correct answer *cappello* (hat). Maybe, the answer of UNIOR4NLP was due to the overestimation of

the link between *proiettili* and the clue *cowboy* while it underestimated the link between this clue and the word *cappello*. We believe that this happened because *cappello* occurs in more contexts than *proiettili*. On the other hand, our system gave the correct answer *cappello* because it was strongly linked with the word sequence *da cowboy* (like cowboys) since this sequence almost always occurs in the MWE *cappello da cowboy* (cowboy hat).

The last game instances that we will analyze is the following:

```
CLUES: andare; musica; oc-
chi; mano; buona
ANSWER: palla
```

To this game instance, our system answered *palla* (ball) and UNIOR4NLP answered *pallino* (cue ball | dot). We suppose that this error is caused by the MWE *andare a pallino* (right on cue) that appear in the online dictionary “Il Nuovo De Mauro” (De Mauro, 2016) which was employed by UNIOR4NLP as linguistic resource. UNIOR4NLP considered a co-occurrence in this dictionary as strong as 200 co-occurrences in the Italian corpora so this link obtained a higher PMI than that between *andare* and *palla* but, actually, the MWE *andare in palla* (be confused) is much more common than *andare a pallino*.

6 Conclusions

We described and tested Robospierre, a system developed to solve the word game “La Ghigliottina” (the guillotine). The result of the tests showed that, even if its result were below state-of-the-art, it was able to solve some game instances that the state-of-the-art system did not solved.

In the future, we plan to improve the extraction of the links in the MWEs extracting them from a bigger corpus. We also intend to assign at every element in the list of proverbs and in the lists of titles a score that represents how much that element is known.

Reference

- Agrawal Rakesh and Srikant Ramakrishnan. 1994. “Fast algorithms for mining association rules.” *Proc. 20th int. conf. very large data bases, VLDB. Vol. 1215*.
- Basile Pierpaolo, de Gemmis Marco, Lops Pasquale, and Semeraro Giovanni. 2016. “Solving a complex

- language game by using knowledge-based word associations discovery”. *IEEE Transactions on Computational Intelligence and AI in Games* 8(1), pages 13–26.
- Basile Pierpaolo, de Gemmis Marco, Siciliani Lucia, and Semeraro Giovanni. 2018. “Overview of the evalita 2018 solving language games (nlp4fun) task.” In Caselli Tommaso, Novielli Nicole, Patti Viviana, and Rosso Paolo, editors, *Proceedings of the 6th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA’18)*, CEUR.org, Turin, Italy.
- De Mauro Tullio. 2016. *Il Nuovo De Mauro (Online)*. Available at: dizionario.internazionale.it.
- Ferrucci David A., Levas Anthony, Bagchi Sugato, Gondek David, and Mueller Erik T. 2013. “Watson: Beyond jeopardy!” In *Artif. Intell.*, 199 pages 93–105.
- Lyding Verena, Stemle Egon, Borghetti Claudia, Brunello Marco, Castagnoli Sara, Dell’Orletta Felice, Dittmann Henrik, Lenci Alessandro, and Pirrelli Vito. 2014. “The PAISÀ corpus of italian web texts.” In *Proceedings of the 9th Web as Corpus Workshop (WaC-9)*, pages 36–43. Association for Computational Linguistics, Gothenburg, Sweden.
- Sangati Federico, Pascucci Antonio, and Monti Johanna. 2018. “Exploiting multiword expressions to solve ‘La Ghigliottina’”. In Caselli Tommaso, Novielli Nicole, Patti Viviana, and Rosso Paolo, editors, editors, *Proceedings of the 6th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA’18)*, Turin, Italy. CEUR.org.B.
- Semeraro Giovanni, Lops Pasquale, Basile Pierpaolo, and De Gemmis Marco. 2009. “On the tip of my thought: playing the guillotine game.” In *Proceedings of the 21st International Joint Conference on Artificial Intelligence, IJCAI’09*, pages 1543–1548. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.
- Silberstein Max. 2018. NooJ Manual, Available for download at: www.nooj4nlp.net
- Squadrone Luca. 2018. “Computer challenges guillotine: how an artificial player can solve a complex language tv game with web data analysis.” In Caselli Tommaso, Novielli Nicole, Patti Viviana, and Rosso Paolo, editors, *Proceedings of the 6th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA’18)*, Turin, Italy. CEUR.org.H.
- Vietri Simonetta. 2014. “The italian module for nooj.” In *Proceedings of the First Italian Conference on Computational Linguistics, CLiC-it 2014*. Pisa University Press.
- Yannakakis Georgios N. and Togelius Julian. 2018. *Artificial Intelligence and Games*. Springer.

From Sartre to Frege in Three Steps: A* Search for Enriching Semantic Text Similarity Measures

Davide Colla

University of Turin,
Computer Science Department
davide.colla@unito.it

Enrico Mensa

University of Turin,
Computer Science Department
enrico.mensa@unito.it

Marco Leontino

University of Turin,
Computer Science Department
marco.leontino@unito.it

Daniele P. Radicioni

University of Turin,
Computer Science Department
daniele.radicioni@unito.it

Abstract

English. In this paper we illustrate a preliminary investigation on semantic text similarity. In particular, the proposed approach is aimed at complementing and enriching the categorization results obtained by employing standard distributional resources. We found that the paths connecting entities and concepts from documents at stake provide interesting information on the connections between document pairs. Such semantic browsing device enables further semantic processing, aimed at unveiling contexts and hidden connections (possibly not explicitly mentioned in the documents) between text documents.¹

1 Introduction

In the last few years many efforts have been spent to extract information contained in text documents, and a large number of resources have been developed that allow exploring domain-based knowledge, defining a rich set of specific semantic relationships between nodes (Vrandečić and Krötzsch, 2014; Auer et al., 2007; Navigli and Ponzetto, 2012). Being able to extract and to make available the semantic content of documents is a challenging task, with beneficial impact on different applications, such as document categorisation (Carducci et al., 2019), keyword extraction (Colla et al., 2017), question answering, text summarisation, semantic texts comparison, on building explanations/justifications for similarity judgements (Colla et al., 2018) and more. In this paper we present an approach aimed at extracting

meaningful information contained in text documents, also based on background information contained in an encyclopedic resource such as Wikidata (Vrandečić and Krötzsch, 2014).

Although our approach has been devised on a specific application domain (PhD theses in philosophy), we argue that it can be easily extended to further application settings. The approach focuses on the ability to extract relevant pieces of information from text documents, and to map them onto the nodes of a knowledge graph, obtained from semantic networks representing encyclopedic and lexicographic knowledge. In this way it is possible to compare different documents based on their graphical description, which has a direct anchoring to their semantic content.

We propose a system to assess the similarity between textual documents, hybridising the propositional approach (such as traditional statements expressed through RDF triples) with a distributional description (Harris, 1954) of the nodes contained in the knowledge graph, that are represented with word embeddings (Mikolov et al., 2013; Camacho-Collados et al., 2015; Speer et al., 2017). This step allows to obtain similarity measures (based on vector descriptions, and on path-finding algorithms) and explanations (represented as paths over a semantic network) more focused on the semantic definition of concepts and entities involved in the analysis.

2 Related Work

Surveying the existing approaches requires to briefly introduce the most widely used resources along with their main features.

Resources

BabelNet (BN) is a wide-coverage multilingual semantic network, originally built by integrating

¹Copyright © 2019 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

WordNet (Miller, 1995) and Wikipedia (Navigli and Ponzetto, 2010). NASARI is a vectorial resource whose senses are represented as vectors associated to BabelNet synsets (Camacho-Collados et al., 2015). Wikidata is a knowledge graph based on Wikipedia, whose goal is to overcome problems related to information access by creating new ways for Wikipedia to manage its data on a global scale (Vrandečić and Krötzsch, 2014).

2.1 Approaches to semantic text similarity

Most literature in computing semantic similarity between documents can be arranged into three main classes.

Word-based similarity. Word-based metrics are used to compute the similarity between documents based on their terms; examples of features analysed are common morphological structures (Islam and Inkpen, 2008) and words overlap (Huang et al., 2011) between the texts. In one of the most popular theories on similarity (the Tversky’s contrast model) the similarity of a word pair is defined as a direct function of their common traits (Tversky, 1977). This notion of similarity has been recently adjusted to model human similarity judgments for short texts: the Symmetrical Tversky Ratio Model (Jimenez et al., 2013), and employed to compute semantic similarity between word- and sense-pairs (Mensa et al., 2017; Mensa et al., 2018).

Corpus-based similarity. Corpus-based measures try to identify the degree of similarity between words using information derived from large corpora (Mihalcea et al., 2006; Goma and Fahmy, 2013).

Knowledge-based similarity. Knowledge-based measures try to estimate the degree of semantic similarity between documents by using information drawn from semantic networks (Mihalcea et al., 2006). In most cases only the hierarchical structure of the information contained in the network is considered, without considering the relation types within nodes (Jiang and Conrath, 1997; Richardson et al., 1994); some authors consider the “is-a” relation (Resnik, 1995), but leaving unexploited the more domain-dependent ones. Moreover, only concepts are usually considered, omitting the Named Entities.

An emerging paradigm is that of *knowledge graphs*. Knowledge graph extraction is a challenging task, particularly popular in recent

years (Schuhmacher and Ponzetto, 2014). Several approaches have been developed, e.g., aimed at extracting knowledge graphs from textual corpora, attaining a network focused on the type of documents at hand (Pujara et al., 2013). Such approaches may be affected by scalability and generalisation issues. In the last years many resources representing knowledge in a structured form have been proposed that build on encyclopedic resources (Auer et al., 2007; Suchanek et al., 2007; Vrandečić and Krötzsch, 2014).

As regards as semantic similarity, a framework has been proposed based on entity extraction from documents, providing mappings to knowledge graphs in order to compute semantic similarities between documents (Paul et al., 2016). Their similarity measures are mostly based on the network structure, without introducing other instruments such as embeddings, that are largely acknowledged as relevant in semantic similarity. Hecht et al. (2012) propose a framework endowed with explanatory capabilities from similarity measures based on relations between Wikipedia pages.

3 The System

In this Section we illustrate the generation process of the knowledge graph from Wikidata, which will be instrumental to build paths across documents. Such paths are then used, at a later time, to enrich the similarity scores computed during the classification.

3.1 Knowledge Graph Extraction

The first step consists of the extraction of a knowledge graph related to the given reference domain. Wikidata is then searched for concepts and entities related to the domain being analysed. By starting from the extracted elements, which constitute the basic nodes of the knowledge graph, we still consider Wikidata and look for relevant semantic relationships towards other nodes, not necessarily already extracted in the previous step. The types of relevant relationships depend on the treated domain. Considering the philosophical domain, we selected a set of 30 relations relevant to compare the documents. For example, we considered the relation *movement* that represents the literary, artistic, scientific or philosophical movement, the relation *studentOf* that represents the person who has taught the considered philosopher, and the relation *influencedBy* that represents the person’s

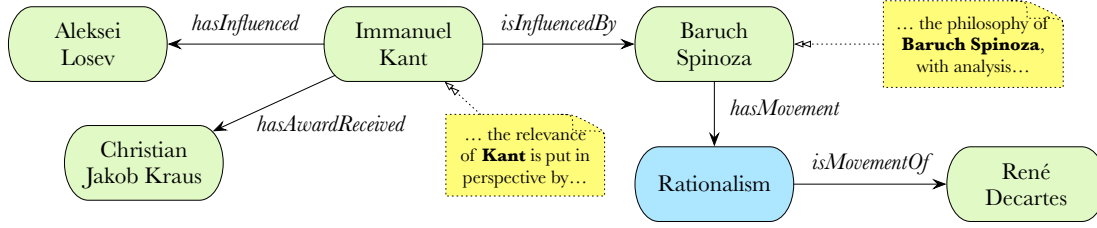


Figure 1: A small portion of the knowledge graph extracted from Wikidata, related to the philosophical domain; nodes represent BabelSynsets (concepts or NEs), rectangles represent documents.

idea from which the considered philosopher’s idea has been influenced. In this way, we obtain a graph where each node is a concept or entity extracted from Wikidata; such nodes are connected with edges labeled with specific semantic relations.

The obtained graph is then mapped onto BabelNet. At the end of the first stage, the knowledge graph represents the relevant domain knowledge (Figure 1) encoded through BabelNet nodes, that are connected through the rich set of relations available in Wikidata. Each text document can be linked to the knowledge graph, thereby allowing to make semantic comparisons by analysing the possible paths connecting document pairs.

Without loss of generality, we considered the philosophical domain, and extracted a knowledge graph containing 22, 672 nodes and 135, 910 typed edges; Wikidata entities were mapped onto BabelNet approximately in the 90% of cases.

3.2 Information extraction and semantic similarity

The second step consists in connecting the documents to the obtained knowledge graph. We harvested a set of 475, 383 UK doctoral theses in several disciplines through the Electronic Theses Online Service (ETHOS) of the British National Library.² At first, concepts and entities related to the reference domain were extracted from the considered documents, with a special focus on two different types of information, such as *concepts* and *Named Entities*. *Concepts* are keywords or multi-word expressions representing meaningful items related to the domain (such as, e.g., ‘philosophy-of-mind’, ‘Rationalism’, *etc.*) while *Named Entities* are persons, places or organisations (mostly universities, in the present setting) strongly related to the considered domain. Named entities are extracted using the Stanford CoreNLP NER module (Manning et al., 2014) improved with extrac-

tion rules based on morphological and syntactical patterns, considering for example sequences of words starting with a capital letter or associated to a particular Part-Of-Speech pattern. Similarly, we extract relevant concepts based on particular PoS patterns (such as NOUN-PREPOSITION-NOUN, thereby recognizing, for example, *philosophy of mind*).

We are aware that we are not considering the problem of word sense disambiguation (Navigli, 2009; Tripodi and Pelillo, 2017). The underlying assumption is that as long as we are concerned with a narrow domain, this is a less severe problem: e.g., if we recognise the person *Kant* in a document related to philosophy, probably the person cited is the philosopher whose name is *Immanuel Kant* (please refer to Figure 1), rather than the less philosophical Gujarati poet, playwright and essayist Kavi Kant.³

By mapping concepts and Named Entities found in a document onto the graph, we gain a set of *access points* to the knowledge graph. Once acquired the access points to the knowledge graph for a pair of documents, we can compute the semantic similarity between documents by analysing the paths that connect them.

3.3 Building Paths across Documents

The developed framework is used to compute paths between pairs of senses and/or entities featuring two given documents. Each edge in the knowledge graph has associated a semantic relation type (such as, e.g., “*hasAuthor*”, “*influencedBy*”, “*hasMovement*”). Each path intervening between two documents is in the form

$$DOC_1 \xrightarrow{ACCESS} SaulKripke \xrightarrow{influencedBy} LudwigWittgenstein \xrightarrow{influencedBy} BertrandRussell \xrightarrow{influencedBy} BaruchDeSpinoza \xleftarrow{ACCESS} DOC_2$$

²<https://ethos.bl.uk>.

³<https://tinyurl.com/y3s9lsp7>.

In this case we can argue in favor of the relatedness of the two documents based on the chain of relationships illustrating that *Saul Kripke* (from document d_1) has been *influenced-by* Ludwig Wittgenstein, that has been *influenced-by* Bertrand Russell, that in turn has been *influenced-by* Baruch De Spinoza, mentioned in d_2 . The whole set of paths connecting elements from a document d_1 to a document d_2 can be thought of as a form of evidence of the closeness of the two documents: documents with numerous shorter paths connecting them are intuitively more related. Importantly enough, such paths over the knowledge graph do not contain general information (e.g., Kant was a man), but rather they are highly domain-specific (e.g., Oskar Becker had as doctoral student Jürgen Habermas).

A* Search

The computation of the paths is performed via a modified version of the A^* algorithm (Hart et al., 1968). In particular, paths among access nodes are returned in order, from the shortest to the longest one. Given the huge dimension of the network, and since we are guaranteed to retrieve shortest paths first, we stop the search after one second of computation time.

4 Experimentation

In this Section we report the results of a preliminary experimentation: given a dataset of PhD theses, we first explore the effectiveness of standard distributional approaches to compute the semantic similarity between document pairs; we then elaborate on how such results can be complemented and enriched through the computation of paths between entities therein.

Experimental setting We extracted 4 classes of documents (100 for each class) from the EThOS dataset. For each record we retrieved the title and abstract fields, that were used for subsequent processing. We selected documents containing ‘Antibiotics’, ‘Molecular’, ‘Hegel’ or ‘Ethics’ either in their title (in 15 documents per class) or in their abstract (15 documents per class). Each class is featured on average by 163.5 tokens (standard deviation $\sigma = 39.3$), including both title and abstract. The underlying rationale has been that of selecting documents from two broad areas, each one composed by two different sets of data, having to do with medical disciplines and molecular biology in the former case, and with Hegelianism

and the broad theme of ethics in the latter case. Intra-domain classes (that is both ‘Antibiotics’-‘Molecular’ and ‘Hegel’-‘Ethics’) are not supposed to be linearly separable, as it mostly occurs in real problems. Of course, this feature makes more interesting the categorization problem. The dataset was used to compute some descriptive stats (such as inverse document frequency), characterizing the whole collection of considered documents.

From the aforementioned set of 400 documents we randomly chose a subset of 20 documents, 5 documents for each of the 4 classes from those containing the terms (either ‘Antibiotics’, ‘Molecular’, ‘Hegel’ or ‘Ethics’) in the title. This selection strategy was aimed at selecting more clearly individuated documents, exhibiting a higher similarity degree within classes than across classes.⁴

4.1 Investigation on Text Similarity with Standard Distributional Approaches

GLoVe and Word Embedding Similarity

The similarity scores were computed for each document pair with a Word Embedding Similarity approach (Agirre et al., 2016). In particular, each document d has been provided with a vector description averaging the GloVe embeddings t_i (Pennington et al., 2014) for all terms in the title and abstract:

$$\vec{N}_d = \frac{1}{|T_d|} \sum_{t_i \in T_d} \vec{t}_i, \quad (1)$$

where each \vec{t}_i is the GloVe vector for the term t_i . Considering two documents d_1 and d_2 , each one associated to a particular vector \vec{N}_{d_i} , we compare them using the cosine similarity metrics:

$$\text{sim}(\vec{N}_{d_1}, \vec{N}_{d_2}) = \frac{\vec{N}_{d_1} \cdot \vec{N}_{d_2}}{\|\vec{N}_{d_1}\| \|\vec{N}_{d_2}\|}. \quad (2)$$

The obtained similarities between each document pair are reported in Figure 2(a).⁵ The computed distances show that overall this approach is sufficient to discriminate the scientific doctoral theses from the philosophical ones. In particular, the top green triangle shows the correlation scores among antibiotics documents, while the bottom triangle reports the correlation scores among philo-

⁴In future work we will verify such assumptions by involving domain experts in order to validate and/or refine the heuristics employed in the document selection.

⁵The plot was computed using the *corrplot* package in R.

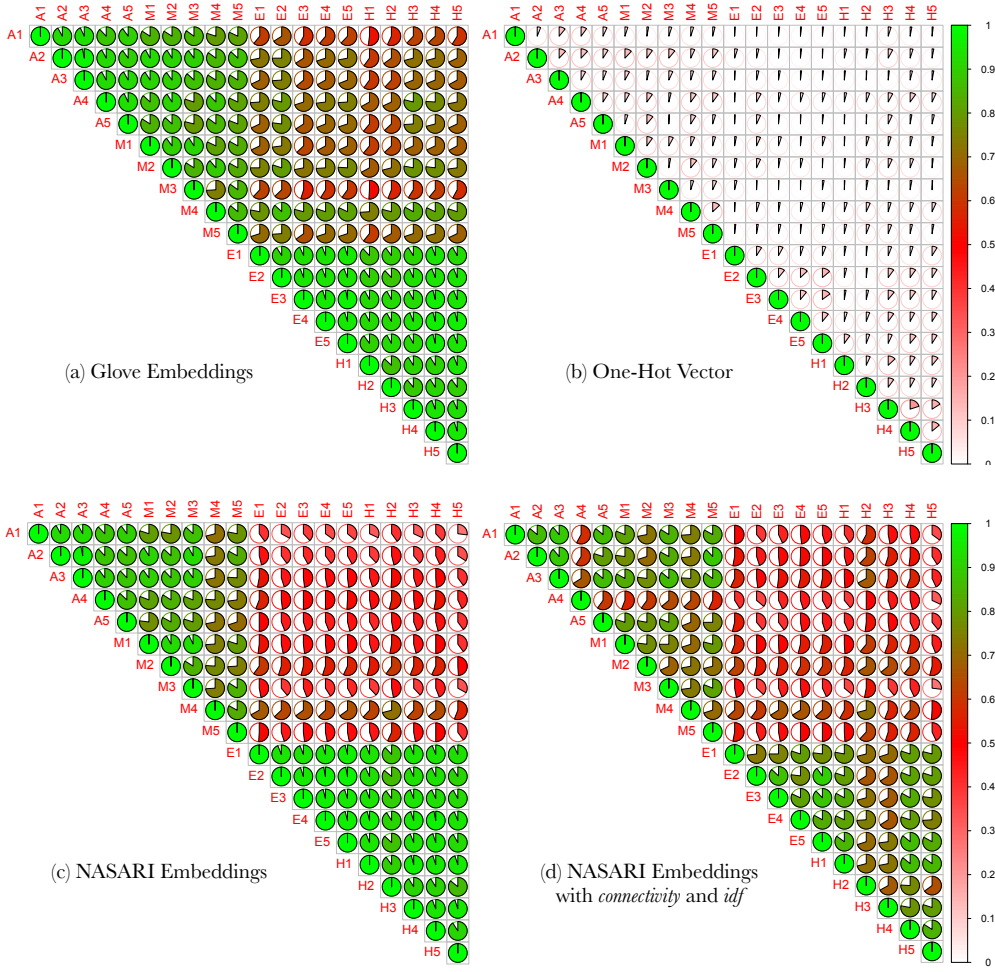


Figure 2: Comparison between correlation scores. Documents have scientific subject (‘A’ for ‘Antibiotics’, ‘M’ for ‘Molecular’ biology), and philosophic subject (‘E’ for ‘Ethics’, ‘H’ for ‘Hegel’).

sophical documents. The red square graphically illustrates the poor correlation between the two classes of documents. On the other side, the subclasses (Hegelism-Ethics and Antibiotics-Molecular) could not be separated. Provided that word embeddings are known to conflate all senses in the description of each term (Camacho-Collados and Pilehvar, 2018), this approach performed surprisingly well in comparison to a baseline based on a one-hot vector representation, only dealing with term-based features (Figure 2(b)).

NASARI and Sense Embedding Similarity

We then explored the hypothesis that semantic knowledge can be beneficial for better separating documents: after performing word sense disambiguation (the BabelFy service was employed (Moro et al., 2014)), we used the NASARI embedded version to compute the vector \vec{N}_d , as the average of all vectors associated to the senses contained in S_d , basically employing the same for-

mula as in Equation 1. We then computed the similarity matrix, displayed in Figure 2(c). It clearly emerges that also NASARI is well suited to solve a classification task when domains are well separated. However, also in this case the adopted approach does not seem to discriminate well within the two main classes: for instance, the square with vertices E1-H1; E5-H1; E5-H5; E1-H5 should be reddish, indicating a lower average similarity between documents pertaining the Hegel and Ethics classes. We experimented in a set of widely varied conditions and parameters, obtaining slightly better similarity scores by weighting NASARI vectors with senses *idf*, and senses connectivity (c , obtained from BabelNet):

$$\vec{N}_d = \frac{1}{|S_d|} \sum_{s_i \in S_d} \vec{s}_i \cdot \log \left(\frac{|S_d|}{H(s_i)} \right) \cdot \left(1 - \frac{1}{c} \right), \quad (3)$$

where $H(s_i)$ is the number of documents containing the sense s_i . The resulting similarities scores

are provided in Figure 2(d).

Documents are in fact too close, and presumably the adopted representation (merging all senses in each document) is not as precise as needed. In this setting, we tried to investigate the documents similarity based on the connections between their underlying sets of senses. Such connections were computed on the aforementioned graph.

4.2 Enriching Text Similarity with Paths across Documents

In order to examine the connections between the considered documents we focused on the philosophical portion of our dataset, and exploited the knowledge graph described in Section 3. The computed paths are not presently used to refine the similarity scores, but only as a suggestion to characterize possible connections between document pairs. The extracted paths contain precious information that can be easily integrated in downstream applications, by providing specific information that can be helpful for domain experts to achieve their objectives (e.g., in semantically browsing text documents, in order to find influence relations across different philosophical schools).

As anticipated, building paths among the fundamental concepts of the documents allows grasping important ties between the documents topics. For instance, one of the extracted paths (between the author ‘Hegel’ and the work ‘Sense and Reference’ (Frege, 1948)) shows the connections between the entities at stake as follows. G.W.F. Hegel *hasMovement* Continental Philosophy, which is in turn the *movementOf* H.L. Bergson, who has been *influencedBy* G. Frege, who finally *hasNotableWork* Sense and Reference. The semantic specificity of this information provides precious insights that allow for a proper consideration of the relevance of the second document w.r.t. the first one. It is worth noting that the fact that Hegel is a continental philosopher is trivial –tacit knowledge– for philosophers, and was most probably left implicit in the thesis abstract, while it can be a relevant piece of information for a system requested to assess the similarity of two philosophical documents. Also, this sort of path over the extracted knowledge graph enables a form of semantic browsing that benefits from the rich set of Wikidata relations paired with the valuable coverage ensured by BabelNet on domain-specific con-

cepts and entities.

The illustrated approach allows the uncovering of insightful and specific connections between documents pairs. However, this preliminary study also pointed out some issues. One key problem is the amount of named entities contained in the considered documents (e.g., E5 only has one access point, while E3 has none). Another issue has to do with the inherently high connectivity of some nodes of the knowledge graph (hubness). For instance, the nodes *Philosophy*, *Plato* and *Aristotle* are very connected, which results in the extraction of some trivial and uninteresting paths among the specific documents. The first issue could be tackled by also considering the main concepts of a document if no entity can be found, whilst the second one could be mitigated by taking into account the connectivity of the nodes as a negative parameter while computing the paths.

5 Conclusions

In this paper we have investigated the possibility of enriching semantic text similarity measures via symbolic and human readable knowledge. We have shown that distributional approaches allow for a satisfactory classification of documents belonging to different topics, however, our preliminary experimentation showed that they are not able to capture the subtle aspects characterizing documents in close areas. As we have argued, exploiting paths over graphs to explore connections between document pairs may be beneficial in making explicit domain-specific links between documents.

As a future work, we could refine the methodology related to the extraction of the concepts in the Knowledge Graph, defining approaches based on specific domain-related ontologies. Two relevant works, to these ends, are the *PhilOnto* ontology, that represents the structure of philosophical literature (Grenon and Smith, 2011), and the *InPho* taxonomy (Buckner et al., 2007), combining automated information retrieval methods with knowledge from domain experts. Both resources will be employed in order to extract a more concise, meaningful and discriminative Knowledge Graph.

Acknowledgments

The authors are grateful to the EThOS staff for their prompt and kind support. Marco Leontino has been supported by the REPOSUM project, BONG.CRT.17.01 funded by Fondazione CRT.

References

- Eneko Agirre, Carmen Banea, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Rada Mihalcea, German Rigau, and Janyce Wiebe. 2016. Semeval-2016 task 1: Semantic textual similarity, monolingual and cross-lingual evaluation. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 497–511.
- Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. 2007. Dbpedia: A nucleus for a web of open data. In *The semantic web*, pages 722–735. Springer.
- Cameron Buckner, Mathias Niepert, and Colin Allen. 2007. Inpho: the indiana philosophy ontology. *APA Newsletters-newsletter on philosophy and computers*, 7(1):26–28.
- Jose Camacho-Collados and Mohammad Taher Pilehvar. 2018. From word to sense embeddings: A survey on vector representations of meaning. *Journal of Artificial Intelligence Research*, 63:743–788.
- José Camacho-Collados, Mohammad Taher Pilehvar, and Roberto Navigli. 2015. NASARI: a novel approach to a semantically-aware representation of items. In *Proceedings of NAACL*, pages 567–577.
- Giulio Carducci, Marco Leontino, Daniele P Radicioni, Guido Bonino, Enrico Pasini, and Paolo Tripodi. 2019. Semantically aware text categorisation for metadata annotation. In *Italian Research Conference on Digital Libraries*, pages 315–330. Springer.
- Davide Colla, Enrico Mensa, and Daniele P Radicioni. 2017. Semantic measures for keywords extraction. In *Conference of the Italian Association for Artificial Intelligence*, pages 128–140. Springer.
- Davide Colla, Enrico Mensa, Daniele P. Radicioni, and Antonio Lieto. 2018. Tell me why: Computational explanation of conceptual similarity judgments. In *Proceedings of the 17th International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems (IPMU), Special Session on Advances on Explainable Artificial Intelligence*, Communications in Computer and Information Science (CCIS), Cham. Springer International Publishing.
- Gottlob Frege. 1948. Sense and reference. *The philosophical review*, 57(3):209–230.
- Wael H Gomaa and Aly A Fahmy. 2013. A survey of text similarity approaches. *International Journal of Computer Applications*, 68(13):13–18.
- Pierre Grenon and Barry Smith. 2011. Foundations of an ontology of philosophy. *Synthese*, 182(2):185–204.
- Zellig S Harris. 1954. Distributional structure. *Word*, 10(2-3):146–162.
- Peter E. Hart, Nils J. Nilsson, and Bertram Raphael. 1968. A formal basis for the heuristic determination of minimum cost paths. *IEEE Transactions on Systems Science and Cybernetics*, SSC-4(2):100–107.
- Brent Hecht, Samuel H Carton, Mahmood Quaderi, Johannes Schöning, Martin Raubal, Darren Gergle, and Doug Downey. 2012. Explanatory semantic relatedness and explicit spatialization for exploratory search. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*, pages 415–424. ACM.
- Cheng-Hui Huang, Jian Yin, and Fang Hou. 2011. A text similarity measurement combining word semantic information with tf-idf method. *Jisuanji Xuebao(Chinese Journal of Computers)*, 34(5):856–864.
- Aminul Islam and Diana Inkpen. 2008. Semantic text similarity using corpus-based word similarity and string similarity. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 2(2):10.
- Jay J Jiang and David W Conrath. 1997. Semantic similarity based on corpus statistics and lexical taxonomy. *arXiv preprint cmp-lg/9709008*.
- Sergio Jimenez, Claudia Becerra, Alexander Gelbukh, Av Juan Dios Bátiz, and Av Mendizábal. 2013. Softcardinality-core: Improving text overlap with distributional measures for semantic textual similarity. In *Proceedings of *SEM 2013*, volume 1, pages 194–201.
- Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60.
- Enrico Mensa, Daniele P. Radicioni, and Antonio Lieto. 2017. Merali at semeval-2017 task 2 subtask 1: a cognitively inspired approach. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 236–240, Vancouver, Canada, August. Association for Computational Linguistics.
- Enrico Mensa, Daniele P Radicioni, and Antonio Lieto. 2018. Cover: a linguistic resource combining common sense and lexicographic information. *Language Resources and Evaluation*, 52(4):921–948.
- Rada Mihalcea, Courtney Corley, Carlo Strapparava, et al. 2006. Corpus-based and knowledge-based measures of text semantic similarity. In *AAAI*, volume 6, pages 775–780.
- Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

- George A Miller. 1995. WordNet: a lexical database for English. *Communications of the ACM*, 38(11):39–41.
- Andrea Moro, Alessandro Raganato, and Roberto Navigli. 2014. Entity linking meets word sense disambiguation: a unified approach. *Transactions of the Association for Computational Linguistics*, 2:231–244.
- Roberto Navigli and Simone Paolo Ponzetto. 2010. BabelNet: Building a very large multilingual semantic network. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 216–225. Association for Computational Linguistics.
- Roberto Navigli and Simone Paolo Ponzetto. 2012. BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artif. Intell.*, 193:217–250.
- Roberto Navigli. 2009. Word sense disambiguation: A survey. *ACM Computing Surveys (CSUR)*, 41(2):10.
- Christian Paul, Achim Rettinger, Aditya Mogadala, Craig A Knoblock, and Pedro Szekely. 2016. Efficient graph-based document similarity. In *European Semantic Web Conference*, pages 334–349. Springer.
- Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.
- Jay Pujara, Hui Miao, Lise Getoor, and William Cohen. 2013. Knowledge graph identification. In *International Semantic Web Conference*, pages 542–557. Springer.
- Philip Resnik. 1995. Using information content to evaluate semantic similarity in a taxonomy. *arXiv preprint cmp-lg/9511007*.
- Ray Richardson, A Smeaton, and John Murphy. 1994. Using wordnet as a knowledge base for measuring semantic similarity between words.
- Michael Schuhmacher and Simone Paolo Ponzetto. 2014. Knowledge-based graph document modeling. In *Proceedings of the 7th ACM international conference on Web search and data mining*, pages 543–552. ACM.
- Robert Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *AAAI*, pages 4444–4451.
- Fabian M Suchanek, Gjergji Kasneci, and Gerhard Weikum. 2007. Yago: a core of semantic knowledge. In *Proceedings of the 16th international conference on World Wide Web*, pages 697–706. ACM.
- Rocco Tripodi and Marcello Pelillo. 2017. A game-theoretic approach to word sense disambiguation. *Computational Linguistics*, 43(1):31–70.
- Amos Tversky. 1977. Features of similarity. *Psychological review*, 84(4):327.
- Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: A free collaborative knowledge base. *Communications of the ACM*, 57(10).

Is “*manovra*” really “*del popolo*”? Linguistic Insights into Twitter Reactions to the Annual Italian Budget Law¹

Claudia Roberta Combei
University of Bologna
claudia.combei2@unibo.it

Abstract

English. Relying on linguistic cues obtained by means of structural topic modeling as well as descriptive lexical analyses, this study contributes to the general understanding of the Twitter users’ response to the annual Italian budget law approved at the end of December 2018. Some topics contained in the dataset of tweets are procedural or generic, but besides those, it often emerges that Twitter users expressed their concern with respect to the provisions of this law. Supportive attitudes seem to be less frequent. This paper also advocates that findings from inductive studies on Twitter data should be interpreted with caution, since the nature of tweets might not be adequate for drawing far-reaching generalisations.

1 Introduction

In the last decade, Internet has revolutionized human communication and interaction. And among all forms of digitally-mediated communication, social media stand out as one of the most effective. As Boulianne (2017) points out, the effects of social media depend on their nature of use (e.g. source of information; one-to-one/one-to-many/many-to-many communication; networking and relationship-building; expression of opinions; etc.).

Nowadays, potentially everyone with a computer or a mobile device having access to the internet can write and share contents which may be viewed and debated immediately by other people.

The impact of a social media post may be huge, and unlike other prior forms of communication, it can easily cross borders in just a few seconds. In fact, social media make things happen faster than ever before. For instance, Facebook and Twitter were crucial in allowing the Arab uprisings or the Romanian anti-corruption protests to happen more efficiently and on a larger scale.

2 Tweets and politics

Besides their essential role in information dissemination, networking, and people mobilization, social media are also important indicators and predictors of their users’ opinions, sentiments and attitudes. In fact, various studies have explored people’s reactions towards social, economic, and political issues, by analysing social media posts (e.g. Burnap et al., 2014; Gaspar et al., 2016; Nesi et al., 2018), especially tweets, since they are easily retrievable by means of APIs.

With over 6,000 tweets posted every second, corresponding to roughly 350,000 per minute, 500 million per day, and around 200 billion per year, Twitter has become one of the main tools of communication worldwide (Internet Live Stats, 2019). The number of tweets written daily seems to be correlated to things happening in the real world, and, as a matter of fact, it was shown that important events generate high number of tweets (cf. Hughes and Palen, 2009), something that is generally reflected also on the Twitter “trends”. Based on Hootsuite’s (2019) report, each month, in Italy there are almost 2.5 million active users² of Twitter, a datum that confirms the popularity of this network among various layers of Italian audience.

¹ Copyright © 2019 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

² Users that write or share at least one tweet every month are defined “active”.

This means that Twitter may represent an easily exploitable opportunity for politicians in their attempt to reinforce communication with potential voters in what might be defined as a permanent digitally-mediated electoral campaign. Additionally, it has been suggested that Twitter could be used to model and predict public opinion and behaviour regarding political events, such as electoral campaigns (e.g. Coletto et al., 2015; Kalampokis et al., 2017). In fact, Ott (2017: 59) claims that Twitter may be the ideal tool for the afore-mentioned purposes since, it “*privileges discourse that is simple, impulsive, and uncivil.*”

While indeed tweets have been widely used to analyse public opinion and political discussions in all its forms, several methodological considerations are dutiful. First of all, Twitter users do not represent an optimal sample for public opinion or voting population, especially due to their higher than average level of education and political sophistication, as well as a generally younger age (cf. Gayo-Avello, 2013; Barberá et al., 2015). As a matter of fact, we believe it is more accurate to define Twitter users as a potential share of electorate. Secondly, the language of tweets is characterised by succinctness and sometimes informality, colloquialism, irony, and susceptibility to rumour, all of which are aspects that render the results of large-scale analyses hard to interpret and generalise.

3 Aims and motivations

Acknowledging all the limitations mentioned above, this inductive exploratory study aims to contribute to the growing body of literature examining Twitter and its increasingly prominent role in online communication by studying its application in the context of political discourse. In particular, the linguistic approach presented here is providing insights into tweets regarding the discussion and the approval of the annual Italian budget law (in Italian “*legge finanziaria*” and/or “*legge di bilancio*”). This law was also often labelled as “the manoeuvre” (in Italian “*la manovra*”) and “the people’s manoeuvre” (in Italian “*la manovra del popolo*”) by its proponents – in particular *Movimento 5 Stelle* (abbreviated *M5S*) –, mainly due to some of its populist provisions (e.g. the citizen's basic income and pension).

By means of structural topic modelling (cf. Roberts et al., 2014) and descriptive analyses (i.e. terminology extraction of multi-keywords and word sketches), we are interested in grasping the Twitter users’ attitudes towards the budget law in a significant moment for the first populist Government in the eurozone, namely the coalition formed by *Lega* and *M5S*.

This topic is worth studying since the two parties displayed differences in economic, fiscal, infrastructural, and social policies both in the electoral campaign for the 2018 general elections as well as during the first months of government. For instance, *Lega* supported the flat taxation on incomes, while *M5S* the citizen's basic income (“*reddito di cittadinanza*” in Italian). However, these measures, although slightly modified, as well as the amendment to the 2011 pensions reform (“*quota 100*” in Italian) were included in the coalition agreement and subsequently in the draft for the annual budget law. The bill also contained various other economic and fiscal provisions (e.g. taxes on digital services; new VAT rates; reducing military expenses and the Italian contribution to United Nations; new labour measures; environmental incentives; etc.)³.

We believe that the textual material contained in tweets may be promising in providing hints on how Twitter users – a fraction of the Italian voters – reacted to the provisions of the budget law. Linguistic insights into tweets might be able to guide us in understanding whether the so-called “*manovra del popolo*” was perceived by Twitter user as representing indeed the people’s interest.

4 Data

Although in the Western world there are three mainstream social media networks (i.e. Facebook, Instagram, and Twitter), in this paper we analyse Twitter posts, primarily as a consequence of data availability. Indeed, unlike other tools for social media, Twitter APIs for R (R Core Team, 2018) allow scholars to collect large quantities of tweets and their related metadata in a rather effortless way.

Using the *rtweet* package (Kearney, 2019) for R and Twitter’s developer account, we collected a dataset of 167,259 Twitter posts, for a total of 6.5 million tokens, consisting in tweets and retweets

³ The full text of the annual Italian budget law (*Legge 30 dicembre 2018, n. 145 – Bilancio di previsione dello Stato per l'anno finanziario 2019 e bilancio pluriennale per il triennio 2019-2021*) was published on the Official Gazette of the Italian Republic (GU n.302 31-

12-2018 - Suppl. Ordinario n. 62) and it is available online at this webpage: https://www.gazzettaufficiale.it/atto/stampa/serie_generale/originario (accessed on the 1st of June 2019).

related to the Italian budget law. Moreover, we extracted 88 metadata describing the tweet (i.e. character length, device used, number of retweets, etc.) and the user (i.e. username, location, gender, etc.). In order to capture the most important phases of the Twitter discussion about the annual budget law and considering the one-week rate limit for tweets extraction imposed by the Standard Search API⁴, the data were collected weekly from the 27th of November 2018 through the 8th of January 2019, for a total of 43 consecutive days. The hashtags used as keywords in the queries represented all the names given to the budget bill by Italian political actors, the press, and the public opinion: “#leggedibilancio”, “#leggefinanziaria”, “#manovra”, “#manovradibilancio”, “#manovraeconomica”, “#manovradelpopolo”, and “#manovrafinanziaria”. This guaranteed a large coverage of Twitter users and tweet typologies. Some of the afore-mentioned hashtags (e.g. “#manovra”, “#manovradelpopolo”) were also trending at the end of December.

To avoid duplicates, we discarded all retweets and all posts that contained quotes of other tweets. The removal process was obtained by filtering the dataset, thus selecting only tweets whose values for “is_retweet” and “is_quote” corresponded to “FALSE”. Duplicates other than retweets and quotes were removed with R’s base functions *duplicated* – which identified duplicated tweets – and *unique* – which extracted unique tweets. Since the aim of this study is to uncover the reactions of the Italian voters active on Twitter, we removed the tweets written by political actors. To do so, we defined a list containing the Twitter usernames of the members of the Italian Parliament, as well as those of the official national and local party profiles; this list was used to automatically filter and remove tweets published by the unwanted profiles. We decided to keep tweets from news agencies, online newspapers, and television channels, since they could represent vectors of information exchange regarding the topic analysed in this study. The final dataset contained 20,891 tweets.

Tokens	701,986
Words	414,803
Types	75,485
Lemmas	31,947

Table 1: Dataset statistics.

4.1 Pre-processing

Since the tweets and their metadata would have been used for lexical analyses and structural topic modelling⁵, we performed several pre-processing steps: defining a “stop words” list for Italian consisting of roughly 1,000 lexically empty or uninformative words (i.e. prepositions, conjunctions, auxiliary verbs, etc.); uniformizing, normalising and cleaning the texts with various corpus processing functions available on the R packages *quanteda* (Benoit et al., 2018), *tm* (Feinerer, Hornik, and Meyer, 2008), and *qdapRegex* (Rinker, 2017). Hashtags at the beginning and inside the tweet sentences were kept and decomposed into words (i.e. from “#trasportipubblici” to “trasporti pubblici”), while those after the final point were removed, since most of the times they represented one of the keywords used for extracting tweets. Numbers, punctuation, sequences made up of a single character, and excessive white spaces were removed as well. In order to further use temporal metadata as a covariate for the topical prevalence, the “created_at” metadatum was divided it into date and hour.

5 Analyses and results

As a result of the ever-growing interest and availability of text data – often unstructured –, various statistical and machine-assisted approaches for the analysis of textual material have been proposed. In this paper we are employing the Structural Topic Model (STM) – a generative model of word counts – (cf. Roberts et al., 2014) in R to discover topics from tweets on the annual Italian budget bill and to estimate their relationship to temporal metadata.

Similarly to Latent Dirichlet Allocation (cf. Blei, Ng, and Jordan, 2003) and Correlated Topic Model (cf. Blei and Lafferty, 2007), in the STM approach, a topic represents a mixture over words where each word has a probability of pertaining to a topic, whilst a document is a mixture over topics, therefore a specific document can consist of various topics. The sum of the topic proportions across topics for a specific document as well as the sum of word probabilities for a given topic both equal to 1. The main innovation of STM is the possibility to model topical prevalence and topical content⁶ as a function of metadata. Here we are

⁴ A description of the Standard Search API for Twitter is available at this webpage: <https://developer.twitter.com/en/docs/tweets/search/api-reference/get-search-tweets.html> (accessed on the 1st of June 2019).

⁵ Considering the scope of this paper and the analyses proposed, emoticons and emojis were left out.

⁶ The topical prevalence shows the frequency with which a specific topic is discussed, while the topical

using the date covariate to explain topical prevalence over time.

5.1 Topics

After having employed the STM’s *searchK* function to perform several tests, such as held-out likelihood and residual analysis, the ideal number of topics seemed to be between 10 and 14. Additionally, STM gave the possibility to set the type of initialization, so here the spectral one was chosen, since previous studies had proven its stability and consistence (cf. Roberts, Stewart, and Tingley, 2016). All results presented in this paragraph are based on a *K* of 10. The date of the tweet was used as a prevalence covariate; as a word profile we opted for the highest probability. We did not use the stemming function on STM since it did not perform well on Italian.

Figure 1 in Appendix shows the topics related to the annual Italian budget law as they emerged from the analysis of tweets. Each topic was further classified into one category (i.e. EU & Confidence, Main Measures, Criticism & Concern, Government vs. Opposition, Procedures – Generic, Support). This classification was based on the correlations obtained from a hierarchical clustering representation performed with the *plot* function of the *stmCorrViz* package (Coppola et al., 2016), on the review of the most characterising words, and on the examination of the most exemplar documents, namely the tweets that had the highest proportion of words associated with the topic.

Although we do not claim to model public opinion from tweets, interestingly, the topics managed to echo various issues regarding the budget law. Judging by the expected topic proportions, one could order the most prevalent topics as follows: Topics 9, 8, and 3 (sum of topic proportions: 0.29) reflect disapproval and doubts towards the provisions of the budget law; Topics 1 and 7 (sum of topic proportions: 0.22) describe the difficult negotiation with the European Union (EU) and the threat of an infringement procedure; Topics 10 and 2 (sum of topic proportions: 0.19) depict the main measures contained in the budget bill; Topic 6 (topic proportion: 0.13) illustrates the support to the budget bill and to the Government; Topic 5 (topic proportion: 0.11) refers to the procedures regarding the discussion, the vote, and the approval of the budget law; and Topic 4 (topic proportion: 0.06) reveals the conflict between the

Government and the oppositions on the provisions of the law.

After having calculated the estimated effects of the temporal covariate on topical prevalence, a plot displaying this variation was created. Figure 2 in Appendix shows how the afore-mentioned topics varied over the 43 days considered. Topics are ordered as a function of their expected proportions.

Firstly, there emerged that the variation was not particularly strong, except for some topics. For instance, Topic 9 had a peak at the end of December/the beginning of January, suggesting that Twitter users might have written tweets of concern soon after the approval of the annual Italian budget law. On the other hand, Topic 6, which contained mostly tweets of support towards the measures of the budget bill seemed to be prevalent primarily at the end of November and in mid-December. The procedural topic was generally prevalent at the end of December, a timeframe corresponding to the vote and approval of the law. The two topics summarising the negotiations with the EU, the confidence, and the possible infringement procedure were pervasive during the entire period considered, with some peaks in early- and mid-December. Topic 4 that regarded the disagreement between the Government and the opposition was constant over time, and so were the topics delineating the main measures of the law.

5.2 Descriptive lexical analyses

We were also interested in performing descriptive lexical analyses on tweets. First of all, with the terminology extraction tool on Sketch Engine (Kilgariff et al., 2014) we obtained multi-keywords – able to convey more insights than single words on the issues examined – that appear more frequently in our dataset than in the reference corpus (i.e. Italian Web 2016 – itTenTen16, cf. Jakubíček et al., 2013, for TenTen corpora). If we exclude the hashtags used as keywords for tweets extraction, these are the 30 most representative syntagmas in our dataset:

Syntagma	Translation into English
<i>reddito di cittadinanza</i>	the citizen’s basic income
<i>procedura di infrazione</i>	infringement procedure
<i>clausole di salvaguardia</i>	safeguard clauses

content represents the words used to discuss about that topic (cf. Roberts et al., 2014: 1068).

<i>voto di fiducia</i>	confidence vote
<i>blocco assunzioni</i>	hiring freeze
<i>professioni sanitarie senza titolo</i>	health professions without a degree
<i>flat tax</i>	flat tax
<i>commissione bilancio</i>	budget committee
<i>gilet azzurri</i>	blue vests
<i>taglio pensioni</i>	pension cuts
<i>scatoletta di tonno</i>	tuna can
<i>governi precedenti</i>	previous governments
<i>pensioni minime</i>	minimum pensions
<i>scatola chiusa</i>	black box
<i>nuove tasse</i>	new taxes
<i>promesse elettorali</i>	campaign promises
<i>fasce deboli</i>	vulnerable citizens
<i>deficit strutturale</i>	structural deficit
<i>accordo tecnico</i>	technical arrangement
<i>braccio di ferro</i>	trial of strength
<i>appalti senza gara</i>	no-bid contracts
<i>assurdità totale</i>	total nonsense
<i>terrorismo mediatico</i>	media terrorism
<i>auto inquinanti</i>	polluting cars
<i>più tasse</i>	more taxes
<i>governo sovranista</i>	sovereignist government
<i>manovra contro il popolo</i>	manoeuvre against the people
<i>false promesse</i>	false promises
<i>IVA sui tartufi</i>	VAT for truffles
<i>popolo italiano</i>	Italian people

Table 2: The most representative syntagmas in the dataset.

It is clear that various multi-word expressions referred to procedural aspects, such as those reflecting the vote and the approval of the budget law (e.g. “confidence vote”), while others were used to list its measures, especially fiscal and economic policies (e.g. “the citizen’s basic income”, “flat tax”, etc.). Nevertheless, various syntagmas seemed to express doubts with respect to the provisions of this law. In fact, often, the words chosen by many Twitter users to express their criticism were rather strong (e.g. “total nonsense”, “black box”, “sovereignist government”, etc.).

These concerns and rather negative reactions to the budget bill were reflected also in the word sketches (i.e. visual representations of collocations and word combinations obtained on Sketch Engine) for the words “manovra” and “legge”.

Generally, three different scenarios are distinguishable.

First of all, there were several neutral verbs, nouns, and modifiers associated to the budget law, most of which regarding its procedural aspects. The most frequent (i.e. frequency ≥ 10.81 per million) are listed below:

Word/Syntagma	Translation into English
<i>scrivere</i>	write
<i>cambiare</i>	change
<i>modificare</i>	modify
<i>discutere</i>	discuss
<i>approvare</i>	approve
<i>contenere</i>	contain
<i>prevedere</i>	consist
<i>varare</i>	launch
<i>votare</i>	vote
<i>passare</i>	pass
<i>riscrivere</i>	rewrite
<i>promulgare</i>	promulgate
<i>gialloverde</i>	yellow-green
<i>economica</i>	economic
<i>finanziaria</i>	financial
<i>populista</i>	populist
<i>discussione</i>	discussion
<i>commissione</i>	commission
<i>bilancio</i>	budget

Table 3: Neutral associations.

Next, some positive evaluations of the budget law emerged. The most frequent (i.e. frequency ≥ 10.81 per million) are listed below:

Word/Syntagma	Translation into English
<i>favorire (l'innovazione)</i>	favour (innovation)
<i>grande</i>	big
<i>buona</i>	good
<i>bella</i>	beautiful
<i>significativa</i>	significant
<i>del popolo</i>	of the people
<i>del cambiamento</i>	of the change
<i>per i cittadini</i>	for the citizens
<i>per la crescita</i>	for the growth

Table 4: Positive associations.

Nonetheless, several word associations seemed to suggest negative reactions to the budget law. The most frequent (i.e. frequency ≥ 10.81 per million) are shown below:

Word/Syntagma	Translation into English
<i>recessiva</i>	recessive
<i>piena di errori</i>	full of errors

<i>dannosa</i>	dangerous
<i>cattiva</i>	bad
<i>iniqua</i>	unfair
<i>scellerata</i>	wicked
<i>sbagliata</i>	wrong
<i>snaturata</i>	wretched
<i>taroccata</i>	false
<i>vuota</i>	empty
<i>assurda</i>	absurd
<i>folle</i>	deranged
<i>truffa</i>	fraud
<i>contro il popolo</i>	against the people
<i>del popolino</i>	of the masses
<i>del cappio</i>	of the noose
<i>da lacrime</i>	tearful
<i>scontro</i>	dispute
<i>protesta</i>	protest
<i>vergogna</i>	shame
<i>bocciatura</i>	failure
<i>della povertà</i>	of the poverty
<i>dell'assistenzi-</i> <i>alismo</i>	of welfarism
<i>buio</i>	dark
<i>diminuire</i>	diminish
<i>tagliare</i>	cut

Table 5: Criticism associations.

Finally, using the *tm*'s *findAssocs* function, we calculated the associations of the lemma “*manovra*” in the term-document matrix; some of the afore-mentioned criticism words (e.g. “absurd”, “recessive”, “bad”) had a correlation higher than 0.03, suggesting a rather frequent co-occurrence.

6 Conclusions

This paper explored the Twitter users’ reactions to the annual Italian budget bill. STM outputs and descriptive lexical analyses showed that tweets concerned various aspects associated to the object of this study. Apart from talking about procedural and generic issues, users expressed their doubts and disapproval with respect to the measures of the budget law. Generally, tweets supporting this law were less frequent. The findings of this study, although preliminary, might be seen as indicators of what subsequently turned out to be a failure for the first Conte government. Still, as reiterated throughout the paper, the results might not reflect the real attitudes of the Italian voting population, since Twitter users tend to be younger and to have an above the average level of education and political sophistication (cf. Barberá et al., 2015). Moreover, tweets, by nature, might not be suitable

for drawing steady generalizations, even if the prospects they offer for content and discourse analysis are indeed significant. Further research on this topic might include the investigation of Twitter user’s reactions by means of sentiment analysis.

References

- Pablo Barberá, John T. Jost, Jonathan Nagler, Joshua A. Tucker, and Richard Bonneau. 2015. Tweeting from Left to Right. *Psychological Science*, 26(10):1531–1542.
- Kenneth Benoit, Kohei Watanabe, Haiyan Wang, Paul Nulty, Adam Obeng, Stefan Müller, and Akitaka Matsuo. 2018. quanteda: An R package for the quantitative analysis of textual data. *Journal of Open Source Software*, 3(30):774.
- David M. Blei and John D. Lafferty. 2007. A correlated topic model of Science. *The Annals of Applied Statistics*, 1(1):17–35.
- David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3(3):993–1022.
- Shelley Boulianne. 2017. Revolution in the making? Social media effects across the globe. *Information, Communication & Society*, 22(1):39–54.
- Pete Burnap, Matthew L. Williams, Luke Sloan, Omer Rana, William Housley, Adam Edwards, Vincent Knight, Rob Procter, and Alex Voss. 2014. Tweeting the terror: modelling the social media reaction to the Woolwich terrorist attack. *Social Network Analysis and Mining*, 4(206):1–14.
- Mauro Coletto, Claudio Lucchese, Salvatore Orlando, and Raffaele Perego. 2015. Italian Information Retrieval Workshop - IIR 2015. In *Proceedings of the 6th Italian Information Retrieval Workshop*, Cagliari. CEUR Workshop Proceedings.
- Antonio Coppola, Margaret E. Roberts, Brandon M. Stewart, and Dustin Tingley. 2016. *stmCorrViz: A Tool for Structural Topic Model Visualizations*. R package version 1.3. Retrieved from <https://cran.r-project.org/web/packages/stmCorrViz/index.html/> (accessed on the 1st of June 2019).
- Ingo Feinerer, Kurt Hornik, and David Meyer. 2008. Text Mining Infrastructure in R. *Journal of Statistical Software*, 25(5):1–54.
- Rui Gaspar, Cláudia Pedro, Panos Panagiotopoulos, and Beate Seibt. 2016. Beyond positive or negative: Qualitative sentiment analysis of social media reactions to unexpected stressful events. *Computers in Human Behavior*, 56:179–191.
- Daniel Gayo-Avello. 2013. A Meta-Analysis of State-of-the-Art Electoral Prediction from Twitter

- Data. *Social Science Computer Review*, 31(6):649–679.
- Hootsuite Media Inc. 2019. Digital in 2019. Retrieved from <https://hootsuite.com/it/risorse/digital-in-2019-italy> (accessed on the 1st of June 2019).
- Internet Live Stats. 2019. Twitter Usage Statistics. Retrieved from <https://www.internetlivestats.com/twitter-statistics/> (accessed on the 1st of June 2019).
- Miloš Jakubiček, Adam Kilgarriř, Vojtěch Kovář, Pavel Rychly, and Vít Suchomel. 2013. The TenTen Corpus Family. In *7th International Corpus Linguistics Conference CL 2013*. Lancaster, 125–127.
- Evangelos Kalampokis, Areti Karamanou, Efthimios Tambouris, and Konstantinos Tarabanis. 2017. On Predicting Election Results using Twitter and Linked Open Data: The Case of the UK 2010 Election. *Journal of Universal Computer Science*, 23(3):280–303.
- Adam Kilgarriř, Vít Baisa, Jan Buřta, Miloš Jakubiček, Vojtěch Kovář, Jan Michelfeit, Pavel Rychlý, and Vít Suchomel. 2014. The Sketch Engine: ten years on. *Lexicography*, 1(1):7–36.
- Michael W. Kearney. 2019. *rtweet: Collecting Twitter Data*. R package version 0.6.9 Retrieved from <https://cran.r-project.org/package=rtweet> (accessed on the 1st of June 2019).
- Amanda Lee Hughes and Leysia Palen. 2009. Twitter adoption and use in mass convergence and emergency events. *International Journal of Emergency Management*, 6(3/4):248–260.
- Paolo Nesi, Gianni Pantaleo, Irene Paoli, and Imad Zaza. 2018. Assessing the reTweet proneness of tweets: predictive models for retweeting. *Multimedia Tools and Applications*, 77(20):26371–26396.
- Brian L. Ott. 2017. The age of Twitter: Donald J. Trump and the politics of debasement. *Critical Studies in Media Communication*, 34(1):59–68.
- R Core Team. 2018. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. Retrieved from <http://www.R-project.org/> (accessed on the 1st of June 2019).
- Tyler W. Rinker. 2017. *qdapRegex: Regular Expression Removal, Extraction, and Replacement Tools*. R package version 0.7.2. University at Buffalo. Buffalo, New York. Retrieved from <http://github.com/trinker/qdapRegex/> (accessed on the 1st of June 2019).
- Margaret E. Roberts, Brandon M. Stewart, Dustin Tingley, Christopher Lucas, Jetson Leder-Luis, Shana Kushner Gadarian, Bethany Albertson, and David G. Rand. 2014. Structural Topic Models for Open-Ended Survey Responses. *American Journal of Political Science*, 58(4):1064–1082.
- Margaret E. Roberts, Brandon M. Stewart, and Dustin Tingley. 2016. Navigating the Local Modes of Big Data: The Case of Topic Models. In R. Michael Alvarez (editor), *Computational Social Science: Discovery and Prediction (Analytical Methods for Social Research)*, 51–97. Cambridge University Press., Cambridge.

Appendix

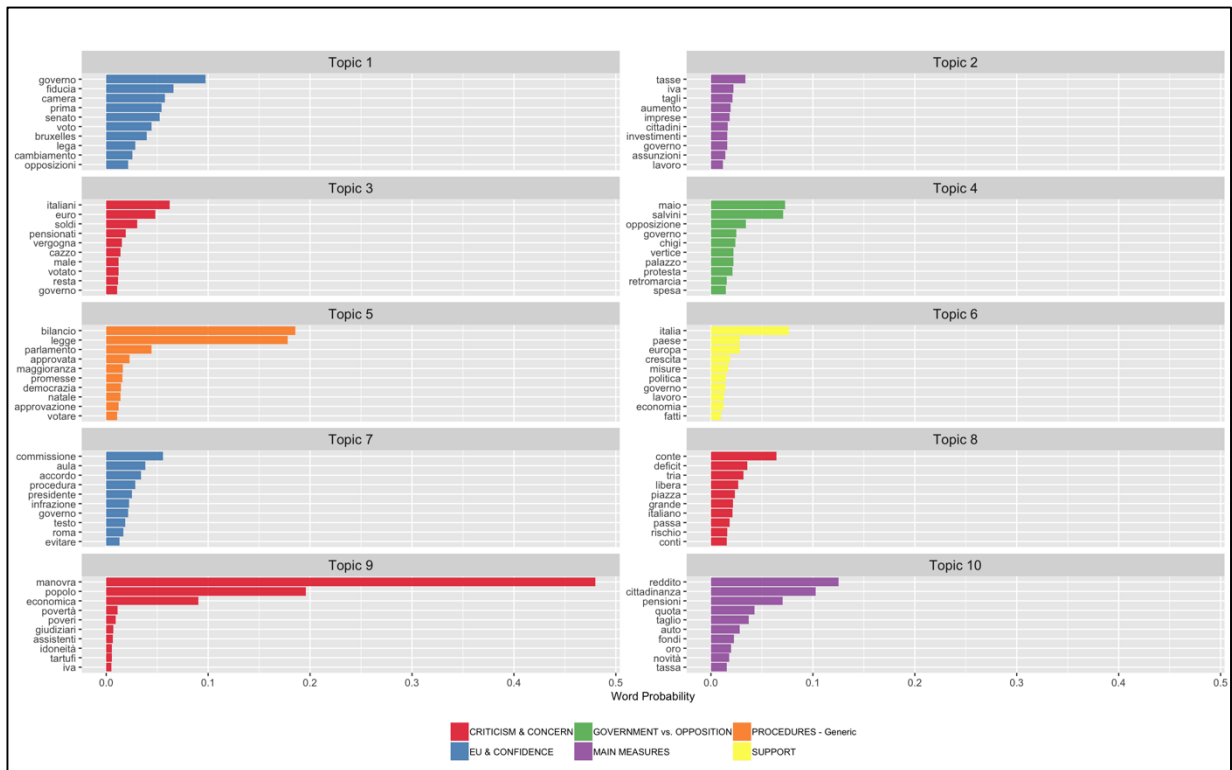


Figure 1: Topics and word probabilities.

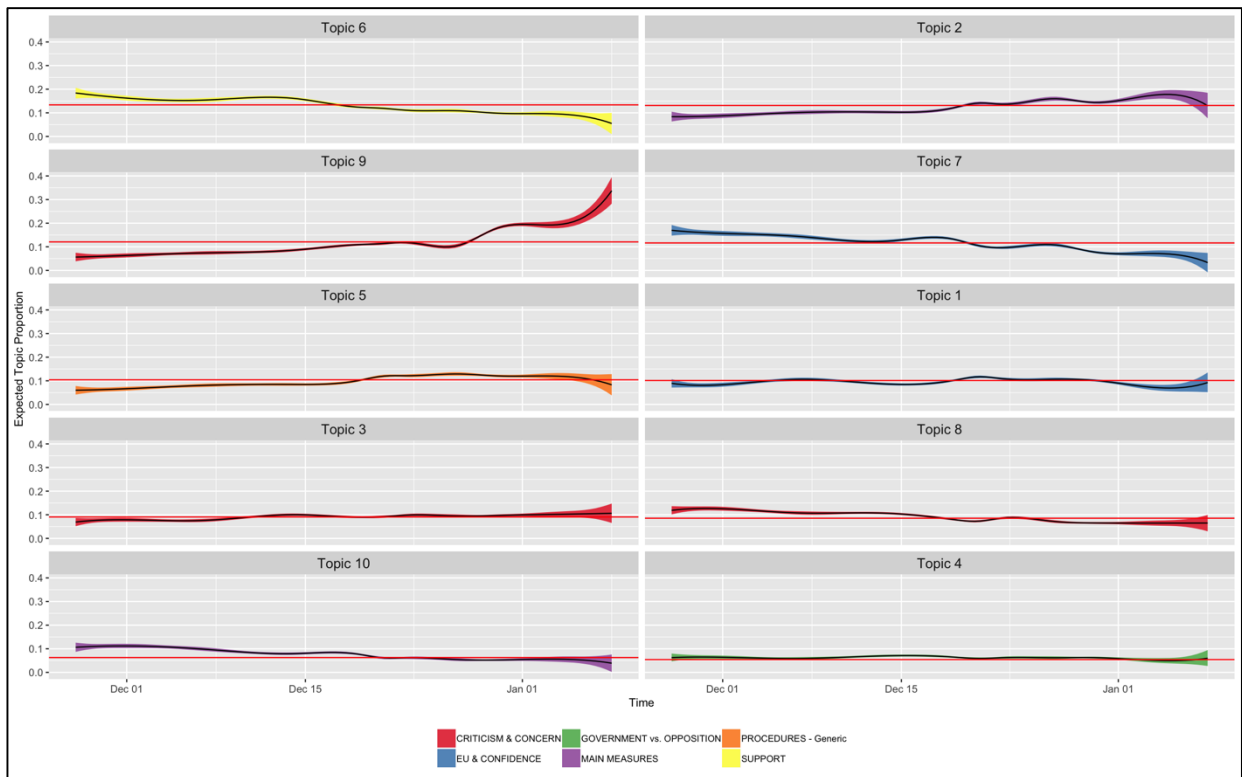


Figure 2: Variation of topic proportions over time.

Cross-Platform Evaluation for Italian Hate Speech Detection

Michele Corazza[†], Stefano Menini[‡],
Elena Cabrio[†], Sara Tonelli[‡], Serena Villata[†]

[†]Université Côte d'Azur, CNRS, Inria, I3S, France

[‡]Fondazione Bruno Kessler, Trento, Italy

michele.corazza@inria.fr

{menini, satonelli}@fbk.eu

{elena.cabrio, serena.villata}@unice.fr

Abstract

English. Despite the number of approaches recently proposed in NLP for detecting abusive language on social networks, the issue of developing hate speech detection systems that are robust across different platforms is still an unsolved problem. In this paper we perform a comparative evaluation on datasets for hate speech detection in Italian, extracted from four different social media platforms, i.e. Facebook, Twitter, Instagram and WhatsApp. We show that combining such platform-dependent datasets to take advantage of training data developed for other platforms is beneficial, although their impact varies depending on the social network under consideration.¹

Italiano. *Nonostante si osservi un crescente interesse per approcci che identifichino il linguaggio offensivo sui social network attraverso l’NLP, la necessità di sviluppare sistemi che mantengano una buona performance anche su piattaforme diverse è ancora un tema di ricerca aperto. In questo contributo presentiamo una valutazione comparativa su dataset per l’identificazione di linguaggio d’odio provenienti da quattro diverse piattaforme: Facebook, Twitter, Instagram and WhatsApp. Lo studio dimostra che, combinando dataset diversi per aumentare i dati di training, migliora le performance di classificazione, anche se l’impatto varia a seconda della piattaforma considerata.*

1 Introduction

Given the well-acknowledged rise in the presence of toxic and abusive speech on social media platforms like Twitter and Facebook, there have been several efforts within the Natural Language Processing community to deal with such problem, since the computational analysis of language can be used to quickly identify offenses and ease the removal of abusive messages. Several workshops (Waseem et al., 2017; Fišer et al., 2018) and evaluation campaigns (Fersini et al., 2018; Bosco et al., 2018; Wiegand et al., 2018) have been recently organized to discuss existing approaches to hate speech detection, propose shared tasks and foster the development of benchmarks for system evaluation.

However, most of the available datasets and approaches for hate speech detection proposed so far concern the English language, and even more frequently they target a single social media platform (mainly Twitter). In low-resource scenarios it is therefore common to have smaller datasets for specific platforms, raising research questions such as: would it be advisable to combine such platform-dependent datasets to take advantage of training data developed for other platforms? Should such data just be added to the training set or they should be selected in some way? And what happens if training data are available only for one platform and not for the other?

In this paper we address all the above questions focusing on hate speech detection for Italian. After identifying a modular neural architecture that is rather stable and well-performing across different languages and platforms (Corazza et al., to appear), we perform our comparative evaluation on freely available datasets for hate speech detection in Italian, extracted from four different social media platform, i.e. Facebook, Twitter, Instagram and Whatsapp. In particular, we

¹Copyright ©2019 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

test the same model while altering only some features and pre-processing aspects. Besides, we use a multi-platform training set but test on data taken from the single platforms. We show that the proposed solution of combining platform-dependent datasets in the training phase is beneficial for all platforms but Twitter, for which results obtained by training on tweets only outperform those obtained with a training on the mixed dataset.

2 Related work

In 2018, the first *Hate Speech Detection* (HaSpeeDe) task for Italian (Bosco et al., 2018) has been organized at EVALITA-2018², the evaluation campaign for NLP and speech processing tools for Italian. The task consists in automatically annotating messages from Twitter and Facebook, with a boolean value indicating the presence (or not) of hate speech. Two cross-platform tasks (Cross-HaSpeeDe) were also proposed, where the training was done on platform-specific data (Facebook or Twitter) and the test on data from another platform (Twitter or Facebook). In general, as expected, results obtained for Cross-HaSpeeDe were lower compared to those obtained for the in-domain tasks, due to the heterogeneous nature of the datasets provided for the task, both in terms of class distribution and data composition. Indeed, not only are Facebook posts in the task dataset longer, but they are also on average more likely to contain hate speech (68% hate posts in the Facebook test set vs. 32% in the Twitter one). This led to a performance drop, with the best system scoring 0.8288 F1 on in-domain Facebook data, and 0.6068 when the same model is tested on Twitter data (Cimino et al., 2018).

The best performing systems on the cross-tasks were ItaNLP (Cimino et al., 2018) when training on Twitter data and testing on Facebook, and Inria-FBK (Corazza et al., 2018) in the other configuration. The former adopts a newly-introduced approach based on a 2-layer BiLSTM which exploits multi-task learning with additional data from the 2016 SENTIPOLC task³. The latter, instead, uses a simple recurrent model with one hidden layer of size 500, a GRU of size 200 and no dropout.

The Cross-HaSpeeDe tasks and the analysis of system performance in a cross-platform scenario

are the starting point of this study. The task summary presented in (Bosco et al., 2018) listed some remarks on the elements affecting the system robustness that led us to extend the cross-platform experiments to new platforms, including also WhatsApp and Instagram data. To our knowledge, there have not been attempts to develop Italian systems for hate speech detection on these two platforms, probably because of the lack of suitable datasets. We therefore annotate our own Instagram data for the task, while we take advantage of a recently developed dataset for cyberbullying detection to test our system on WhatsApp.

3 Data and linguistic resources

In the following, we present the datasets used to train and test our system and their annotations (Section 3.1). Then, we describe the word embeddings (Section 3.2) we have used in our experiments.

3.1 Datasets

Twitter dataset released for the HaSpeeDe (Hate Speech Detection) shared task organized at EVALITA 2018. This dataset includes a total amount of 4,000 tweets (2,704 negative and 1,296 positive instances, i.e. containing hate speech), comprising for each tweet the respective annotation, as can be seen in Example 1. The two classes considered in the annotation are “hateful post” or “not”.

1. Annotation: hateful.

altro che profughi? sono zavorre e tutti uomini (EN: other than refugees? they are ballast and all men).

Facebook dataset also released for the HaSpeeDe (Hate Speech Detection) shared task. It consists of 4,000 Facebook comments collected from 99 posts crawled from web pages (1,941 negative, and 2,059 positive instances), comprising for each comment the respective annotation, as can be seen in Example 2. The two classes considered in the annotation are “hateful post” or “not”.

2. Annotation: hateful.

Matteo serve un colpo di stato. Qua tra poco dovremo andare in giro tutti armati come in America. (EN: Matteo, we need a coup. Soon we will have to go around armed as in the U.S.).

²<http://www.evalita.it/2018>

³<http://www.di.unito.it/~tutreeb/sentipolc-evalita16/index.html>

Whatsapp dataset collected to study pre-teen cyberbullying (Sprugnoli et al., 2018). Such dataset has been collected through a WhatsApp experimentation with Italian lower secondary school students and contains 10 chats, subsequently annotated according to different dimensions as the roles of the participants (e.g. bully, victim) and the presence of cyberbullying expressions in the message, distinguished between different classes of insults, discrimination, sexual talk and aggressive statements. The annotation is carried out at token level. To create additional training instances for our model, we join subsequent sentences of the same author (to avoid cases in which the user writes one word per message) resulting in 1,640 messages (595 positive instances). We consider as positive instances of hate speech the ones in which at least one token was annotated as a cyberbullying expression, as in Example 3).

3. Annotation: Cyberbullying expression.

fai schifo, ciccione! (EN: you suck, fat guy).

Instagram dataset includes a total amount of 6,710 messages, which we randomly collected from Instagram focusing on students' profiles (6,510 negative and 200 positive instances) identified through the monitoring system described in (Menini et al., 2019). Since no Instagram datasets in Italian were available, and we wanted to include this platform to our study, we manually annotated them as "hateful post" (as in Example 4) or "not".

4. Annotation: hateful.

Sei una troglodita (EN: you are a caveman).

3.2 Word Embeddings

In our experiments we test two types of embeddings, with the goal to compare generic with social media-specific ones. In both cases, we rely on Fasttext embeddings (Bojanowski et al., 2017), since they include both word and subword information, tackling the issue of out-of-vocabulary words, which are very common in social media data:

- **Generic embeddings:** we use embedding spaces obtained directly from the Fasttext website⁴ for Italian. In particular, we use the Italian embeddings trained on Common Crawl and Wikipedia (Grave et al., 2018) with size 300. A binary Fasttext model is also available and was therefore used;

- **Domain-specific embeddings:** we trained Fasttext embeddings from a sample of Italian tweets (Basile and Nissim, 2013), with embedding size of 300. We used the binary version of the model.

4 System Description

Since our goal is to compare the effect of various features, word embeddings, pre-processing techniques on hate speech detection applied to different platforms, we use a modular neural architecture for binary classification that is able to support both word-level and message-level features. The components are chosen to support the processing of social-media specific language.

4.1 Modular neural architecture

We use a modular neural architecture (see Figure 1) in Keras (Chollet and others, 2015). The architecture that constitutes the base for all the different models uses a single feed forward hidden layer of 500 neurons, with a ReLu activation and a single output with a sigmoid activation. The loss used to train the model is binary cross-entropy. We choose this particular architecture because it showed good performance in the EVALITA shared task for cross-platform hate speech detection, as well as in other hate speech detection tasks for German and English (Corazza et al., to appear). The architecture is built to support both word-level (i.e. embeddings) and message-level features. In particular, we use a recurrent layer to learn an encoding (x_n in the Figure) derived from word embeddings, obtained as the output of the recurrent layer at the last timestep. This encoding gets then concatenated with the other selected features, obtaining a vector of message-level features.

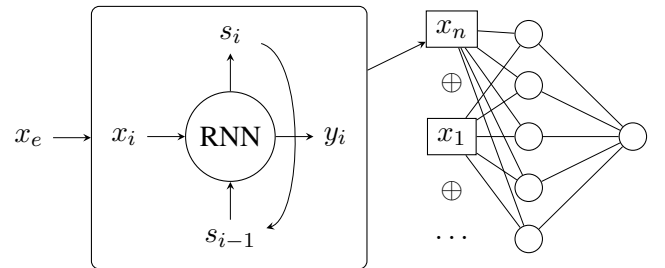


Figure 1: Modular neural architecture for Italian hate speech detection

⁴[urlhttps://fasttext.cc/docs/en/crawl-vectors.html](https://fasttext.cc/docs/en/crawl-vectors.html)

4.2 Preprocessing

The language used in social media platforms has some peculiarities with respect to standard language, as for example the presence of URLs, "@" user mentions, emojis and hashtags. We therefore run the following pre-processing steps:

- **URL and mention replacement:** both urls and mentions are replaced by the strings "URL" and "username" respectively;
- **Hashtag splitting:** Since hashtags often provide important semantic content, we wanted to test how splitting them into single words would impact on the performance of the classifier. To this end, we use the Ekphrasis tool (Baziotis et al., 2017) to do hashtag splitting and evaluate the classifier performance with and without splitting. Since the aforementioned tool only supports English, it has been adapted to Italian by using language-specific Google ngrams.⁵

4.3 Features

- **Word Embeddings:** We evaluate the contribution of word embeddings extracted from social media data, compared with the performance obtained using generic embedding spaces, as described in Section 3.2.
- **Emoji transcription:** We evaluate the impact of keeping emojis or transcribing them in plain text. To this purpose, we use the official plaintext descriptions of the emojis (from the unicode consortium website), translated to Italian with Google translate and then manually corrected, as a substitute for emojis
- **Hurtlex:** We assess the impact of using a lexicon of hurtful words (Bassignana et al., 2018), created starting from the Italian hate lexicon developed by the linguist Tullio De Mauro, organized in 17 categories. This is used to associate to the messages a score for 'hurtfulness'
- **Social media specific features:** We consider a number of metrics related to the language used in social media platforms. In particular,

we measure the number of hashtags and mentions, the number of exclamation and question marks, the number of emojis, the number of words written in uppercase

5 Experimental Setup

In order to be able to compare the results obtained while experimenting with different training datasets and features, we used fixed hyperparameters, derived from our best submission at EVALITA 2018 for the cross-platform task that involved training on Facebook data and testing on Twitter. In particular, we used a GRU (Cho et al., 2014) of size 200 as the recurrent layer and we applied no dropout to the feed-forward layer. Additionally, we used the provided test set for the two Evalita tasks, using 20% of the development set for validation. For Instagram and WhatsApp, since no standard test set is available, we split the whole dataset using 60% of it for training, while the remaining 40% is split in half and used for validation and testing. For this purpose, we use the *train_test_split* function provided by sklearn (Pedregosa et al., 2011), using 42 as seed for the random number generator.

One of our goals was to establish whether merging data from multiple social media platforms can be used to improve performance on single platform test sets. In particular, we used the following datasets for training:

- **Multi-platform:** we merge all the datasets mentioned in Section 3 for training.
- **Multi-platform filtered by length:** we use the same datasets mentioned before, but only considered instances with a length lower or equal to 280 characters, ignoring URLs and user mentions. This was done to match Twitter length restrictions.
- **Same Platform:** for each of the datasets, we trained and tested the model on data from the same platform.

In addition to the experiments performed on different datasets, we also compare the system performance obtained by using different embeddings. In particular, we train the system by using Italian Fasttext word embeddings trained on Common-Crawl and Wikipedia, and Fasttext word embeddings trained by us on a sample of Italian tweets

⁵<http://storage.googleapis.com/books/ngrams/books/datasetv2.html>

Platform	Training set	Embeddings	Features	Emoji Transcription	F1 no hate	F1 hate	Macro AVG
Instagram	Multi Platform	Twitter	Social	Yes	0.984	0.432	0.708
	Single Platform			Yes	0.981	0.424	0.702
Facebook	Multi Platform	Twitter	Social	Yes	0.773	0.871	0.822
	Single Platform			Yes	0.733	0.892	0.812
WhatsApp	Multi Platform	Twitter	Social	Yes	0.852	0.739	0.796
	Single Platform			Yes	0.814	0.694	0.754
Twitter	Single Platform	Twitter	Hurtlex	No	0.879	0.717	0.798
	Filtered Multi Platform	Twitter	Hurtlex	No	0.858	0.720	0.789
	Multi Platform	Twitter	Hurtlex	No	0.851	0.712	0.782

Table 1: Classification results

(Basile and Nissim, 2013), with an embedding size of 300. As described in Section 4.3, we also train our models including either social-media or Hurtlex features. Finally, we compare classification performance with and without emoji transcription.

6 Results

For each platform, we report in Table 1 the best performing configuration considering embedding type, features and emoji transcription. We also report the performance obtained by merging all training data (*Multi-platform*), using only platform-specific training data (*Single platform*) and filtering training instances > 280 characters (*Filtered Multi platform*) when testing on Twitter.

For Instagram, Facebook and Whatsapp, the best performing configuration is identical. They all use emoji transcription, Twitter embeddings and social-specific features. Using multi-platform training data is also helpful, and all the best performing models on the aforementioned datasets use data obtained from multiple sources. However, the only substantial improvement can be observed in the WhatsApp dataset, probably because it is the smallest one, and the classifier benefits from more training data.

The results obtained on the Twitter test set differ from the aforementioned ones in several ways. First of all, the in-domain training set is the best performing one, while the restricted length dataset is slightly better than the non restricted one. These results suggest that learning to detect hate speech on the short length interactions that happen on Twitter does not benefit from using data from other platforms. This effect can be at least partially mitigated by restricting the length of the social interactions considered and retaining only the training instances that are more similar to Twitter ones.

Another remark concerning only Twitter is that

Hurtlex is in this case more useful than social network specific features. While the precise cause for this would require more investigation, one possible explanation is the fact that Twitter is known for having a relatively lenient approach to content moderation. This would let more hurtful words slip in, increasing the effectiveness of Hurtlex as a feature, in addition to word embeddings. Additionally, emoji transcription seems to be less useful for Twitter than for other platforms. This might be explained with the fact that the Twitter dataset has relatively less emojis when compared to the others.

One final outtake confirmed by the results is the fact that embeddings trained on social media platforms (in this case Twitter) always outperform general-purpose embeddings. This shows that the language used on social platforms has peculiarities that might not be present in generic corpora, and that it is therefore advisable to use domain-specific resources.

7 Conclusions

In this paper, we examined the impact of using datasets from multiple platforms in order to classify hate speech on social media. While the results of our experiments successfully demonstrated that using data from multiple sources helps the performance of our model in most cases, the resulting improvement is not always sizeable enough to be useful. Additionally, when dealing with tweets, using data from other social platforms slightly decreases performance, even when we filter the data to contain only short sequences of text. As for future work, further experiments could be performed, by testing all possible combinations of training sources and test sets. This way, we could establish what social platforms share more traits when it comes to hate speech, allowing for better detection systems. At the moment, however, the

size of the datasets varies too broadly to allow for a fair comparison, and we would need to extend some of the datasets. Finally, another approach could be tested, where a model trained on Facebook is used for longer sequences of text, while the Twitter model is applied to the shorter ones.

Acknowledgments

Part of this work was funded by the CREEP project (<http://creep-project.eu/>), a Digital Wellbeing Activity supported by EIT Digital in 2018 and 2019. This research was also supported by the HATEMETER project (<http://hatemeter.eu/>) within the EU Rights, Equality and Citizenship Programme 2014-2020.

References

- Valerio Basile and Malvina Nissim. 2013. Sentiment analysis on italian tweets. In *Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 100–107, Atlanta.
- Elisa Bassignana, Valerio Basile, and Viviana Patti. 2018. Hurltex: A multilingual lexicon of words to hurt. In *5th Italian Conference on Computational Linguistics, CLiC-it 2018*, volume 2253, pages 1–6. CEUR-WS.
- Christos Baziotis, Nikos Pelekis, and Christos Doukheridis. 2017. DataStories at SemEval-2017 Task 4: Deep LSTM with Attention for Message-level and Topic-based Sentiment Analysis. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 747–754, Vancouver, Canada, August. Association for Computational Linguistics.
- Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.
- Cristina Bosco, Felice Dell’Orletta, Fabio Poletto, Manuela Sanguinetti, and Maurizio Tesconi. 2018. Overview of the EVALITA 2018 hate speech detection task. In *Proceedings of the Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2018) co-located with the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018)*, Turin, Italy.
- Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734. Association for Computational Linguistics.
- François Chollet et al. 2015. Keras. <https://github.com/fchollet/keras>.
- Andrea Cimino, Lorenzo De Mattei, and Felice Dell’Orletta. 2018. Multi-task learning in deep neural networks at EVALITA 2018. In *Proceedings of the Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2018) co-located with the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018)*, Turin, Italy.
- Michele Corazza, Stefano Menini, Pinar Arslan, Rachele Sprugnoli, Elena Cabrio, Sara Tonelli, and Serena Villata. 2018. Comparing different supervised approaches to hate speech detection. In *Proceedings of the Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2018) co-located with the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018)*, Turin, Italy.
- Michele Corazza, Stefano Menini, Elena Cabrio, Sara Tonelli, and Serena Villata. to appear. Robust Hate Speech Detection: A Cross-Language Evaluation. *Transactions on Internet Technology*.
- Elisabetta Fersini, Paolo Rosso, and Maria Anzovino. 2018. Overview of the task on automatic misogyny identification at ibereval 2018. In *IberEval@SEPLN*, volume 2150 of *CEUR Workshop Proceedings*, pages 214–228. CEUR-WS.org.
- Darja Fišer, Ruihong Huang, Vinodkumar Prabhakaran, Rob Voigt, Zeerak Waseem, and Jacqueline Wernimont. 2018. Proceedings of the 2nd workshop on abusive language online (alw2). In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*. Association for Computational Linguistics.
- Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. 2018. Learning word vectors for 157 languages. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*.
- Stefano Menini, Giovanni Moretti, Michele Corazza, Elena Cabrio, Sara Tonelli, and Serena Villata. 2019. A system to monitor cyberbullying based on message classification and social network analysis. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 105–110, Florence, Italy, August. Association for Computational Linguistics.
- F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

- Rachele Sprugnoli, Stefano Menini, Sara Tonelli, Filippo Oncini, and Enrico Piras. 2018. Creating a whatsapp dataset to study pre-teen cyberbullying. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 51–59. Association for Computational Linguistics.
- Zeera Waseem, Wendy Hui Kyong Chung, Dirk Hovy, and Joel Tetreault. 2017. Proceedings of the first workshop on abusive language online. In *Proceedings of the First Workshop on Abusive Language Online*. Association for Computational Linguistics.
- Michael Wiegand, Melanie Siegel, and Josef Ruppenhofer. 2018. Overview of the germeval 2018 shared task on the identification of offensive language. In *Proceedings of GermEval 2018, 14th Conference on Natural Language Processing (KONVENS 2018)*.

An Open Science System for Text Mining

Gianpaolo Coro

ISTI-CNR

via Moruzzi 1 Pisa, Italy

coro@isti.cnr.it

Giancarlo Panichi

ISTI-CNR

via Moruzzi 1 Pisa, Italy

panichi@isti.cnr.it

Pasquale Pagano

ISTI-CNR

via Moruzzi 1 Pisa, Italy

pagano@isti.cnr.it

Abstract

Text mining (TM) techniques can extract high-quality information from big data through complex system architectures. However, these techniques are usually difficult to discover, install, and combine. Further, modern approaches to Science (e.g. Open Science) introduce new requirements to guarantee reproducibility, repeatability, and re-usability of methods and results as well as their longevity and sustainability. In this paper, we present a distributed system (NLPHub) that publishes and combines several state-of-the-art text mining services for named entities, events, and keywords recognition. NLPHub makes the integrated methods compliant with Open Science requirements and manages heterogeneous access policies to the methods. In the paper, we assess the benefits and the performance of NLPHub on the I-CAB corpus¹.

1 Introduction

Today, text mining operates within the challenges introduced by big data and new Science paradigms, which impose to manage large volumes, high production rate, heterogeneous complexity, and unreliable content, while ensuring data and methods longevity through re-use in complex models and processes chains. Among the new paradigms, Open Science (OS) focusses on the implementation in computer systems of the three "R"s of the scientific method: Reproducibility, Repeatability, and Re-usability (Hey et al., 2009; EU Commission, 2016). The systems envisaged by OS, are based on Web services networks that support big data processing and the open publication

of results. Although text mining techniques exist that can tackle big data experiments (Gandomi and Haider, 2015; Amado et al., 2018), few examples that incorporate OS concepts can be found (Linthicum, 2017). For example, common text mining "cloud" services do not allow easy repeatability of the experiments by different users and are usually domain-specific and thus poorly re-usable (Bontcheva and Derczynski, 2016; Adedugbe et al., 2018). Available multi-domain systems do not use communication standards (Bontcheva and Derczynski, 2016; Wei et al., 2016), and the few OS-oriented initiatives that use text mining focus specifically on documents preservation and cataloguing (OpenMinTeD, 2019; OpenAire, 2019).

In this paper, we present a multi-domain text mining system (*NLPHub*) that is compliant with OS and combines multiple and heterogeneous processes. NLPHub is based on an e-Infrastructure (e-I), i.e. a network of hardware and software resources that allow remote users and services to collaborate while supporting data-intensive Science through cloud computing (Pollock and Williams, 2010; Andronico et al., 2011). Currently, NLPHub integrates 30 state-of-the-art text mining services and methods to recognize fragments of a text (*annotations*) associated with named abstract or physical objects (named entities), spatiotemporal events, and keywords. These integrated processes cover overall 5 languages (English, Italian, German, French, and Spanish), requested by the European projects this software is involved in (i.e. (Parthenos, 2019; SoBigData, 2019; Ariadne, 2019)). These processes come from different providers that have different access policies, and the e-I is used both to manage this heterogeneity and to possibly speed up the processing through cloud computing. NLPHub uses the Web Processing Service standard (WPS, (Schut and Whiteside, 2007)) to describe all integrated processes, and the Prov-O XML

¹Copyright © 2019 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

ontological standard (Lebo et al., 2013) to track the complete set of input, output, and parameters used for the computations (*provenance*). Overall, these features enable OS-compliance and we show that the orchestration mechanism implemented by NLPHub adds effectiveness and efficiency to the connected methods. The name "NLPHub" refers to the forthcoming extensions of this platform to other text mining methods (e.g. sentiment analysis and opinion mining), and natural language processing tasks (e.g. text-to-speech and speech processing).

2 Methods and tools

2.1 E-Infrastructure and Cloud Computing Platform

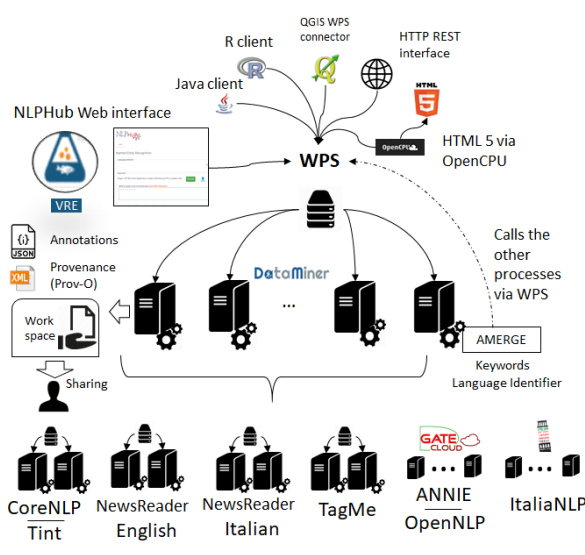


Figure 1: Overall architectural schema of the NLPHub.

NLPHub uses the open-source D4Science e-I (Candela et al., 2013; Assante et al., 2019), which currently supports applications in many domains through the integration of a distributed storage system, a cloud computing platform, online collaborative tools, and catalogues of metadata and geospatial data. D4Science supports the creation of Virtual Research Environments (VREs) (Assante et al., 2016), i.e. Web-based environments fostering collaboration and data sharing between users and managing heterogeneous data and services access policies. D4Science grants each user with access to a private online file system (the *Workspace*) that uses a high-availability distributed storage system behind the scenes, and en-

ables folders creation and sharing between VRE users. Through VREs and accounting and security services, D4Science is able to manage heterogeneous access policies by granting free access to open services in public VREs, and controlled/private access to non-open services in private or moderated VREs. D4Science includes a cloud computing platform named DataMiner (Coro et al., 2015; Coro et al., 2017) that currently hosts ~400 processes and makes all integrated processes available under the WPS standard (Figure 1). WPS is supported by third-party software and allows standardising a process' input, its parameterisation and output. DataMiner executes the processes in a cloud computing cluster of 15 machines with Ubuntu 16.04.4 LTS x86 64 operating system, 16 virtual cores, 32 GB of RAM and 100 GB of disk space. These machines are hosted by the National Research Council of Italy and the Italian Network of the University and Research (GARR). Each process can parallelise an execution either across the machines (using a Map-Reduce approach) or on the cores of one single machine (Coro et al., 2017). After each computation, DataMiner saves - on the user's Workspace- all the information about the input and output data, and the experiment's parameters (computational provenance) using the Prov-O XML standard. In each D4Science VRE, DataMiner offers an online tool to integrate algorithms, which supports many programming languages (Coro et al., 2016). All these features make D4Science useful to develop OS-compliant applications, because WPS and provenance tracking allow repeating and reproducing a computation executed by another user. Also, the possibility to provide a process in multiple VREs focussing on different domains fosters its re-usability (Coro et al., 2017). In this paper, we will use the term "algorithm" to indicate processes running on DataMiner, and "method" to indicate the original processes or services integrated with DataMiner.

2.2 Annotations

NLPHub integrates a number of named entities recognizers (NERs) but also information extraction processes that recognize events, keywords, tokens, and sentences. Overall, we will use the term "annotation" to indicate all the information that NLPHub can extract from a text. The complete list of supported annotations, languages, and processes is reported in the supple-

mentary material, together with the list of all mentioned Web services' endpoints. The ontological classes used for NERs annotations come from the Stanford CoreNLP software. Included non-standard annotations are "Misc" (miscellaneous concepts that cannot be associated with none of the other classes, e.g. "Bachelor of Science"), "Event" (nouns, verbs, or phrases referring to a phenomenon occurring at a certain time and/or space), and "Keyword" (a word or a phrase that is of great importance to understand the text content).

2.3 Integrated Text Mining Methods

NLPHub uses a common JSON format to represent the annotations of every integrated method. This format describes the input text, the NER processes, and the annotations for each NER:

```

1 "text": "input text",
2   "NER1": {
3     "annotations": {
4       "annotation1": [
5         {"indices": [i1, i2] },
6         {"indices": [i3, i4] },
7         ... ,

```

We integrated services and methods with DataMiner through "wrapping algorithms" that transformed the original outputs into this format. We implemented a general workflow in each algorithm to execute the corresponding integrated method, which adopts the following steps: (i) receive an input text file and a list of entities to recognize (among those supported by the language), (ii) pre-process the text by deleting useless characters, (iii) encode the text with UTF-8 encoding, (iv) send the text via HTTP-Post to the corresponding service or execute the method on the local machine directly, if possible, and (v) return the annotation as an NLPHub-compliant JSON document. In the following, we list all the methods currently integrated with NLPHub with reference to Figure 1 for an architectural view.

CoreNLP. The Stanford CoreNLP software (Manning et al., 2014) is an open-source text processing toolkit that supports several languages (Stanford University, 2019). NLPHub integrates CoreNLP as a service instance running within D4Science with English, German, French, and Spanish language packages enabled. Also, the Tint (The Italian NLP Tool) extension for Italian

(Aprosio and Moretti, 2016) was installed as a separate service. Overall, two distinct replicated and balanced virtual machines host these services on machines with 10 GB of RAM and 6 cores.

GATE Cloud. GATE Cloud is a cloud service that offers on-payment text analysis methods as-a-service (GATE Cloud, 2019a; Tablan et al., 2011). NLPHub integrates the GATE Cloud ANNIE NER for English, German, and French within a controlled VRE that accounts for users' requests load. This VRE ensures a fair usage of the services, whose access has been freely granted to D4Science in exchange for enabling OS-oriented features (SoBigData European Project, 2016).

OpenNLP. The Apache OpenNLP library is an open source text processing toolkit mostly based on machine learning models (Kottmann et al., 2011). An OpenNLP-based English NER is available as-a-service on GATE Cloud (GATE Cloud, 2019b) and is included among the free-to-use services granted to D4Science.

ItaliaNLP. ItaliaNLP is a free-to-use service - developed by the "Istituto di Linguistica Computazionale" (ILC-CNR) - hosting a NER method for Italian that combines rule-based and machine learning algorithms (ILC-CNR, 2019; Dell'Orletta et al., 2014).

NewsReader. NewsReader is an advanced events recognizer for 4 languages, developed by the NewsReader European project (Vossen et al., 2016). NewsReader is a formal inferencing system that identifies events by detecting their participants and time-space constraints. Two balanced virtual machines were installed in D4Science for the English and Italian NewsReader versions.

TagMe. TagMe is a service for identifying short phrases (*anchors*) in a text that can be linked to pertinent Wikipedia pages (Ferragina and Scaiella, 2010). TagMe supports 3 languages (English, Italian, and German) and D4Science already hosts its official instances. Since anchors are sequences of words having a recognized meaning within their context, NLPHub interprets them as keywords that can help contextualising and understanding the text.

Keywords NER. Keywords NER is an open-source statistical method that produces tags clouds of verbs and nouns (Coro, 2019a), which was also used by the H-Care award-winning human digital assistant (SpeechTEK 2010, 2019). Tag clouds are extracted through a statistical analysis of part-

of-speech (POS) tags (extracted with TreeTagger, (Schmid, 1995)) and the method can be applied to all the 23 TreeTagger supported languages. Keywords NER is executed directly on the DataMiner machines, and the nouns tags are interpreted as keywords for the NLP Hub scopes, because - by construction - their sequence is useful to understand the topics treated by a text.

Language Identifier. NLP Hub also provides a language identification process (Coro, 2019b), should language information not be specified as input. This process was developed in order to be fast, easily, and quickly extendible to new languages. The algorithm is based on an empirical behaviour of TreeTagger (common to many POS taggers): When TreeTagger is initialised on a certain language, but it processes a text written in another language, it tends to detect many nouns and unstemmed words than verbs and other lexical categories. Thus, the detected language is the one having the most balanced ratio of recognized and stemmed words with respect to other lexical categories. This algorithm is applicable to many languages supported by TreeTagger and can run on the DataMiner machines directly. An estimated accuracy of 95% on 100 sample text files covering the 5 NLP Hub languages was convincing to use this algorithm as an auxiliary tool for the NLP Hub users.

2.4 NLP Hub

On top of the methods and services described so far, we implemented an alignment-merging algorithm (AMERGE) that orchestrates the computations and assembles their outputs. AMERGE receives a user-provided input text, along with the indication of the text language (optionally), and a set of annotations to be extracted (selected among those supported for that language). Then, it concurrently invokes - via WPS - the text processing algorithms that support the input request, and eventually collects the JSON documents coming from them. Finally, it aligns and merges the information to produce one overall sequence represented in JSON format. The issue of merging the heterogeneous connected services' outputs is solved through the use of the DataMiner wrapping algorithms. Another solved issue is the merge of the different intervals identified by several algorithms focusing on the same entities. These intervals may either overlap or be mutually inclusive,

and the alignment algorithm manages all cases through algebraic evaluations, as reported in the following pseudo-code:

```

1 AMERGE Algorithm
2
3 For each annotation  $E$ :
4   Collect all annotations detected
      by the algorithms (intervals
      with text start and end
      positions);
5   Sort the intervals by their
      start position;
6   For each segment  $s_i$ :
7     If  $s_j$  is properly included in
       $s_i$ , process the next  $s_j$ ;
8     If  $s_i$  does not intersect  $s_j$ ,
      brake the loop;
9     If  $s_i$  intersects  $s_j$ , create a
      new segment  $su_i$  as the union
      of the two segments  $\rightarrow$ 
      substitute  $su_i$  to  $s_i$  and
      restart the loop on  $s_j$ ;
10  Save  $s_i$  in the overall list of
      merged intervals  $S$ ;
11  Associate  $S$  to  $E$ ;
12 Return all  $(E, S)$  pairs sets.

```

Since the AMERGE algorithm is a DataMiner algorithm, it is published as-a-service with a RESTful WPS interface. It represents one single access point to the services integrated with NLP Hub. In order to invoke this service, a client should specify an authorization code in the HTTP request that identifies both the invoking user and the VRE (CNR, 2016). The available annotations and methods depend on the VRE. An additional service (NLP Hub-Info) allows retrieving the list of supported entities for a VRE, given a user's authorization code. NLP Hub is also endowed with a free-to-use Web interface (nlp.d4science.org/hub/), based on a public VRE, operating on top of the AMERGE process, which allows interacting with the system and retrieving the annotations in a graphical format.

3 Results

We assessed the NLP Hub performance by using the I-CAB corpus as a reference (Magnini et al., 2006), which contains annotations of the following named-entities categories from 527 Italian newspapers: Person, Location, Organization,

Algorithm	Person				Geopolitical				Location				Organization			
	F-measure	Precision	Recall	Agreement	F-measure	Precision	Recall	Agreement	F-measure	Precision	Recall	Agreement	F-measure	Precision	Recall	Agreement
ItaliaNLP	79%	74%	84%	Excellent	77%	74%	80%	Good	59%	52%	69%	Good	58%	52%	66%	Good
CoreNLP-Tint	85%	78%	93%	Excellent	NA	NA	NA	NA	30%	18%	84%	Marginal	65%	53%	83%	Good
AMERGE	84%	74%	96%	Excellent	77%	74%	80%	Good	31%	19%	88%	Marginal	63%	49%	87%	Good
Keywords NER	20%	12%	56%	Marginal	14%	8%	66%	Marginal	6%	3%	58%	Marginal	22%	13%	66%	Marginal
TagMe	23%	18%	30%	Marginal	33%	22%	67%	Marginal	9%	5%	42%	Marginal	25%	19%	38%	Marginal
AMERGE - Keywords	20%	12%	69%	Marginal	18%	10%	91%	Marginal	6%	3%	74%	Marginal	22%	13%	79%	Marginal

Table 1: Performance assessment of the NLPHub algorithms with respect to the I-CAB corpus annotations.

Geopolitical entity. NLPHub was executed to annotate these same entities plus Keywords (Table 1). The involved algorithms were CoreNLP-Tint, ItaliaNLP, Keywords NER, and TagMe. According to the F-measure, CoreNLP-Tint was the best at recognizing Persons and Organizations, whereas ItaliaNLP - the only one supporting Geopolitical entities - had the highest performance on Locations and a moderately-high performance on Geopolitical entities. Overall, the connected methods showed high performance on specific entities, but there was not one method outperforming the others on all entities. AMERGE had lower but good F-measure and a generally high recall in all cases, which indicates that the connected algorithms include complementary and valuable intervals. The AMERGE-Keywords algorithm had a generally high recall (especially on Geopolitical entities), which means that the extracted keywords include also words from the annotated entities. The associated F-measures indicate that there is overlap with several entities. In turn, this indicates that AMERGE-Keywords could be a valuable source of information in the case of uncertainty about the entities that can be extracted from a text. As a further evaluation, we used Cohen's Kappa (Cohen, 1960) to explore the agreement between the algorithms and the I-CAB annotations. This measure required estimating the overall number of classifiable tokens, thus it is more realistic to refer to Fleiss' Kappa macro classifications rather than to the exact values (Fleiss, 1971). According to Fleiss' labels, all NERs generally have good agreement with I-CAB except for Locations, which are often reported as Geopolitical entities in I-CAB. This evaluation also highlights that AMERGE has good general agreement with manual annotations, and thus can be a valid choice when there is no prior knowledge about the algorithm to use for extracting a certain entity.

4 Conclusions

We have described NLPHub, a distributed system connecting and combining 30 text processing methods for 5 languages that adds Open Science-oriented features to these methods. The advantages of using NLPHub are several, starting from the fact that it provides one single access endpoint to several methods and spares installation and configuration time. Further, it proposes the AMERGE process as a valid option when the best performing algorithm for a certain entity extraction is not known *a priori*. Also, the AMERGE-Keywords annotations can be used when the entities to extract are not known. Indeed, these features would require more investigation, especially through multiple-language experiments, in order to define their full potential and limitations. Finally, NLPHub adds to the original methods features like WPS and Web interfaces, provenance management, results sharing, and access/usage policies control, which make the methods more compliant with Open Science requirements.

The potential users of NLPHub are scholars who want to use NERs but also want to avoid software and hardware-related issues, or automatic agents that need to automatically extract and reuse knowledge from large quantities of texts. For example, NLPHub can be used in automatic ontology population and - since it also supports Events extraction - automatic narratives generation (Petasi et al., 2011; Metilli et al., 2019). Future extensions of NLPHub will involve other text mining methods (e.g. sentiment analysis, opinion mining, and morphological parsing), and additional NLP tasks like text-to-speech and speech processing as-a-service.

Supplementary Material

Supplementary material is available on D4Science at this permanent hyper-link.

References

- [Adedugbe et al.2018] Oluwasegun Adedugbe, Elhadj Benkhelifa, and Russell Campion. 2018. A cloud-driven framework for a holistic approach to semantic annotation. In *2018 Fifth International Conference on Social Networks Analysis, Management and Security (SNAMS)*, pages 128–134. IEEE.
- [Amado et al.2018] Alexandra Amado, Paulo Cortez, Paulo Rita, and Sérgio Moro. 2018. Research trends on big data in marketing: A text mining and topic modeling based literature analysis. *European Research on Management and Business Economics*, 24(1):1–7.
- [Andronico et al.2011] Giuseppe Andronico, Valeria Ardizzone, Roberto Barbera, Bruce Becker, Riccardo Bruno, Antonio Calanducci, Diego Carvalho, Leandro Ciuffo, Marco Fargetta, Emidio Giorgio, et al. 2011. e-infrastructures for e-science: a global view. *Journal of Grid Computing*, 9(2):155–184.
- [Apro시오 and Moretti2016] Alessio Palmero Apro시오 and Giovanni Moretti. 2016. Italy goes to stanford: a collection of corenlp modules for italian. *arXiv preprint arXiv:1609.06204*.
- [Ariadne2019] Ariadne. 2019. The AriadnePlus European Project. <https://ariadne-infrastructure.eu/>.
- [Assante et al.2016] Massimiliano Assante, Leonardo Candela, Donatella Castelli, Gianpaolo Coro, Lucio Lelii, and Pasquale Pagano. 2016. Virtual research environments as-a-service by gcube. *PeerJ Preprints*, 4:e2511v1.
- [Assante et al.2019] Massimiliano Assante, Leonardo Candela, Donatella Castelli, Roberto Cirillo, Gianpaolo Coro, Luca Frosini, Lucio Lelii, Francesco Mangiacrapa, Valentina Marioli, Pasquale Pagano, et al. 2019. The gcube system: Delivering virtual research environments as-a-service. *Future Generation Computer Systems*, 95:445–453.
- [Bontcheva and Derczynski2016] Kalina Bontcheva and Leon Derczynski. 2016. Extracting information from social media with gate. In *Working with Text*, pages 133–158. Elsevier.
- [Candela et al.2013] Leonardo Candela, Donatella Castelli, Gianpaolo Coro, Pasquale Pagano, and Fabio Sinibaldi. 2013. Species distribution modeling in the cloud. *Concurrency and Computation: Practice and Experience*.
- [CNR2016] CNR. 2016. gcube wps thin clients. https://wiki.gcube-system.org/gcube/How_to_Interact_with_the_DataMiner_by_client.
- [Cohen1960] Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.
- [Coro et al.2015] Gianpaolo Coro, Leonardo Candela, Pasquale Pagano, Angela Italiano, and Loredana Liccardo. 2015. Parallelizing the execution of native data mining algorithms for computational biology. *Concurrency and Computation: Practice and Experience*, 27(17):4630–4644.
- [Coro et al.2016] Gianpaolo Coro, Giancarlo Panichi, and Pasquale Pagano. 2016. A web application to publish r scripts as-a-service on a cloud computing platform. *Bollettino di Geofisica Teorica ed Applicata*, 57:51–53.
- [Coro et al.2017] Gianpaolo Coro, Giancarlo Panichi, Paolo Scarponi, and Pasquale Pagano. 2017. Cloud computing in a distributed e-infrastructure using the web processing service standard. *Concurrency and Computation: Practice and Experience*, 29(18):e4219.
- [Coro2019a] Gianpaolo Coro. 2019a. The Keywords Tag Cloud Algorithm. <https://svn.research-infrastructures.eu/public/d4science/gcube/trunk/data-analysis/LatentSemanticAnalysis/>.
- [Coro2019b] Gianpaolo Coro. 2019b. The Language Identifier Algorithm. hyper-link.
- [Dell’Orletta et al.2014] Felice Dell’Orletta, Giulia Venturi, Andrea Cimino, and Simonetta Montemagni. 2014. T2k²: a system for automatically extracting and organizing knowledge from texts. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)*.
- [EU Commission2016] EU Commission. 2016. Open science (open access). <https://ec.europa.eu/programmes/horizon2020/en/h2020-section/open-science-open-access>.
- [Ferragina and Scaiella2010] Paolo Ferragina and Ugo Scaiella. 2010. Tagme: on-the-fly annotation of short text fragments (by wikipedia entities). In *Proceedings of the 19th ACM international conference on Information and knowledge management*, pages 1625–1628. ACM.
- [Fleiss1971] Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.
- [Gandomi and Haider2015] Amir Gandomi and Murata Haider. 2015. Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management*, 35(2):137–144.
- [GATE Cloud2019a] GATE Cloud. 2019a. GATE Cloud: Text Analytics in the Cloud. <https://cloud.gate.ac.uk/>.

- [GATE Cloud2019b] GATE Cloud. 2019b. OpenNLP English Pipeline. <https://cloud.gate.ac.uk/shopfront/displayItem/opennlp-english-pipeline>.
- [Hey et al.2009] Tony Hey, Stewart Tansley, Kristin M Tolle, et al. 2009. *The fourth paradigm: data-intensive scientific discovery*, volume 1. Microsoft research Redmond, WA.
- [ILC-CNR2019] ILC-CNR. 2019. The ItaliaNLP REST Service. <http://api.italianlp.it/docs/>.
- [Kottmann et al.2011] J Kottmann, B Margulies, G Ingersoll, I Drost, J Kosin, J Baldridge, T Goetz, T Morton, W Silva, A Autayeu, et al. 2011. Apache OpenNLP. www.opennlp.apache.org.
- [Lebo et al.2013] Timothy Lebo, Satya Sahoo, Debrah McGuinness, Khalid Belhajjame, James Cheney, David Corsar, Daniel Garijo, Stian Soiland-Reyes, Stephan Zednik, and Jun Zhao. 2013. Prov-o: The prov ontology. *W3C Recommendation*, 30.
- [Linthicum2017] David S Linthicum. 2017. Cloud computing changes data integration forever: What's needed right now. *IEEE Cloud Computing*, 4(3):50–53.
- [Magnini et al.2006] Bernardo Magnini, Emanuele Pianta, Christian Girardi, Matteo Negri, Lorenza Romano, Manuela Speranza, Valentina Bartalesi, and Rachele Sprugnoli. 2006. I-cab: the italian content annotation bank. In *LREC*, pages 963–968. Citeseer.
- [Manning et al.2014] Christopher Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, and David McClosky. 2014. The stanford corenlp natural language processing toolkit. In *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations*, pages 55–60.
- [Metilli et al.2019] Daniele Metilli, Valentina Bartalesi, and Carlo Meghini. 2019. Steps towards a system to extract. In *Proceedings of the Text2Story 2019 Workshop*, page na. Springer.
- [OpenAire2019] OpenAire. 2019. European project supporting Open Access. <https://www.openaire.eu/>.
- [OpenMinTeD2019] OpenMinTeD. 2019. Open Mining INfrastructure for TExt and Data. <https://cordis.europa.eu/project/rcn/194923/factsheet/en>.
- [Parthenos2019] Parthenos. 2019. The Parthenos European Project. <http://www.parthenos-project.eu/>.
- [Petasis et al.2011] Georgios Petasis, Vangelis Karkaletsis, Georgios Paliouras, Anastasia Krithara, and Elias Zavitsanos. 2011. Ontology population and enrichment: State of the art. In *Knowledge-driven multimedia information extraction and ontology evolution*, pages 134–166. Springer-Verlag.
- [Pollock and Williams2010] Neil Pollock and Robin Williams. 2010. E-infrastructures: How do we know and understand them? strategic ethnography and the biography of artefacts. *Computer Supported Cooperative Work (CSCW)*, 19(6):521–556.
- [Schmid1995] Helmut Schmid. 1995. Treetagger - a language independent part-of-speech tagger. *Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart*, 43:28.
- [Schut and Whiteside2007] Peter Schut and A Whiteside. 2007. OpenGIS Web Processing Service. OGC project document <http://www.opengeospatial.org/standards/wps>.
- [SoBigData European Project2016] SoBigData European Project. 2016. Deliverable D2.7 - IP principles and business models. <http://project.sobigdata.eu/material>.
- [SoBigData2019] SoBigData. 2019. The SoBigData European Project. <http://sobigdata.eu/index>.
- [SpeechTEK 20102019] SpeechTEK. 2010. 2019. SpeechTEK 2010 - H-Care Avatar wins People's Choice Award. <http://web.archive.org/web/20160919100019/http://www.speechtek.com/europe2010/avatar/>.
- [Stanford University2019] Stanford University. 2019. Stanford CoreNLP - Human Languages Supported. <https://stanfordnlp.github.io/CoreNLP/>.
- [Tablan et al.2011] Valentin Tablan, Ian Roberts, Hamish Cunningham, and Kalina Bontcheva. 2011. GATE Cloud.net: Cloud Infrastructure for Large-Scale, Open-Source Text Processing. In *UK e-Science All hands Meeting*.
- [Vossen et al.2016] Piek Vossen, Rodrigo Agerri, Itziar Aldabe, Agata Cybulska, Marieke van Erp, Antske Fokkens, Egoitz Laparra, Anne-Lyse Minard, Alessio Palmero Aprosio, German Rigau, et al. 2016. Newsreader: Using knowledge resources in a cross-lingual reading machine to generate more knowledge from massive streams of news. *Knowledge-Based Systems*, 110:60–85.
- [Wei et al.2016] Chih-Hsuan Wei, Robert Leaman, and Zhiyong Lu. 2016. Beyond accuracy: creating interoperable and scalable text-mining web services. *Bioinformatics*, 32(12):1907–1910.

Detecting Irony in Shakespeare's Sonnets with SPARSAR

Rodolfo Delmonte

Department of Linguistic Studies
Ca Foscari University
Ca Bembo - Venezia
delmont@unive.it

Nicolò Busetto

Department of Linguistic Studies
Ca Foscari University
Ca Bembo - Venezia
830070@stud.unive.it

Abstract

English. In this paper we propose a novel approach to irony detection in Shakespeare's Sonnets, a well-known data set that is statistically valuable. In order to produce a meaningful experiment, we created a gold standard by collecting opinions from famous literary critics on the same data focusing on irony. In the experiment, we use SPARSAR a system for English poetry analysis and reciting by TTS. The system produces a deep linguistically based representation at phonetic, syntactic and semantic level. It has been used to detect irony with a novel approach based on phonetic processing and sentiment analysis. At first the evaluation was very disappointing, only 50% of the sonnets matched the gold standard. Eventually, taking advantage of the semantic representation produced by the system at propositional level, the logical structure of the sonnet has been highlighted by computing the discourse relations of the couplet and/or the final quatrain. In this way we managed to improve accuracy by 17% up to 66.88%¹.

Italiano. In questo articolo si propone un nuovo approccio per l'individuazione dell'ironia nei Sonetti di Shakespeare, un dataset che è statisticamente valido. Allo scopo di produrre esperimenti significativi, abbiamo creato un gold standard raccogliendo le opinioni di famosi critici letterari sullo stesso corpus, con l'ironia come tema. Nell'esperimento abbiamo usato SPARSAR un sistema per l'analisi e la

recitazione della poesia inglese con TTS. Il sistema produce una rappresentazione linguistica profonda a livello fonetico, sintattico e semantico. E' stata usata per individuare l'ironia sulla base dell'analisi fonetica e del sentiment. All'inizio la valutazione è stata molto deludente, solo il 50% di tutti i sonetti erano inclusi nel gold standard. Poi sulla base della rappresentazione semantica prodotta dal sistema a livello proposizionale, è stata messa in luce la struttura logica del sonetto calcolando le relazioni del discorso del distico e/o della quartina finale. In questo modo abbiamo ottenuto un miglioramento dell'accuracy del 17% raggiungendo il 66.88%.

1 Introduction

Shakespeare's Sonnets are a collection of 154 poems which is renowned for being full of ironic content (Weiser, 1983), (Weiser, 1987) and for its ambiguity thus sometimes reverting the overall interpretation of the sonnet. Lexical ambiguity, i.e. a word with several meanings, emanates from the way in which the author uses words that can be interpreted in more ways not only because inherently polysemous, but because sometimes the additional meaning it evokes is derived on the basis of the sound, i.e. by homophones (see "eye", "I" in sonnet 152). The sonnets are also full of metaphors which many times require contextualising the content to the historical Elizabethan life and society. Furthermore, the sonnets are full of words related to specific language domains. For instance, there are words related to the language of economy, war, nature and to the discoveries of the modern age, and each of these words may be used as a metaphor of love. Many of the sonnets are organized around a conceptual contrast,

¹Copyright © 2019 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0)

an opposition that runs parallel and then diverges, sometimes with the use of the rhetorical figure of the chiasmus. It is just this contrast that generates irony, sometimes satire, sarcasm, and even parody. Irony may be considered in turn as: what one means using language that normally signifies the opposite, typically for humorous or emphatic effect; a state of affairs or an event that seems contrary to what one expects and is amusing as a result. As to sarcasm this may be regarded the use of irony to mock or convey contempt.(Attardo, 1994) Parody is obtained by using the words or thoughts of a person but adapting them to a ridiculously inappropriate subject. There are several types of irony, though we select verbal irony which, in the strict sense, is saying the opposite of what you mean for outcome, and it depends on the extra-linguistics context. It is important to remark that in many cases, the linguistic structures on which irony is based, may require the use of nonliteral or figurative language, i.e. the use of metaphors.

In our approach we will follow the so-called incongruity presumption or incongruity-resolution presumption. Theories connected to the incongruity presumption are mostly cognitive-based and related to concepts highlighted for instance, in (Attardo, 2000). The focus of theorization under this presumption is that in humorous texts, or broadly speaking in any humorous situation, there is an opposition between two alternative dimensions. As a result, in our study of the sonnets, produced by the contents of manual classification, we have been looking for contrasting situations; while in the sentiment analysis experiment, we have been concerned with a quantitative count of polarity related items.

Computational research on sentiment analysis has been based on the use of shallow features with a binary choice to train statistical model (Carvalho et al., 2009) that, when optimized for a particular task, will produce acceptable performance. However generalizing the model has proven to be a hard task. In addition, the text addressed by recent research has been limited to tweets, which are in no way comparable to the sonnets contain a lot of nonliteral language. The other common approach used to detect irony, in the majority of the cases, is based on polarity detection(Van Hee et al., 2018). Sentiment Analysis(Kim and Hovy, 2004) and (Kao and Jurafsky, 2012) is in fact an indiscriminate labeling of texts either on a lexi-

con basis or on a supervised feature basis where in both cases, it is just a binary - ternary or graded - decision that has to be taken. This is certainly not explanatory of the phenomenon and will not help in understanding what it is that causes humorous reactions to the reading of an ironic piece of text. It certainly is of no help in deciding which phrases, clauses or just multiwords or simply words, contribute to create the ironic meaning (see (Reyes et al., 2012); (Reyes and Rosso, 2013)).

We will not comment here on the work done to produce the gold standard which has already been described in a separate paper (Busetto & Delmonte, 2019 - To appear) but see all the file in the Supplementary materials). We simply say that we considered as ironic or sarcastic all sonnets that have been so defined by at least one of the many literary critics' comments we looked into².

2 The Architecture of SPARSAR: Syntax and Semantics

SPARSAR³ (Delmonte, 2016) builds three representations of the properties and features of each poem: a Phonetic Relational View from the phonological and the phonetic content of each word; a Poetic Relational View where the main poetic devices are addressed, related to rhythm and rhyme, and the overall metrical structure; then a Semantic Relational View where the syntactic, semantic and pragmatic content of the poem is represented, at the lexical semantic level, at the anaphoric level and at the predicate-argument structure. At this level, also the sentiment or overall mood of the poem is computed on the basis of a lean lexically based sentiment analysis. The system uses a modified version of VENSES, a semantically oriented NLP pipeline (Delmonte et al., 2005). It is accompanied by a module that works at sentence level and produces a whole set of analysis both at quantitative, syntactic and semantic level. As regards syntax, the system makes available chunks and dependency structures. Then the system introduces semantics both in the version of a classifier and by isolating verbal complex in order to verify propositional properties, like presence of negation, to compute factuality from a

²We used criticism from a set of authors including (Frye, 1957) (Calimani, 2009) (Melchiori, 1971) (Eagle, 1916) (Marelli, 2015) (Schoenfeldt, 2010) (Weiser, 1987) (Serpieri, 2002) all listed in the reference section.

³the system is freely downloadable from its website <https://sparsar.wordpress.com/>

crosscheck with modality, aspectuality – that is derived from the lexica – and tense. On the other hand, the classifier has two different tasks: separating concrete from abstract nouns, identifying highly ambiguous from singleton concepts (from number of possible meanings from WordNet and other similar repositories). Eventually, the system carries out a sentiment analysis of the poem, thus contributing a three-way classification: neutral, negative, positive that can be used as a powerful tool for prosodically related purposes.

State of the art semantic systems are based on different theories and representations, but the final aim of the workshop was reaching a consensus on what constituted a reasonably complete semantic representation. Semantics in our case not only refers to predicate-argument structure, negation scope, quantified structures, anaphora resolution and other similar items. It is referred essentially to a propositional level analysis, which is the basis for discourse structure and discourse semantics contained in discourse relations. It also paves the way for a deep sentiment or affective analysis of every utterance, which alone can take into account the various contributions that may come from syntactic structures like NPs and APs, where affectively marked words may be contained. Their contribution needs to be computed in a strictly compositional manner with respect to the meaning associated to the main verb, where negation may be lexically expressed or simply lexically incorporated in the verb meaning itself. The system does low level analyses before semantic modules are activated, that is tokenization, sentence splitting, multiword creation from a large lexical database. Then chunking and syntactic constituency parsing which is done using a rule-based recursive transition network: the parser works in a cascaded recursive way to include higher syntactic structures up to sentence and complex sentence level. These structures are then passed to the first semantic mapping algorithm that looks for subcategorization frames in the lexica made available for English, including VerbNet, FrameNet, WordNet and a proprietor lexicon of some 10K entries, with most frequent verbs, adjectives and nouns, containing also a detailed classification of all grammatical or function words. This mapping is done following LFG principles (Bresnan, 1982) (Bresnan, 2001), where c-structure is mapped onto f-structure thus obeying uniqueness, completeness

and coherence. The output of this mapping is a rich dependency structure, which contains information related also to implicit arguments, i.e. subjects of infinitivals, participials and gerundives. LFG representation also has a semantic role associated to each grammatical function, which is used to identify the syntactic head lemma uniquely in the sentence. Finally it takes care of long distance dependencies for relative and interrogative clauses. When fully coherent and complete predicate argument structures have been built, pronominal binding and anaphora resolution algorithms are fired. Coreferential processes are activated at the semantic level: they include a centering algorithm for topic instantiation and memorization that we do using a three-place stack containing a Main Topic, a Secondary Topic and a Potential Topic. Main Topics are chosen as best candidates for free pronominals - as long as morphological features are matching. In order to become a Main Topic, a Potential Topic must be reiterated. Discourse Level computation is done at propositional level by building a vector of features associated to the main verb of each clause. They include information about tense, aspect, negation, adverbial modifiers, modality. These features are then filtered through a set of rules which have the task to classify a proposition as either objective/subjective, factual/nonfactual, foreground/background. In addition, every lexical predicate is evaluated with respect to a class of discourse relations. Eventually, discourse structure is built, according to criteria of clause dependency where a clause can be classified either as coordinate or subordinate. Factuality is used to set apart opinions from facts and subjectivity is also used to contribute positively to the choice of expressing ironic content.

3 The Architecture of SPARSAR: Phonetics and Poetic Devices

The second module is a rule-based system that converts graphemes of each poem into phonetic characters, it divides words into stressed/unstressed syllables and computes rhyming schemes at line and stanza level. To this end it uses grapheme to phoneme translations made available by different sources, amounting to some 500K entries, and include CMU dictionary

⁴, MRC Psycholinguistic Database ⁵, Celex Database (H. et al., 1995), plus a proprieter database made of some 20,000 entries. Out of vocabulary words are computed by means of a prosodic parser implemented in a previous project (Bacalu and Delmonte, 1999) containing a big pronunciation dictionary which covers 170,000 entries approximately. Besides the need to cover the majority of grapheme to phoneme conversions by the use of appropriate dictionaries, remaining problems to be solved are related to ambiguous homographs like “import” (verb) and “import” (noun) and are treated on the basis of their lexical category derived from previous tagging. Eventually there is always a certain number of Out Of Vocabulary (OOV) words. The simplest case is constituted by differences in spelling determined by British vs. American pronunciation. This is taken care of by a dictionary of graphemic correspondances. However, whenever the word is not found the system proceeds by morphological decomposition, splitting at first the word from its prefix and if that still does not work, its derivational suffix. As a last resource, an orthographically based version of the same dictionary is used to try and match the longest possible string in coincidence with current OOVW. Then the remaining portion of word is dealt with by guessing its morphological nature, and if that fails a grapheme-to-phoneme parser is used. Some words thus reconstructed are wayfarer, gangrened, krog, copperplate, splendor, filmy, seraphic, unstarred.

Other words we had to reconstruct are: shrive, slipstream, fossicking, unplotted, corpuscle, thither, wraiths, etc. In some cases, the problem that made the system fail was the presence of a syllable which was not available in VESD, our database of syllable durations. This problem has been coped with partly by manually inserting the missing syllable and by computing its duration from the component phonemes; but also from the closest similar syllable available in the database. We only had to add 12 new syllables for a set of approximately 1000 poems that the system computed. The system has no limitation on

type of poetic and rhetoric devices, however it is dependent on language: Italian line verse requires a certain number of beats and metric accents which are different from the ones contained in an English iambic pentameter. Rules implemented can demote or promote word-stress on a certain syllable depending on selected language, line-level syllable length and contextual information. This includes knowledge about a word being part of a dependency structure either as dependent or as head.

4 The Experiment for the Automatic Annotation of the Sonnets using SPARSAR

The experiment we devised was organized as follows: we downloaded SPARSAR from its dedicated website <https://sparsar.wordpress.com/>. At first, following (Tsur, 1992), pag.15 and (Fonagy, 1971), and on the basis of the complete Phonological description of each word in the poem (see (Delmonte, 2016)), the system creates a relation between sound and mood or attitude by means of the module for sentiment analysis. In particular, it collapses together unvoiced, obstruent consonants with high and back vowels to represent hatred and struggle, mystic obscurity, sad and aggressive mood; the opposite is represented by voiced, sonorants and continuants consonants associated to low and front vowels. These oppositions are then applied to the one created by polarity values, negative vs. positive. We use these quantities to check an existing correlation, by using ratios. Basic relations are reported already in (Delmonte, 2016), where however mood of each sonnet was manually computed. We report here relations intervening between the output of the system, comparing ratios derived from sound relations with those from polarity. As said above, polarity values are computed according to a lexicalized approach to sentiment analysis which takes into account also negation at propositional level (see (Taboada et al., 2011)) A ratio lower than 1 indicates a majority of Negative items, higher than 1 a majority of Positive items. The same would apply to the remaining ratios. We compute the mean value for the three indices – Contrasting Vowels, Contrasting Consonants, Contrasting Voicing to indicate a generic sound related mood, Positive when the mean is higher than 1 and negative when it is lower. We then compare Results for polarity from sentiment

⁴It is available online at <http://www.speech.cs.cmu.edu/cgi-bin/cmudict/>

⁵Previously, data for POS were merged in from a different dictionary (MRC Psycholinguistic Database, <http://lcb.unc.edu/software/multimrc/multimrc.zip>), which uses British English pronunciation)

analysis with those obtained from sound evaluations. We mark sonnets with a clash between the two parameters with 1 and with 0 whenever they converge to the same value. From a perusal of the results, a total of 79 sonnets over 98 have a clash, amounting to a remarkably high percentage of 80%. However when we check the system output with the critics' choice we come up with a different picture: only 77 of all sonnets match with critics opinion, i.e. exactly 50%. This is the list of those 77 sonnets that have been found to match between the critics' list and the list of the sonnets recognized by the system as having some kind of contrast:

**1 2 4 5 6 10 12 14 17 18 19 20 21 27 30 32 33
34 35 37 41 42 47 48 50 56 57 61 65 67 68 69 71
72 74 75 77 78 79 81 82 84 87 92 95 97 98 101
102 104 106 108 109 111 113 114 115 116 123
125 126 127 129 134 136 137 139 142 144 145
146 149 151 152 153 154**

4.1 Extracting Couplets from Logical Structure

Considering the low accuracy reached with the purely quantitative approach, we decided to look into the semantic output of the system. We deemed that one of the possible reasons for the relatively low accuracy of the system could be the use of quantities to generate abstract evaluations: in other words, it is not always the case that a contrast is to be found by counting number of negative vs. positive items present in the sonnet. As to semantic representation created by *SPARSAR*, we are here referring to the logical structure of the Elizabethan sonnet where the argumentation is developed into three sections and the conclusion usually comes in the final couplet. This conclusion may revert the contents of the logical order as defined by the premises. The poet may defer the conclusion in the couplet to complete the logical argumentation by adding some further motivation. But in some cases the couplet is used to provoke surprise in the reader/hearer, accompanied by laughter or by indignation whenever sarcasm is intended. So eventually the opposition may only be present in the final two lines, and be hinted at by presence of discourse markers like "Yet", "But". In that case, it will not be sufficient for the system to ascertain the required quantity for a contrast, unless some specific rule is inserted that triggers such unexpected, unpredictable ending. To

this purpose, we proceeded by extracting manually those failed - we list them in the Appendix - that the system found without (sufficient) contrast, contrary to the decision of the critics.⁶

After a careful perusal of the couplet of each such sonnet we came up with a double list. The result is that for 26 sonnets the couplet is a clear indicator of the subversion of mood, which may go from negative to positive, if the rest of the sonnet was mostly negative; or from positive to negative in the opposite case. As said above, the trigger for the reverted mood was to be found in the presence of a discourse marker at the beginning of the first (sometimes the second) line of the couplet. Appropriate discourse markers for mood reversal are adversatives, like "but", but also concessives, like "yet" and resultatives like "so, then". This only applies to 13 of the sonnets, the remaining couplets are characterized by presence of negation and negative items (while the rest of the poem has a majority of positive items). This rule was added to the system which raised accuracy on all sonnets to 66.88%. Here below the list of 26 reclassified sonnets:

**3, 7, 8, 9, 13, 22, 40, 43, 49, 53, 58, 59, 60, 70,
73, 80, 120, 130, 131, 132, 133, 138, 140, 141,
148, 150**

The remaining sonnets require the system to look at the previous and last stanza where again an appropriate discourse marker - or a negation plus negative items - must be present to introduce the reversal of mood. However, this additional modification of the system was not fully successful and was abandoned. The list of these 19 sonnets is this:

**15, 16, 25, 26, 29, 31, 36, 55, 62, 85, 86, 88, 89,
91, 93, 94, 121, 124, 143**

5 Conclusion

In this paper we have presented work carried out to annotate and experiment with the theme of irony in Shakespeare's sonnets. The gold standard for the experiment has been created by collecting comments produced by literary critics on the presence of some kind of thematic, semantic and syntactic

⁶What we found is a list of 45 sonnets: 3, 7, 8, 9, 13, 15, 16, 22, 25, 26, 29, 31, 36, 40, 43, 49, 53, 55, 58, 59, 60, 62, 70, 73, 80, 85, 86, 88, 89, 91, 93, 94, 120, 121, 124, 130, 131, 132, 133, 138, 140, 141, 143, 148, 150

opposition in the sonnets as to produce some sort or irony. We have used the system available on the web, SPARSAR, to produce an automatic evaluation based on two parameters, phonetic features collapsed according to the theory that treats certain sounds to induce a negative rather than a positive mood. The second parameter is polarity, derived from the output of the module for sentiment analysis available in the system. From a comparison between the critics' choices and the system's the result was at first rather disappointing, it stopped at 50% of all sonnets. We then produced a new and much richer experiment by considering the logical structure of the sonnet and the content of the couplet by means of sentiment analysis, discourse markers and discourse relations. This allowed us to reach a final accuracy of 68.88%.

References

- Salvatore Attardo. 1994. *Linguistic Theories of Humor*. Mouton de Gruyter, Berlin New York.
- Salvatore Attardo. 2000. Irony as relevant inappropriateness. *Journal of Pragmatics*, 84(32).
- Ciprian Bacalu and Rodolfo Delmonte. 1999. Prosodic modeling for syllable structures from the vesd - venice english syllable database. In *Aspetti Computazionale in Fonetica, Linguistica e Didattica delle Lingue: Modelli e Algoritmi - Atti 9 Convegno GFS-AIA*, pages 147–160, Venezia. GFS-AIA.
- Joan Bresnan, editor. 1982. *The Mental Representation of Grammatical Relations*. The MIT Press, Cambridge MA.
- Joan Bresnan. 2001. *Lexical-Functional Syntax*. Blackwell Publishing, Oxford.
- Dario Calimani. 2009. *William Shakespeare, I sonetti della menzogna*. Carrocci, Roma.
- P. Carvalho, L. Sarmento, M. Silva, and E. de Oliveira. 2009. Clues for detecting irony in user-generated contents: oh...!! it's so easy;-). In *Proceeding of the 1st international CIKM workshop on Topic-sentiment analysis for mass opinion*, pages 53–56. ACM.
- Rodolfo Delmonte, Sara Tonelli, Marco Aldo Piccolino Boniforti, Antonella Bristot, and Emanuele Pianta. 2005. Venses – a linguistically-based system for semantic evaluation. In *Machine Learning Challenges*, pages 344–371, Berlin. Springer.
- Rodolfo Delmonte. 2016. Exploring shakespeare's sonnets with sparsar. volume 4, pages 61–95.
- R.L. Eagle. 1916. *New light on the enigmas of Shakespeare's Sonnets*. John Long Limited, London.
- Ivan Fonagy. 1971. The functions of vocal style. In Seymour Chatman, editor, *Literary Style: A Symposium*, pages 159–174. Oxford UP, London.
- Northrop Frye. 1957. *Anatomy of Criticism: Four Essays*. Princeton University Press.
- Baayen R. H., R. Piepenbrock, and L. Gulikers. 1995. *The CELEX Lexical Database (CD-ROM)*. Linguistic Data Consortium.
- Justine Kao and Dan Jurafsky. 2012. A computational analysis of style, affect, and imagery in contemporary poetry. In *Proceedings of NAACL Workshop on Computational Linguistics for Literature*, pages 8–17, Stroudsburg, PA, USA. ACL.
- S.-M. Kim and E. Hovy. 2004. Determining the sentiment of opinions. In *Proceedings of the 20th international conference on computational linguistics - COLING*, pages 1367–1373, Stroudsburg, PA, USA. ACL.
- Maria Antonietta Marelli. 2015. *William Shakespeare, I Sonetti – con testo a fronte*. Garzanti, Milano.
- Giorgio Melchiori. 1971. *Shakespeare's Sonnets*. Adriatica Editrice, Bari.
- Antonio Reyes and Paolo Rosso. 2013. On the difficulty of automatically detecting irony: beyond a simple case of negation. *Knowledge and Information Systems*, 40:595–614.
- Antonio Reyes, Paolo Rosso, and Davide Buscaldi. 2012. From humor recognition to irony detection: The figurative language of social media. *Data & Knowledge Engineering*.
- Michael Schoenfeldt. 2010. *Cambridge introduction to Shakespeare's poetry*. Cambridge University Press, Cambridge.
- Alessandro Serpieri. 2002. *Polifonia Shakespeariana*. Bulzoni, Roma.
- M. Taboada, J. Brooke, M. Tofiloski, K. Voll, and M. Stede. 2011. Lexicon-based methods for sentiment analysis. *Computational Linguistics*, 37(2):267–307.
- Reuven Tsur. 1992. *What Makes Sound Patterns Expressive: The Poetic Mode of Speech-Perception*. Duke UP, Durham N.C.
- Cynthia Van Hee, Els Lefever, and Véronique Hoste. 2018. Semeval-2018 task 3: Irony detection in english tweets.). In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 39–50. ACL.
- David K. Weiser. 1983. Shakespearean irony: The 'sonnets'. *Neuphilologische Mitteilungen*, 84(4):456–469.
- David K. Weiser. 1987. *Mind in Character – Shakespeare's Speaker in the Sonnets*. The University of Missouri Press.

APPENDIX

List of couplets and quatrains from sonnets which contain a discourse marker for reverted logical structure

A Section 1: Couplets Reverting the Logical Sequence

Sonnet 3 But if thou live remembered not to be,
Die single and thine image dies with thee.

Sonnet 7 So thou, thyself out-going in thy noon,
Unlooked on diest unless thou get a son.

Sonnet 8 Whose speechless song, being many,
seeming one, Sings this to thee: "Thou single wilt
prove none."

Sonnet 9 No love toward others in that bosom
sits That on himself such murd'rous shame com-
mits.

Sonnet 22 Presume not on thy heart when mine
is slain; Thou gav'st me thine not to give back
again.

Sonnet 40 Lascivious grace, in whom all ill
well shows, Kill me with spites; yet we must not
be foes.

Sonnet 43 All days are nights to see till I see
thee, And nights bright days when dreams do show
thee me.

Sonnet 49 To leave poor me, thou hast the
strength of laws, Since why to love I can allege
no cause.

Sonnet 53 In all external grace you have some
part, But you like none, none you, for constant
heart.

Sonnet 58 I am to wait, though waiting so be
hell, Not blame your pleasure, be it ill or well.

Sonnet 59 O sure I am the wits of former days
To subjects worse have giv'n admiring praise.

Sonnet 60 And yet to times in hope my verse
shall stand, Praising thy worth, despite his cruel
hand.

Sonnet 70 If some suspect of ill masked not
thy show, Then thou alone kingdoms of hearts
shouldst owe.

Sonnet 73 This thou perceiv'st, which makes
thy love more strong, To love that well which thou
must leave ere long.

Sonnet 80 Then, if he thrive and I be cast away,
The worst was this: my love was my decay.

Sonnet 120 But that your trespass now becomes
a fee; Mine ransoms yours, and yours must ransom
me.

Sonnet 130 And yet, by heaven, I think my love
as rare As any she belied with false compare.

Sonnet 131 In nothing art thou black save in
thy deeds, And thence this slander, as I think, pro-
ceeds.

Sonnet 132 Then will I swear beauty herself is
black, And all they foul that thy complexion lack.

Sonnet 133 And yet thou wilt, for I being pent
in thee, Perforce am thine, and all that is in me.

Sonnet 138 Therefore I lie with her, and she
with me, And in our faults by lies we flattered be.

Sonnet 140 That I may not be so, nor thou be-
lied, Bear thine eyes straight, though thy proud
heart go wide.

Sonnet 141 Only my plague thus far I count my
gain, That she that makes me sin awards me pain.

Sonnet 148 O cunning love! With tears thou
keep'st me blind, Lest eyes well seeing thy foul
faults should find.

Sonnet 150 If thy unworthiness raised love in
me, More worthy I to be beloved of thee.

B Section 2: Couplet + (Part of) Previous Stanza

Sonnet 15 Then the conceit of this inconstant
stay Sets you, most rich in youth, before my
sight, Where wasteful time debateth with decay,
To change your day of youth to sullied night; And
all in war with time for love of you, As he takes
from you, I engraft you new.

Sonnet 16 So should the lines of life that life re-
pair Which this time's pencil or my pupil pen Nei-
ther in inward worth nor outward fair Can make
you live yourself in eyes of men. To give away
yourself keeps yourself still, And you must live,
drawn by your own sweet skill.

Sonnet 25 The painful warrior famousèd for
worth, After a thousand victories once foiled, Is
from the book of honor razèd quite, And all the
rest forgot for which he toiled. Then happy I that
love and am belovèd Where I may not remove nor
be removèd.

Sonnet 26 But that I hope some good conceit
of thine In thy soul's thought, all naked, will be-
stow it. Till whatsoever star that guides my mov-
ing Points on me graciously with fair aspect And
puts apparel on my tattered loving, To show me
worthy of thy sweet respect. Then may I dare to
boast how I do love thee; Till then, not show my
head where thou mayst prove me.

Sonnet 29 Yet in these thoughts myself almost

despising, Haply I think on thee, and then my state, Like to the lark at break of day arising From sullen earth, sings hymns at heaven's gate. For thy sweet love remembered such wealth brings That then I scorn to change my state with kings.

Sonnet 31 But things removed that hidden in thee lie. Thou art the grave where buried love doth live, Hung with the trophies of my lovers gone, Who all their parts of me to thee did give; That due of many now is thine alone. Their images I loved I view in thee, And thou, all they, hast all the all of me.

Sonnet 36 I may not evermore acknowledge thee, Lest my bewailèd guilt should do thee shame; Nor thou with public kindness honor me, Unless thou take that honor from thy name. But do not so; I love thee in such sort, As, thou being mine, mine is thy good report.

Sonnet 55 Even in the eyes of all posterity That wear this world out to the ending doom. So till the judgment that yourself arise, You live in this, and dwell in lovers' eyes.

Sonnet 62 But when my glass shows me myself indeed, Beated and chopped with tanned antiquity, Mine own self-love quite contrary I read; Self so self-loving were iniquity. 'Tis thee, myself, that for myself I praise, Painting my age with beauty of thy days.

Sonnet 85 But that is in my thought, whose love to you, Though words come hindmost, holds his rank before. Then others for the breath of words respect, Me for my dumb thoughts, speaking in effect.

Sonnet 86 As victors of my silence cannot boast. I was not sick of any fear from thence; But when your countenance filled up his line, Then lacked I matter, that enfeebled mine.

Sonnet 88 The injuries that to myself I do, Doing thee vantage, double vantage me. Such is my love, to thee I so belong, That for thy right myself will bear all wrong.

Sonnet 89 Thy sweet belovèd name no more shall dwell, Lest I, too much profane, should do it wrong And haply of our old acquaintance tell. For thee against myself I'll vow debate, For I must ne'er love him whom thou dost hate.

Sonnet 91 But these particulars are not my measure; All these I better in one general best. Thy love is better than high birth to me, Richer than wealth, prouder than garments' cost, Of more delight than hawks or horses be; And having thee,

of all men's pride I boast; Wretched in this alone, that thou mayst take All this away, and me most wretched make.

Sonnet 93 But heav'n in thy creation did decree That in thy face sweet love should ever dwell; Whate'er thy thoughts or thy heart's workings be, Thy looks should nothing thence but sweetness tell. How like Eve's apple doth thy beauty grow, If thy sweet virtue answer not thy show.

Sonnet 94 But if that flow'r with base infection meet, The basest weed outbraves his dignity. For sweetest things turn sourest by their deeds; Lilies that fester smell far worse than weeds.

Sonnet 121 Which in their wills count bad what I think good? No, I am that I am, and they that level At my abuses reckon up their own; I may be straight, though they themselves be bevel. By their rank thoughts my deeds must not be shown, Unless this general evil they maintain: All men are bad, and in their badness reign.

Sonnet 124 That it nor grows with heat nor drowns with showers. To this I witness call the fools of time, Which die for goodness, who have lived for crime.

Sonnet 143 So run'st thou after that which flies from thee, Whilst I, thy babe, chase thee afar behind. But if thou catch thy hope, turn back to me, And play the mother's part, kiss me, be kind. So will I pray that thou mayst have thy Will, If thou turn back and my loud crying still.

Towards an Italian Learner Treebank in Universal Dependencies

Elisa Di Nuovo

Dipartimento di Lingue e Letterature
Straniere e Culture Moderne
University of Turin
elisa.dinuovo@unito.it

Cristina Bosco

Alessandro Mazzei
Manuela Sanguinetti
Dipartimento di Informatica
University of Turin
{bosco,mazzei,msanguin}@di.unito.it

Abstract

In this paper we describe the preliminary work on a novel treebank which includes texts written by learners of Italian drawn from the VALICO corpus. Data processing mostly involved the application of Universal Dependencies formalism and error annotation. First, we parsed the texts on UDPipe trained on the existent Italian UD treebanks, then we manually corrected them. The particular focus of this paper is on a one-hundred-sentence sample of the collection, used as a case study to define an annotation scheme for identifying the linguistic phenomena characterizing learners' interlanguage.

1 Introduction

The increasing interest in Learner Corpora (henceforth LC) is twofold motivated. On the one hand, LC are an especially valuable source of knowledge for interlanguage varieties. They allow in-depth comparisons of non-native varieties, helping to elucidate the properties of the interlanguage developed by learners with different mother tongues and learning levels. For this reason, LC are important resources enabling data-driven studies exploited within several research areas, such as Second Language Acquisition, Foreign Language Teaching, Contrastive Interlanguage Analysis, Computer-aided Error Analysis, Computer-Assisted Language Learning and L2 Lexicography (e.g. (Pravec, 2002; Granger, 2008; McEnery and Xiao, 2011)). On the other hand, LC have raised considerable computational interest, which is closely related to their usefulness in tasks such as Native Language Identification (Jarvis

and Paquot, 2015; Malmasi, 2016), Grammatical-Error Detection and Correction (Leacock et al., 2015; Ng et al., 2014), and Automated Essay Scoring (Higgins et al., 2015).

In this paper we describe the development of a novel learner Italian treebank, i.e. VALICO-UD, in which Universal Dependencies (UD) formalism is tied to error annotation. The considerations of the annotation process, carried out on a set of one hundred sentences selected from a subcorpus of VALICO¹ (see Table 1) (Corino and Marelllo, 2017), allowed us to test a pilot scheme which pinpoints some of the features of L2 Italian.

This paper is organized as follows: in Section 2 we provide an overview of LC, focusing on Italian resources in particular; in Section 3 we present the data and the error annotation of VALICO-UD; in Section 4 we offer some examples of how we applied literal annotation to the learner sentences (LS) and, finally, in Section 6 we present conclusion and future work.

2 Related work

LC, also called interlanguage or L2 corpora, are collections of data produced by foreign or second language learners (Granger, 2008). Most LC projects were launched in the nineties and focused mainly on learner English (Tono, 2003), but recently we have witnessed an increasing interest in LC for other target languages. This has contributed to the establishment of learner corpus research (Tono, 2003).

LC can be enriched with Part of Speech (PoS) tagging, syntactic, semantic, discourse structure and error-tagging (with explicit or implicit target hypotheses²) annotation (Garside et al., 1997). To provide linguistic annotation, NLP tools are often used (Huang et al., 2018) and combined with

Copyright © 2019 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

¹<http://www.valico.org/>

²A reconstructed LS on which error identification is based (Reznicek et al., 2013).

human post-editing in order to overcome issues arising from the failures of the automatic analysis (Geertzen et al., 2013; Granger et al., 2009; Dahlmeier et al., 2013).

Among the 14 learner Italian corpora registered in the *Learner Corpora around the World* list³, the majority are in the form of plain texts, or they only annotate PoS (COLI, LOCCLI and CAIL2⁴, and VALICO), while only MERLIN (Boyd et al., 2014) annotates syntax and errors (with explicit target hypotheses).

Although MERLIN contains 816 texts written in non-native Italian (Boyd et al., 2014), they are not balanced for learners’ mother tongue and are not annotated using a standard annotation for syntax, which would allow comparisons with other resources. To fill this gap, we decided to develop VALICO-UD, a L1-balanced resource developed within the UD formalism, thus providing a greater potential for contrastive analysis. Indeed, a UD-annotated LC can be compared with other LC (therefore different interlanguages) or also with native corpora of the L1 involved. For all these reasons, we decided to develop this new learner Italian treebank within the UD formalism. References were the English and Chinese experiences, respectively the English Second Language (ESL) (Berzak et al., 2016) and the Chinese Foreign Language (CFL) (Lee et al., 2017) treebanks.

The scholars involved in the annotation of the ESL and CFL treebanks decided to follow a well-established line of work, for which learner language analysis is centered upon morpho-syntactic surface evidence. This is motivated by various studies, e.g. (Díaz-Negrillo et al., 2010; Ragheb and Dickinson, 2012), in which the difference between morphological and distributional PoS is stressed. We decided to follow this line of research annotating discrepancies between morphological and distributional PoS, as described in the next sections. However, in lieu of carrying out manual annotation from scratch, such as in the ESL, we combined automatic annotation and manual post-editing (as shown in the next section).

3 Data and annotation

The data of VALICO-UD are drawn from the VALICO corpus (Corino and Marello, 2017), a

³<https://uclouvain.be/en/research-institutes/ilc/cecl/learner-corpora-around-the-world.html>.

⁴COLI, LOCCLI and CAIL2 are developed at Università per Stranieri di Perugia and coordinated by Stefania Spina.

collection of non-native Italian texts elicited by comic strips proposed to the learners. It consists of a selection of narrative and descriptive texts providing a large variety of structures beyond simple presentative/existential constructions.

The portion of VALICO that we selected for the treebank is made up of 237 texts (2,261 LS) organized in four sections as shown in Table 1.

L1	# Texts	# LS Tokens
English (EN)	60	8,285
French (FR)	59	7,301
German (DE)	58	7,417
Spanish (ES)	60	7,365
EN+FR+DE+ES	237	30,368

Table 1: VALICO-UD in figures – LS section.

Although the unpredictability and variation of a learner product, in terms of vocabulary, morphology and syntax, makes parsing a LC an especially challenging task (Corino and Russo, 2016; Díaz-Negrillo et al., 2010), it is highly recommendable for smoothly retrieving interlanguage features. Due to this peculiarity of interlanguage, keeping separated the LS from its specifically built target hypothesis (TH) is highly recommended (Lüdeling et al., 2005).

Our annotation scheme for learner Italian uses the inventory of the Italian UD PoS tags and dependency relations (Bosco et al., 2013; Bosco et al., 2014) and the related guidelines. In addition, we tried to follow as much as possible the ESL treebank to have comparable resources.

First, we trained UDPipe (Straka et al., 2016) on the Italian UD corpora, which include standard texts, ISDT (Bosco et al., 2014), and Twitter posts, POSTWITA-UD (Sanguinetti et al., 2018). Second, we automatically parsed VALICO-UD. Third, we manually corrected the treebank. This step is currently ongoing and we envision the treebank to be released in the UD repository in a few months.

For each sentence in VALICO-UD we provide two distinct versions both annotated in UD and tied to an error encoding system (see Section 3.1): one version for the LS and the other for its TH. The latter will differ from the former only when some errors occur. As a trial for this scheme, we selected one hundred sentences (i.e. sample set) containing each at least one error to be annotated.

```
# sent_id = NameSurname00135LS
# text = Può essere un rubadore perche ha la cara chiusa e minacciata.
# err = Può essere un (RN) (i) rubadore (i) (c) rubatore (c) (RN)
(MI) (i) perche (i) (c) perche (c) (MI) ha la (FNL) (i) cara (i)
(c) faccia (c) (FNL) chiusa e (DJ) (i) minacciata (i)
(c) minacciosa (c) (DJ).
# segment =
# typo = 8 ADJ, 11 VERB
# foreign = 8 NOUN
# context = 4 NOUN
1 Può potere AUX VM - 4 aux
2 essere essere AUX V - 4 cop
3 un uno DET RI - 4 det
4 rubadore rubadore NOUN S - 0 root
5 perche perché CONJ CS - 6 mark
6 ha avere VERB V - 4 advcl
7 la il DET RD - 8 det
8 cara caro NOUN S - 6 obj
9 chiusa chiuso ADJ A - 8 amod
10 e e CCONJ CC - 11 cc
11 minacciata minacciato ADJ A - 9 conj
12 . PUNCT FS - 4 punct
```

```
# sent_id = NameSurname00135TH
# text = Può essere un rubatore perché ha la faccia chiusa e minacciosa.
# err = Può essere un (RN) (i) rubadore (i) (c) rubatore (c) (RN)
(MI) (i) perche (i) (c) perche (c) (MI) ha la (FNL) (i) cara (i)
(c) faccia (c) (FNL) chiusa e (DJ) (i) minacciata (i)
(c) minacciosa (c) (DJ).
# segment =
# typo = 8 ADJ, 11 VERB
# foreign = 8 NOUN
# context = 4 NOUN
1 Può potere AUX VM - 4 aux
2 essere essere AUX V - 4 cop
3 un uno DET RI - 4 det
4 rubatore rubatore NOUN S - 0 root
5 perché perché CONJ CS - 6 mark
6 ha avere VERB V - 4 advcl
7 la il DET RD - 8 det
8 faccia faccia NOUN S - 6 obj
9 chiusa chiuso ADJ A - 8 amod
10 e e CCONJ CC - 11 cc
11 minacciosa minaccioso ADJ A - 9 conj
12 . PUNCT FS - 4 punct
```

Figure 1: Example of two CoNLL-U trees of the LS (left) and TH (right) number #35: *He-can to-be a thief because he-has the face closed and threaten_PP*.

3.1 Error Annotation

In writing the TH we decided to adhere as much as possible to the LS and to focus on linguistic correctness (e.g. grammaticality) rather than linguistic appropriateness (e.g. register) (Reznicek et al., 2013)⁵. For this reason, sometimes we sacrificed naturalness for the sake of adherence to the LS. This principle was applied also to lexical errors requiring replacement. For instance, in Figure 1, the term “rubadore” in the LS was replaced with “rubatore” and not with its more common synonym “ladro”, *thief*.⁶ With this principle in mind, we decided to correct words if they are not present neither in the VINCA corpus⁷ (the reference corpus specifically compiled for VALICO and containing texts based on the same comic strips but written by Italian native speakers) nor in our reference dictionary, *Il Nuovo Vocabolario di Base della Lingua Italiana* (De Mauro, 2016). In fact, the VINCA corpus is quite small and the language used sounds quite unnatural though being produced by speakers whose mother tongue is namely Italian (see Corino and Marelllo (2017, p. 12)).

Once the target hypotheses are written, we applied to them a coding system based on Nicholls (2003), which was used also in the NUCLE (Dahlmeier et al., 2013) and FCE (Yannakoudakis et al., 2011) corpora. Our system follows Nicholls’s same principle: “the first letter repre-

sents the *general type of error* (e.g. wrong form, omission), while the second letter identifies the *word class of the required word*”.

To provide a finer-grained description of errors, we used a large variety of letters in the first and second position (e.g. I: inflection, X: auxiliary) and a third letter which encodes information about some grammatical features (e.g. T: tense, M: mood, G: gender) (Simone, 2008, pp. 303–346) and other phenomena involved (e.g. capitalization, language transfer and government). Finally, Nicholls included a catch-all code (CE: complex error) to cover complex, multiple errors. In our sample set, we did not use it because we managed to describe all errors encountered using nested XML tags. However, we do not exclude that, applying the error codes to the whole corpus, we might find particularly complex errors which need to be marked using this code.

Figure 1 shows an annotation example of a LS along with its corresponding TH in the typical CoNLL-U format and with the resource-specific fields used to encode the error information. The **sent_id** field contains the identification code of the sentence: in the example, NameSurname001 (anonymized here) indicates the unique identifier of the text and refers to the transcribers name and surname; the following two-digit number, 35 in the example, indicates the position of the sentence in the text; finally, LS or TH indicates learner sentence and target hypothesis, respectively. The **text** field contains the uncoded sentence (which can be the learner sentence or the target hypothesis). The **err** field contains the error annotation based on

⁵In the future we plan to provide a second TH, focusing on linguistic appropriateness.

⁶Although “rubadore” is reported and marked as obsolete in the Italian Dictionary Olivetti, “rubatore” is the variant reported in De Mauro (2016), our reference dictionary.

⁷<http://www.valico.org/vinca.html>

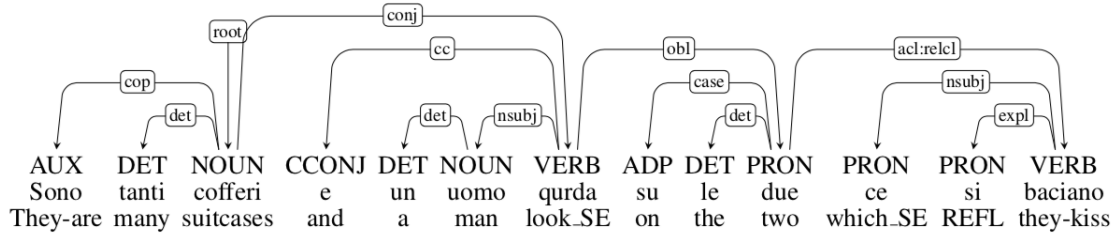


Figure 2: LS #10.

<SVS#><MAX><i> Sono </i><c> Ci Sono </c></MAX><i> Ci Sono </i><c> Ci sono </c> </SVS#>
 <IDG#><FNL><i> tanti cofferi </i><c> tanti valigie </c> </FNL><i> tanti valigie </i><c> tante valigie
 </c></IDG#> e un uomo <MAR><i></i><c> che </c></MAR> <SV><i> qurda </i><c> guarda </c></SV> <UT>
 <IDG> <i> sulle </i><c> sui </c></IDG> <i> sui </i><c> i </c></UT> due <SAR><i> ce </i><c> che </c></SAR>
 si baciano.

Figure 3: Error-annotated sentence #10.

the coding scheme introduced above. The **foreign** field includes the index and the PoS of the words which are considered errors due to language transfer. The **context** field contains the index and the PoS of the words which need replacement due to wrong context-bound lexical choices⁸. Finally, in line with the ESL, we used the **segment** field when a sentence was wrongly divided and the **typo** field to indicate PoS distributional-morphological discrepancies.

In the error-annotated sentence (the “err” field mentioned above), we report the wrong form(s) inside the <i>_</i> tag and the corrected form(s) inside the <c>_</c> tag. Figure 3 shows three examples of nested tag and two examples of *cascade* errors (i.e. an error which is due to the correction of another token) (Andorno and Rastelli, 2009, p. 52). The <MAX>_</MAX> tag at the beginning of the sentence, for example, indicates a missing existential-construction pronoun, i.e. “Sono” (are) instead of “Ci sono” (there are). After the insertion of the missing pronoun “Ci”, the capital “S” in “Sono” needs to be changed into a lowercase “s”: this is a case in which we have a cascade capitalization error and we mark it adding a hashtag after the normal error code, as in <SVS#>_</SVS#>. Another cascade error is found in the next nested tag: we have an Inflection Determiner Gender error which is caused by the correction of the expression “tanti cofferi”, involving a determiner and a noun (“cofferi” is a

German word adapted to Italian and meaning luggage); thus, we have a cascade <IDG#>_</IDG#> tag which embeds a <FNL>_</FNL> tag (Form Noun Language_transfer). The next three tags, <MAR>_</MAR>, <SAR>_</SAR> and <SV>_</SV>, indicate Missing pronoun (A) Relative (“che”, *that*), Spelling pronoun Relative (“ce” instead of “che”) and Spelling Verb errors (“qurda” instead of “guarda”, *look*), respectively. There is, finally, another example of nested tag involving an Inflection Determiner Gender and an Unnecessary preposition errors; this has been used to indicate the multiple-step shift from the LS “sulle” (*on the Fem.Pl*) to its TH counterpart “i” (*the Masc.Pl*): the shift involved a change in the gender of the article (from feminine to masculine) and the drop of the preposition “su” (*on*), mistakenly used in the LS.

In order to ensure consistency across different annotators, the error annotation guidelines provide a hierarchical order to be applied when dealing with nested tags. We organized the errors in a pyramid with at the bottom mechanical errors (i.e. tokenization, capitalization, spelling and punctuation) and, proceeding towards the apex, morphological (derivation and inflection), lexical (form and replace), and syntactic (missing, unnecessary and word order) errors. For example, following this hierarchical order, mechanical errors should be corrected before a syntactic error. However, cascade errors make an exception and change the correction order, as we seen in Figure 3 in which we have a cascade capitalization error (SVS#) caused by a missing pronoun error (MAX)

⁸Only those choices in which there is no mismatch between distributional and morphological PoS are registered in this field.

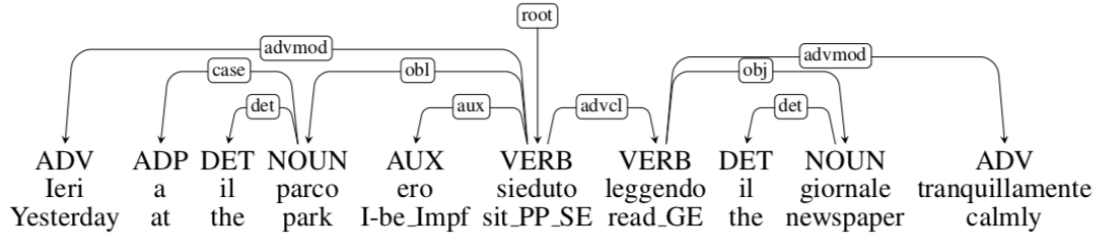


Figure 4: LS #88.

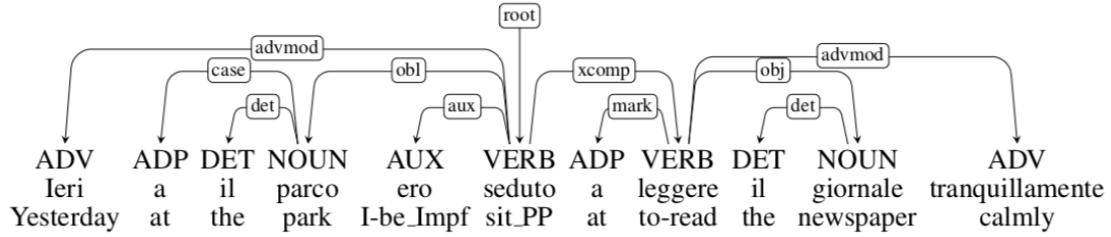


Figure 5: TH #88.

and a cascade inflection error (IDG#) due to a lexical error (FNL).

In the LS sample set, containing 1,860 tokens, we marked 496 errors (which represent 26,66% of the LS sample set tokens) distributed as shown in Table 2.

Error category	Tag	# occ	% tot
Derivation	D	24	4.84%
Form	F	71	14.31%
Inflection	I	72	14.51%
Spelling	S	92	18.55%
Word segmentation	T	16	3.22%
Word order	W	15	3.02%
Missing word	M	76	15.32%
Unnecessary word	U	55	11.09%
Replace word	R	75	15.12%
Total	—	496	—

Table 2: Error categories as encoded in the first letter (general error type) and their distribution in the sample set.

4 From VALICO to VALICO-UD

In this Section we describe how we applied literal annotation to the (morpho-)syntactic structure of the LS in particular, relying on the Universal Dependencies scheme.

Literal Annotation

We annotated UD PoS and relations sticking as

much as possible to the literal reading of the learner sentence, thereby creating a treebank in line with the two existing learner treebanks in the UD framework (ESL and CFL).

Argument Structure: When some extraneous or unnecessary prepositions occur, we annotate the dependencies accordingly. Figure 2 shows a LS in which the verb “guardare”, *look*, is used as an intransitive verb, thus we annotate its direct object as an oblique⁹.

Missing or Unnecessary Words: We annotate literally when there are missing or unnecessary words. In the example in Figure 2 the clitic pronoun “ci” is missing, thus we treated “sono” as a copular verb. There are other cases in which the clitic pronoun “ci” is mistakenly combined with the verb *to be* forming an existential clause, and consequently causing a distributional mismatch (e.g. LS: “[...] non *ci era pericoloso o violento*”, TH: “[...] non *era pericoloso o violento*”¹⁰). In these cases we mark in the “typo” field the morphological PoS and in the PoS column the distributional PoS, cf. Figure 1.

Extraneous Word Forms: When the learner misuses existent word forms, we annotate them literally. In Figure 4, the learner used a gerund, “leggendo” (*reading*), instead of the infinitive “a

⁹In all the examples SE stands for spelling error, REFL for reflexive pronoun, PP for past participle, GE for gerund and Impf for imperfect tense.

¹⁰LS: “[...] not *there it-be Impf dangerous or violent*”, TH: “[...] not *it-be Impf dangerous or violent*”.

leggere” (*to read*). We then labeled it as an adverbial clause in the LS (Figure 4) and as an open clausal complement in the TH (Figure 5).

Exceptions to Literal Annotation

Spelling: Some examples of spelling errors are presented in Figure 2. We lemmatize and PoS-tag them referring to their correct versions, similarly to Andorno and Rastelli (2009, p. 58). Thus, “ce” was treated as “che”, *which*,¹¹ and “qurda” as “guarda” *look*.

Word Formation: We do not treat literally valid words that are contextually implausible. We consider them differently depending on the PoS of the intended word: if the intended word has the same PoS we signal it in the “context” field (e.g. LS: “[...] salvando una ragazza *indefessa*”, TH: “[...] salvando una ragazza *indifesa*”¹²), if it is different in the “typo” field (cf. Figure 1).

Nonexistent Words: In cases in which the learner wrote a word which does not exist in Italian and it is arguably a foreign word, we signal it in the “foreign” field¹³. In the example in Figure 1 the word “cara” (i.e. an adjective translatable into *beloved*) is arguably a transfer from the Spanish noun meaning *face*. In this case we lemmatize it with the correct lemma of “cara”. In addition, in the “typo” field we mark the occurring mismatch between distributional and morphological PoS.

Word Tokenization: If one word is mistakenly segmented into two, we use the “goeswith” relation, as germane to UD annotation guidelines¹⁴. If two words are mistakenly segmented into one, we use X as PoS and decide the relation on a case-by-case basis. For example in LS: “[...] butta tutto *perterra*”, TH: “[...] butta tutto *per terra*”¹⁵ we assigned to “perterra” PoS ‘X’ and dependency relation ‘obl’.

5 Inter-Annotator Agreement

As stated above, the complete manual revision of the treebank is still in progress; however, with the aim of assessing the annotation quality of this preliminary sample set, as well as the quality of the annotation guidelines (especially the ones con-

cerning the LS section) both LS and TH sections were annotated by two independent annotators. The inter-annotator agreement was then computed, considering two measures in particular: UAS (Unlabeled Attachment Score) and LAS (Labeled Attachment Score) for the assignment of both parent node and dependency relation, and the Cohen’s kappa coefficient (Cohen, 1960) for dependency relations only (similarly to Lynn (2016)). UAS and LAS were computed with the script provided in the second CoNLL shared task on multilingual parsing (Zeman et al., 2018)¹⁶. The results are reported in Table 3, and though showing slightly higher results for the TH set, overall they are very close across the sets. Especially as regards the LS section, this is evidence of the guidelines clarity and of the annotators’ consistency, even when dealing with non-canonical syntactic structures.

set	UAS	LAS	kappa
LS	92.11%	88.63%	0.8988
TH	92.47%	88.88%	0.9068

Table 3: Agreement results on the sample set of both LS and TH.

6 Conclusion and future work

In this paper we introduced VALICO-UD and proposed an annotation scheme suitable for texts of learner Italian encompassing both UD and error annotation. Our scheme follows the principle of “literal annotation” and takes PoS and dependency morphological-distributional mismatches into account. Our error tag set seems adequate to book-mark errors, providing also a fine-grained description of some of them.

There are a number of possible applications for the monolingual parallel treebank proposed in this paper. In the near future, we plan to apply the tree edit distance to LS and TH to measure linguistic competence. Recently, the tree edit distance has been applied to various tasks (Emms, 2008; Tsarfaty et al., 2011; Plank et al., 2015), and a study has formalized the notion of *syntactic anisomorphism* (Ponti et al., 2018). We aim to explore a correlation between these notions and the linguistic competence to describe the achievements of foreign language learners.

¹¹When “ce” is used instead of “c’è”, *there is*, we treat it as a single token and mark it as root, in line with what we would have done if it were “c’è”.

¹²LS: “[...] saving a *untiring* girl”, TH: “[...] saving a *vulnerable* girl”.

¹³The lemma will be its Italian (quasi-)equivalent.

¹⁴<https://universaldependencies.org/u/overview/typos.html>

¹⁵[...] he-throw everything *on the ground*.

¹⁶<http://universaldependencies.org/conll18/evaluation.html>

References

- Cecilia Maria Andorno and Stefano Rastelli. 2009. Un'annotazione orientata alla ricerca acquisizionale. In Cecilia Maria Andorno and Stefano Rastelli, editors, *Corpora di italiano L2: tecnologie, metodi, spunti teorici*, pages 49–70. Guerra.
- Yevgeni Berzak, Jessica Kenney, Carolyn Spadine, Jing Xian Wang, Lucia Lam, Keiko Sophie Mori, Sebastian Garza, and Boris Katz. 2016. Universal Dependencies for Learner English. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 737–746.
- Cristina Bosco, Montemagni Simonetta, and Simi Maria. 2013. Converting Italian Treebanks: Towards an Italian Stanford Dependency Treebank. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 61–69.
- Cristina Bosco, Felice Dell'Orletta, Simonetta Montemagni, Manuela Sanguinetti, and Maria Simi. 2014. The EVALITA 2014 Dependency Parsing Task. In *Proceedings of EVALITA 2014 Evaluation of NLP and Speech Tools for Italian*, pages 1–8.
- Adriane Boyd, Jirka Hana, Lionel Nicolas, Detmar Meurers, Katrin Wisniewski, Andrea Abel, Karin Schöne, Barbora Stindlová, and Chiara Vettori. 2014. The MERLIN corpus: Learner Language and the CEFR. In *Conference on Language Resources and Evaluation*, pages 1281–1288.
- Jacob Cohen. 1960. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20(1):37–46.
- Elisa Corino and Carla Marelllo. 2017. *Italiano di stranieri. I corpora VALICO e VINCA*. Guerra.
- Elisa Corino and Claudio Russo. 2016. Parsing di Corpora di Apprendenti di Italiano: un Primo Studio su VALICO. In *Proceedings of the 3rd Italian Conference on Computational Linguistics, CLiC-it 2016*, pages 105–110.
- Daniel Dahlmeier, Hwee Tou Ng, and Siew Mei Wu. 2013. Building a large annotated corpus of learner English: The NUS corpus of learner English. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 22–31.
- Tullio De Mauro. 2016. *Il Nuovo Vocabolario di Base della Lingua Italiana*. Internazionale, <http://www.internazionale.it/opinione/tullio-de-mauro/2016/12/23/il-nuovo-vocabolario-di-base-della-lingua-italiana>.
- Ana Díaz-Negrillo, Detmar Meurers, Salvador Valera, and Holger Wunsch. 2010. Towards Interlanguage POS Annotation for Effective Learner Corpora in SLA and FLT. *Language Forum*, 36(1-2):139–154.
- Martin Emms. 2008. Tree Distance and Some Other Variants of Evalb. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation*, pages 1373–1379.
- Roger Garside, Geoffrey N. Leech, and Tony McEnery. 1997. *Corpus Annotation: Linguistic Information from Computer Text Corpora*. Taylor & Francis.
- Jeroen Geertzen, Theodora Alexopoulou, and Anna Korhonen. 2013. Automatic Linguistic Annotation of Large Scale L2 Databases: The EF-Cambridge Open Language Database (EFCAMDAT). In *Proceedings of the 31st Second Language Research Forum*, pages 240–254.
- Sylviane Granger, Estelle Dagneaux, Fanny Meunier, and Magali Paquot. 2009. *International Corpus of Learner English*. Louvain University Press.
- Sylviane Granger. 2008. Learner Corpora. In Anke Lüdeling and Merja Kytö, editors, *Corpus Linguistics*, volume 1, pages 259–275. Walter de Gruyter.
- Derrick Higgins, Chaitanya Ramineni, and Klaus Zechner. 2015. Learner Corpora and Automated Scoring. In Sylviane Granger, Gaëtanelle Gilquin, and Fanny Meunier, editors, *The Cambridge Handbook of Learner Corpus Research*, pages 587–604. Cambridge University Press.
- Yan Huang, Akira Murakami, Theodora Alexopoulou, and Anna Korhonen. 2018. Dependency Parsing of Learner English. *International Journal of Corpus Linguistics*, 23(1):28–54.
- Scott Jarvis and Magali Paquot. 2015. Learner Corpora and Native Language Identification. In Sylviane Granger, Gaëtanelle Gilquin, and Fanny Meunier, editors, *The Cambridge Handbook of Learner Corpus Research*, pages 605–628. Cambridge University Press.
- Claudia Leacock, Martin Chodorow, and Joel Tetrault. 2015. Automatic Grammar- and Spell-Checking for Language Learners. In Sylviane Granger, Gaëtanelle Gilquin, and Fanny Meunier, editors, *The Cambridge Handbook of Learner Corpus Research*, pages 567–586. Cambridge University Press.
- John Lee, Herman Leung, and Keying Li. 2017. Towards Universal Dependencies for Learner Chinese. In *Proceedings of the NoDaLiDa 2017 Workshop on Universal Dependencies (UDW 2017)*, pages 67–71.
- Anke Lüdeling, Maik Walter, Emil Kroymann, and Peter Adolphs. 2005. Multi-level error annotation in learner corpora. In *Proceedings of Corpus Linguistics 2005*, volume 1, pages 14–17.
- Teresa Lynn. 2016. *Irish Dependency Treebanking and Parsing*. Ph.D. thesis, Dublin City University, Ireland and Macquarie University, Sydney, Australia.

- Shervin Malmasi. 2016. *Native Language Identification: explorations and applications*. Ph.D. thesis, Macquarie University, Sydney, Australia.
- Tony McEnery and Richard Xiao. 2011. What corpora can offer in language teaching and learning. In Eli Hinkel, editor, *Handbook of Research in Second Language Teaching and Learning*, volume 2, pages 364–380. Routledge.
- Hwee Tou Ng, Siew Mei Wu, Ted Briscoe, Christian Hadiwinoto, Raymond Hendy Susanto, and Christopher Bryant. 2014. The CoNLL-2014 shared task on grammatical error correction. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning: Shared Task*, pages 1–14.
- Diane Nicholls. 2003. The Cambridge Learner Corpus: Error coding and analysis for lexicography and ELT. In *Proceedings of the Corpus Linguistics 2003 Conference*, volume 16, pages 572–581.
- Barbara Plank, Héctor Martínez Alonso, Željko Agić, Danijela Merkle, and Anders Søgaard. 2015. Do dependency parsing metrics correlate with human judgments? In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning*, pages 315–320.
- Edoardo Maria Ponti, Roi Reichart, Anna Korhonen, and Ivan Vulić. 2018. Isomorphic transfer of syntactic structures in cross-lingual NLP. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pages 1531–1542.
- Norma A. Pravec. 2002. Survey of Learner Corpora. *ICAME journal*, 26(1):8–14.
- Marwa Ragheb and Markus Dickinson. 2012. Defining syntax for learner language annotation. In *Proceedings of COLING 2012: Posters*, pages 965–974.
- Marc Reznicek, Anke Lüdeling, and Hagen Hirschmann. 2013. Competing target hypotheses in the falko corpus. In Ana Díaz-Negrillo, Nicolas Ballier, and Paul Thompson, editors, *Automatic treatment and analysis of learner corpus data*, volume 59, pages 101–123. John Benjamins Publishing Company.
- Manuela Sanguinetti, Cristina Bosco, Alberto Lavelli, Alessandro Mazzei, Oronzo Antonelli, and Fabio Tamburini. 2018. PoSTWITA-UD: an Italian Twitter Treebank in Universal Dependencies. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation*, pages 1768–1775.
- Raffaele Simone. 2008. *Fondamenti di linguistica*. Laterza.
- Milan Straka, Jan Hajic, and Jana Straková. 2016. Udpipes: Trainable pipeline for processing conll-u files performing tokenization, morphological analysis, pos tagging and parsing. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation*, pages 88–99.
- Yukio Tono. 2003. Learner corpora: design, development and applications. In *Proceedings of the Corpus Linguistics 2003 conference*, pages 800–809.
- Reut Tsarfaty, Joakim Nivre, and Evelina Andersson. 2011. Evaluating dependency parsing: Robust and heuristics-free cross-annotation evaluation. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 385–396.
- Helen Yannakoudakis, Ted Briscoe, and Ben Medlock. 2011. A new dataset and method for automatically grading esol texts. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 180–189.
- Daniel Zeman, Jan Hajič, Martin Popel, Martin Potthast, Milan Straka, Filip Ginter, Joakim Nivre, and Slav Petrov. 2018. CoNLL 2018 shared task: Multilingual parsing from raw text to universal dependencies. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 1–19.

Building an Italian Written-Spoken Parallel Corpus: a Pilot Study

Elisa Dominutti, Lucia Pifferi

Università di Pisa

elisa.dominutti@gmail.com

luciapiff@gmail.com

Felice Dell’Orletta, Simonetta Montemagni

Valeria Quochi

ILC-CNR

name.surnamen@ilc.cnr.it

Abstract

This paper presents a pilot study towards the creation of a monolingual written-spoken parallel corpus in Italian, featuring two main novelties in the general landscape of spoken corpora: the alignment with the written counterpart of the same content and the spoken variety dealt with, represented by transcriptions of radio news broadcasting.

1 Introduction

Nowadays, the contrast between written and spoken language does no longer represent a clear-cut opposition. The emergence of modern communication technologies such as radio, television and new (digital) media led to important changes in the analysis of the diamesic variation. Under this view, the opposition spoken vs. written language is reformulated in terms of a continuum with prototypical written and spoken language at the extreme poles and within which a cline of intermediate linguistic varieties can be recognised, mixing, to a different extent, features of the two. Nencioni (1976) defined the extreme poles of this continuum as the *parlato-parlato* (‘spoken-spoken’) variety, i.e. casual, spontaneous conversation, and the *scritto-scritto* (‘written-written’) variety, i.e. planned, formal, written language. Besides the typical contexts envisaging the use of spoken language—which require all participants to be present in the same environment, that the conversation is held in turns and that speakers make sure their messages are getting across—different contexts can be imagined: among them, the radio and television language which, despite being spoken, present traces of textual organisation recall-

ing the written language. Nencioni (1976) qualifies this variety of language use as *parlato-scritto* (‘spoken-written’), a label that emphasises its hybrid nature characterised by the co-occurrence of traits typical of both written and spoken language. From a different perspective, Ong (1982) refers to this variety as ‘secondary orality’, i.e. “an orality not antecedent to writing and print, as primary orality is, but consequent and dependent upon writing and print”.

In addition to this socio-linguistic interest, the issue also bears relevance for computational approaches as it has a substantial impact on the perceived naturalness of human-machine interaction. Indeed, one of the reasons why speech synthesis applications still produce unnatural speech, apart from bad prosody is that written language is generally not suitable, i.e. comprehensible, direct and effective, in spoken contexts (Kaji et al., 2004). With the rise and quick spread of Virtual Reality (VR) and Augmented-Reality (AR) applications, moreover, the mismatch between written and spoken language styles brings about serious technological limitations because unnaturalness of the virtual agents translates into bad human comprehension and/or distrust in those agents altogether. It is thus no longer sufficient to pass a written message to the speech synthesizer, but such a message needs to be transformed in a form suitable to be spoken in the specific context of use. In order to be able to do this, corpus data is needed such as a monolingual parallel aligned corpus of written and spoken texts about the same content. A corpus designed in this way is of fundamental importance for: a) investigating the features of the *parlato-scritto* language variety, its similarities and differences with respect to the written language; and b) for creating the prerequisites for the design and development of tools for monitoring the communicative effectiveness of texts with respect to their production mode and for support-

Copyright © 2019 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

ing the semi-automatic generation or transformation of texts to be delivered orally. Such a corpus represents an important novel contribution in the area of language corpora; generally in fact corpora target either written or spoken language. Some corpora indeed also include sections with transcriptions of spoken language: see for instance the Brown corpus for English. On the front of spoken corpora, large corpora of spoken Italian were produced, some aiming at specific purposes, like CiT (*Corpus di Italiano Trasmesso*) (Spina, 2000) or LIR (Maraschio et al., 2004), while others aiming at representing Italian in a wider perspective like C-ORAL-ROM (Cresti and Moneglia, 2005). Some of them take into account only a few aspects of the linguistic variability, mainly the diaphasic and in some cases diamesic dimension.

Our *Corpus Italiano Parallelo Parlato Scritto* ('Spoken Written Italian Parallel Corpus', henceforth CIPPS) features two fundamental novelties in the general landscape of spoken corpora: the alignment with a written counterpart of the same content and the type of spoken variety dealt with.

2 Background and related works

Notwithstanding the differences between written and spoken language styles and the impact it bears on human-machine interaction, little computational work has been devoted to develop data and methods for "transforming" a written text in a text suitable for a specific spoken context.

Previous works mostly deal with the transformation of spoken language into grammatically valid, correct written language that can be parsed by standard NLP tools—see for instance Marimuthu and Devi (2014) and Giuliani et al. (2014). However, the rise and spread of VR and AR applications seem to call for the need to appropriately tackle also the other direction, i.e. the transformation of written into (diamesically) appropriate spoken language, which presents different challenges¹.

Few studies have been devoted to the automatic transformation or generation of suitable spoken language, mostly on Japanese. Among these, Murata and Isahara (2001) describe an interesting model to perform different kinds of paraphrasing tasks, that is to transform sentences according to

different predefined criteria. Interestingly, in their experiments both on sentence compression and on transformation from written language to spoken language they manage to apply the same algorithm applied to different data and obtain good results. For the latter experiment, they used a monolingual parallel corpus of academic papers and transcripts of oral presentations and built a system that learns re-writing rules according to the defined criteria. In the former case re-writing rules were learnt from dictionaries.

Kaji and colleagues (2004; 2005) worked on the transformation of written language to spoken language style in Japanese, approaching the issue as a lexical paraphrasing problem, for which they constructed an ad-hoc written-spoken web corpora focused on the connotational differences related to the *suitability for orality* of expressions. Their method learns predicate paraphrases from a dictionary and then uses the corpus to statistically determine whether an expression is suitable to be spoken.

More recently, Matsubara and Hayashi (2012) report about an application for generating spontaneous news speech in a news speech delivery service. They approach the issue as a text generation task and develop a rule-based system for automatically generating news speech scripts—to be read via speech synthesis—starting from newspaper articles. Their approach however focuses on a specific stylistic difference peculiar to Japanese hardly portable to other languages and does not involve any kind of parallel aligned data.

3 Pilot corpus creation

In this work we describe our first attempts at building a parallel written-spoken corpus that might ultimately be useful to train a system for the transformation of written text into text suitable to be spoken. We focus on two different language varieties within the spoken-written language continuum, mentioned in section 1, namely radio spoken language and newspaper written language. This focus was dictated both by the need to neutralize the effects possibly deriving from considering different topics, textual genres and/or communication contexts, and by the practical need of finding readily available data to run the pilot. Thus the present data-set is built by aligning newspaper articles, taken as representatives of the written-written variety and news broadcasting via radio,

¹VR/AR is currently a hot topic especially in both educational and industrial-training contexts (Akçayır and Akçayır, 2017; Żywicki et al., 2018; Gattullo et al., 2019; Heinz et al., 2019; Albayrak et al., 2019).

Day	Num of news	Average lenght
13/05/2003	150	479
15/05/2003	144	523
17/05/2003	148	480
23/05/1995	119	578
25/05/1995	125	547
27/05/1995	124	549
Tot	810	526

Table 1: Written corpus

Day	Num of news	Average lenght
13/05/2003	365	60
15/05/2003	321	57
17/05/2003	156	73
23/05/1995	1184	66
25/05/1995	1106	60
27/05/1995	598	83
Tot	3730	66.5

Table 2: Spoken corpus

taken as representatives of the spoken–written variety.

3.1 Data selection and preparation

Given the goals defined above, our first step was to collect the materials for building the pilot data-set.

For the spoken data-set we chose the *Lessico di italiano Radiofonico* corpus (LIR)(Maraschio et al., 2004)², which consists in transcriptions of various Italian radio broadcast channels sampled in 1995 and 2003 and contains various types of annotations among which: broadcaster, text genre, speaker, communication type, self-corrections, breaks, etc. In particular, we selected the transcriptions of radio news by Radio RAI1, Radio RAI2 and Radio RAI3³ which amount to 6 days altogether: the 23rd, 25th, 27th May 1995, and the 13th, 15th and 17th 2003.

The written data-set was created by taking all news articles published in La Repubblica on the same dates⁴. Tables 1 and 2 report the figures of the data-sets.

In the case of the spoken corpus extensive extraction and cleaning work was required because the original transcriptions include many different genres (e.g. advertisements, interviews, entertainment,...) and several different annotation tags.

3.2 Spoken corpus cleaning

From the selected days of the LIR corpus we needed to extract only the transcriptions of news text. The original texts in fact contain several types of annotations, all in a proprietary tagging format, and news are easily recognisable. So, for each day mentioned, we created a data-set by collating the news of the different radio broadcasters,

thus obtaining 6 spoken data-sets, one for each day. These were subsequently cleaned by using regular expressions that removed all annotation tags, which provided us with raw text data for the alignment experiment.

In Table 2 we can see the number of news extracted for each day and their average length in terms of tokens. Interestingly, but not surprisingly, we observe that newspaper articles on average are longer than radio news.

4 Alignment methodology

Once we gathered, cleaned and normalised the relevant data, we proceeded to align written and spoken texts on the basis of topic and semantic equivalence. Since the spoken transcriptions do not have an explicit marking of sentence boundaries, for the time being alignment is performed at text level; we leave sentence-level alignment for future work.

Given the six spoken data-sets and their corresponding written ones we experimented with two different methods to perform their alignment. One is based on the *Jaccard index* (Jaccard henceforth), the other method on *cosine similarity* (Cosine henceforth). Both algorithms followed one common preliminary step: for each data-set we took into consideration only nouns, verbs, adjectives and numerals, i.e. semantically heavy words.

The first method calculates similarity using the Jaccard index as a statistical index. In general, this coefficient measures the similarity of two samples through the ratio between the size of the intersection and the size of the union of the sample sets; so, in this case, the numerator is given by the overlap of words of the two documents, i.e. the number of relevant words present in both. The denominator instead is the sum of the relevant words of both documents. The computation can be represented

²Source: <http://www.accademiadellacrusca.it/it/attivita/lessico-frequenza-dellitaliano-radiofonico-lir>

³The news transcriptions of the other broadcasters were too short for our purposes.

⁴source: <https://ricerca.repubblica.it/>

as follows:

$$J(A, B) = \frac{|\text{overlapping words in A, B}|}{|\text{words A} + \text{words B}|} \quad (1)$$

The range of acceptable values stands between 0 (for the couples of documents that have no words in common) and 0,5 (for the couples of documents with the highest similarity, i.e. with all relevant words in common).

The second method computes the cosine similarity between a vector representing all the relevant words in a spoken text and a vector representing a written text. Each vector contains a number of components identical to the amount of relevant words contained in the texts, the value of each component being the *TFiDF* value of the corresponding word in the represented text. Once all vectors were built, we compared each spoken-vector with every written-vector and computed their cosine similarity. Finally, considering values of similarity in decreasing order we reorganised the pairs and completed document-alignment. The range of acceptable values for the Cosine method stands between 0 and 1, with values close to 1.0 indicating strong similarity.

4.1 Alignment evaluation

The two methods illustrated above produced twelve output files, six for each method, all ranked on the basis of their similarity score in decreasing order. For each of them we considered the first one hundred spoken-written text pairs and manually evaluated their alignments on a binary scale with respect to their information content. News about the same topics, events or facts were considered good alignments. We decided to stop the evaluation at the first one hundred pairs, because after this threshold the recognised alignments were no longer significant (i.e. algorithms aligned pairs of documents with different topics).

On the 1200 manually assessed pairs we then calculated the **accuracy** of the two methods. We considered accuracy as the ratio between the number of aligned pairs in particular range of distance values and the total number of couples in the same range.

The graphics in Figures 1 and 2 show method accuracy for each range of similarity values, using both the 1995 and 2003 data. For example, in the range of values between 0,1 and 0,2, the Cosine method has an accuracy of 6% with the 1995 data and 22% with the 2003 data. As we advance in the

higher similarity bands, we notice a growing trend for both methods, but while for Cosine we observe a gradual growth, the Jaccard method shows a faster rise. Moreover, we notice that most of the alignments occur in the lowest similarity range of value, while in the higher similarity ranges we found very few alignments (see Table 3 and 4 for details).

Remembering that the range of admissible values are different for the two methods let us focus on the results.

Cosine alignment evaluation Cosine for both data-sets has an accuracy of 100% in the range of values 0,8-0,7 and 0,6-0,5, while for the range 0,2-0,3 it has an accuracy of 6% for 1995's data-sets and 22% for 2003's data. Figure 1 shows a gap between 0.7 and 0.6 for 2003's data. That is because, for this data-set, the cosine method did not assign values in this range. Overall, Cosine total accuracy is 61%, 53% on 1995 data and 69% on 2003 data.

Jaccard alignment evaluation In the range 0,3-0,2 the Jaccard method has an accuracy of 100% on both datasets; while for the 1995 data it drops to 53% in the range 0,2-0,1 and to 47% in the range 0,1-0,6. For the 2003 data in the range 0-2,01 the accuracy is 86%, which decreases to 44,8% in the range 0,1-0,07. Also in this case, as reported in Table 4, we have few alignments in higher distances despite the number of lower ones.

Overall, Jaccard total accuracy is 50%, 50% on 1995 data and 51% on 2003 data.

According to this evaluation, Cosine using *TFiDF* values is the best method for aligning our data.

Here is an example of text pairs with high cosine similarity values (0,7-0,8):

[Spoken]: [...] il diario di Paul
Mccartney [...] rottura con i Beatles
è stato riconsegnato [...] al cantante
il giorno dopo il concerto dei fori
imperiali [...] Mccartney ha avuto
modo di rileggere quel preziosissimo
diario stracolmo di ricordi e ha
confermato l'autenticità [...] alcune frasi portano il segno della
storia "Arriva John per discutere lo
scioglimento della partnership" giugno
millenovecentosettanta la fine dei
Beatles

[Written]: [...] il diario di Paul McCartney [...] rottura con i Beatles è stato riconsegnato [...] al cantante, il giorno dopo il concerto dei fori imperiali. [...] sir Paul ha avuto modo di rileggere quel preziosissimo diario stracolmo di ricordi, e ha confermato l'autenticità dell'agenda. [...] alcune frasi portano il segno della storia: ``arriva John per discutere lo scioglimento della partnership''. giugno 1970, la fine dei Beatles. [...]

What follows instead is an example of a good alignment with lower cosine similarity values (0,3-0,2)⁵:

[Spoken]: se non mi attaccassero non mi difenderei [...] spiega Berlusconi [...] "Io sono un moderato" ripete il premier "Mi difendo da teoremi folli che non attaccano me ma il presidente del consiglio" [...]

[Written]: Berlusconi al contrattacco "Denuncerò chi mi offende". [...] E aggiunge che le accuse contro di lui si basano su "Teoremi folli". Teoremi ai quali [...] "Ho dato la risposta più moderata, contenuta e misurata che si potesse dare". [...]

The first example is also an example of high Jaccard similarity values (0,3-0,2).

In general, with both methods, the pairs of documents correctly aligned in the lower ranges of similarity show considerable differences in terms of lexical items and possibly linguistic structures, and thus represent a very interesting set of pairs for future investigation. Regarding higher ranges, we find a greater lexical overlap and a lower variation in linguistic structure. Comparing the pairs correctly aligned by the two methods we counted 77 identical ones, while the number of different pairs derived from Jaccard is 220, and from Cosine 260. In total we obtained 557 different correctly aligned pairs.

5 Pilot corpus profiling

The final pilot CIPPS corpus consists of 557 text pairs corresponding to the correctly aligned and manually validated pairs of spoken and written

⁵For reasons of space the example texts have been arbitrarily shortened.

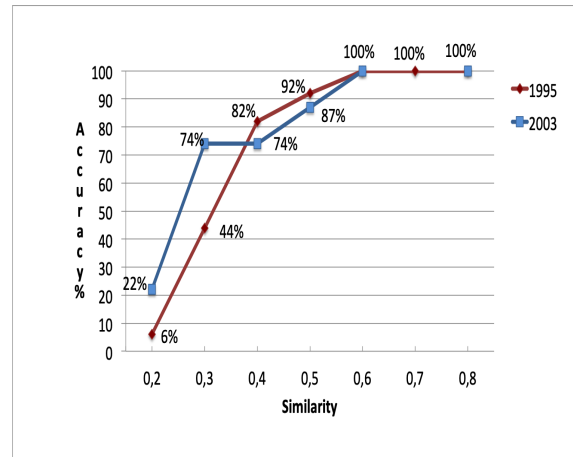


Figure 1: Cosine accuracy

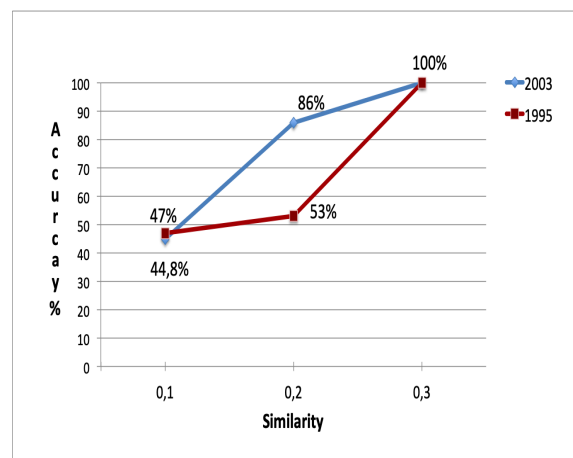


Figure 2: Jaccard accuracy

Distance	1995		2003	
	Correct	Tot	Correct	Tot
0,8-0,7	1	1	2	2
0,7-0,6	3	3	0	0
0,6-0,5	6	6	4	4
0,5-0,4	12	13	26	30
0,4-0,3	45	55	45	61
0,3-0,2	90	206	123	167
0,2-0,1	1	16	8	36
TOT	158	300	208	300

Table 3: Cosine Accuracy (1995-2003)

Distance	1995		2003	
	Correct	Tot	Correct	Tot
0,3-0,2	5	5	3	3
0,2-0,1	41	77	37	43
0,1-0,065	103	218	114	254
TOT	149	300	154	300

Table 4: Jaccard accuracy (1995-2003)

documents resulting from both alignment methods. It can thus be taken as a gold-standard corpus of content aligned text pairs of news for the dates and years mentioned in section 3.1.

This section reports on our preliminary contrastive analysis of CIPPS using Monitor-IT (Montemagni, 2013), so as to establish basic linguistic profiling of the two language varieties represented in the corpus. This analysis was done with a specific view to investigating similarities and differences in the distribution of multi-level linguistic cues (we focus here on lexical and morpho-syntactic features) both within the corpus and against prototypical written and spoken language (in the future, we plan to extend this analysis to the underlying syntactic structure).

Let us first compare the two sections of the CIPPS corpus. On the one hand, highly correlated features between the CIPPS written and spoken sections concern the distribution of nouns (both common and proper) and adjectives as well as verbal forms used in the third person singular; the correlation was calculated with the Spearman's Correlation Coefficient ($p\text{-value} \leq 0.05$). On the other hand, statistically significant different features across the spoken and written corpus sections detected with the Wilcoxon test ($p\text{-value} \leq 0.05$) include specific verbal forms, deictic elements and determiners, prepositions and acronyms, as well as lexical richness (measured in terms of Token/Type Ratio). In particular, if verbal moods such as gerundive, subjunctive, infinitive and conditional are typically associated with written articles, the 1st and 2nd person of verbs in both singular and plural forms are typical of the spoken news reports. Demonstrative determiners and pronouns represent significant features of the spoken variety, whereas acronyms and lexical richness measured in terms of Token-Type Ratio characterise the written CIPPS section.

For what concerns the comparison of the linguistic profiling results sketched above with what we know from the literature about features of spoken vs. written language, we observe that the widely acknowledged fact that spoken language is less complex than written language is declinated here in quite a peculiar way. Differently from the 'spoken-spoken' variety characterised by a reduced number of nouns and consequently by a lower noun/verb ratio (ranging between 0,80 and 1, (Montemagni, 2013)), the 'spoken-written' va-

riety shares with prototypical written language a twice higher noun/verb ratio, which, according to Biber (1988), is typical of informative texts. On the other hand, it shares with prototypical spoken language the more frequent use of deictic elements, of 1st/2nd person reference in verbal forms, lexical repetition.

These findings, which need to be further elaborated and explored, confirm the hybrid nature of the spoken language variety represented in the CIPPS corpus, which is in line with the trend reported in the literature that the language of the radio shares features with both spontaneous oral and written language varieties.

6 Conclusions and Future work

In this paper we have presented our first experiments towards the creation of the CIPPS, a monolingual written-spoken parallel aligned corpus. The data for this pilot was drawn from existing corpora and archives, it was automatically aligned on the basis of two statistical methods and finally manually validated. To the best of our knowledge, this is the first attempt to build such a corpus and more research is needed to improve its potentials and increase its magnitude.

Among the open issues to be approached first is the lack of punctuation in the spoken part of the corpus, which makes automatic alignment with the written counterpart too coarse. As mentioned in the introduction, a corpus like ours might also be precious as a training set for the development of a system for transforming written into suitable spoken texts. Although little work has been done in this direction, the time is now ripe to tackle the challenge and we plan to start experimenting with both paraphrasing methods—as mentioned in section 1—and with monolingual machine translation, taking inspiration from Quirk et al. (2004) and Wubben et al. (2012). In this perspective, however, the first necessary step is to increase corpus size and improve alignment.

Acknowledgments

This work was partially supported by the 2-year project ADA, Automatic Data and documents Analysis to enhance human-based processes, funded by Regione Toscana (BANDO POR FESR 2014-2020).

References

- Murat Akçayır and Gökçe Akçayır. 2017. Advantages and challenges associated with augmented reality for education: A systematic review of the literature. *Educational Research Review*, 20:1 – 11.
- M. S. Albayrak, A. Öner, I. M. Atakli, and H. K. Ekenel. 2019. Personalized training in fast-food restaurants using augmented reality glasses. In *2019 International Symposium on Educational Technology (ISET)*, pages 129–133, July.
- Douglas Biber. 1988. *Variation across speech and writing*. Cambridge University Press.
- Emanuela Cresti and Massimo Moneglia. 2005. *C-ORAL-ROM, Integrated Reference Corpora for Spoken Romance Languages*. John Benjamins.
- M. Gattullo, V. Dalena, A. Evangelista, A. E. Uva, M. Fiorentino, A. Boccaccio, M. Ruta, and J. L. Gabbard. 2019. A context-aware technical information manager for presentation in augmented reality. In *2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, pages 939–940, March.
- Manuel Giuliani, Thomas Marschall, and Amy Isard. 2014. Using ellipsis detection and word similarity for transformation of spoken language into grammatically valid sentences. In *Proceedings of the SIGDIAL 2014 Conference, The 15th Annual Meeting of the Special Interest Group on Discourse and Dialogue, 18-20 June 2014, Philadelphia, PA, USA*, pages 243–250.
- Mario Heinz, Sebastian Büttner, and Carsten Röcker. 2019. Exploring training modes for industrial augmented reality learning. In *Proceedings of the 12th ACM International Conference on Pervasive Technologies Related to Assistive Environments, PETRA 2019, Island of Rhodes, Greece, June 5-7, 2019*, pages 398–401.
- Nobuhiro Kaji and Sadao Kurohashi. 2005. Lexical choice via topic adaptation for paraphrasing written language to spoken language. In *Natural Language Processing - IJCNLP 2005, Second International Joint Conference, Jeju Island, Korea, October 11-13, 2005, Proceedings*, pages 981–992.
- Nobuhiro Kaji, Masashi Okamoto, and Sadao Kurohashi. 2004. Paraphrasing predicates from written language to spoken language using the web. In *Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics, HLT-NAACL 2004, Boston, Massachusetts, USA, May 2-7, 2004*, pages 241–248.
- Nicoletta Maraschio, Stefania Stefanelli, Stefania Bucchini, and Marco Biffi. 2004. Dal corpus lir: prove e confronti lessicali. In Federico Albano Leoni, Francesco Cutugno, Massimo Pettorino, and Renata Savy, editors, *Atti del Convegno Nazionale “Il Parlato Italiano”*, page 36.
- K Marimuthu and Sobha Lalitha Devi. 2014. Automatic conversion of dialectal tamil text to standard written tamil text using fst. In *Proceedings of the 2014 Joint Meeting of SIGMORPHON and SIGFSM, Baltimore, Maryland, USA, June 27, 2014*, pages 37–45.
- Shigeki Matsubara and Yukiko Hayashi. 2012. Personalization of news speech delivery service based on transformation from written language to spoken language. In Toyohide Watanabe, Junzo Watada, Naohisa Takahashi, Robert J. Howlett, and Lakhmi C. Jain, editors, *Intelligent Interactive Multimedia: Systems and Services*, pages 449–457, Berlin, Heidelberg. Springer Berlin Heidelberg.
- Simonetta Montemagni. 2013. Tecnologie linguistico-computazionali e monitoraggio della lingua italiana. *Studi italiani di linguistica teorica ed applicata*, (XLII(1)):145–172.
- Masaki Murata and Hitoshi Isahara. 2001. Universal model for paraphrasing - using transformation based on a defined criteria. *CoRR*, cs.CL/0112005.
- Giovanni Nencioni. 1976. Parlato-parlato, parlato-scritto, parlato-recitato. *Strumenti critici*, (29).
- Walter J. Ong. 1982. *Orality and Literacy: The Technologizing of the Word*. Methuen.
- Chris Quirk, Chris Brockett, and William B. Dolan. 2004. Monolingual machine translation for paraphrase generation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing, EMNLP 2004, A meeting of SIGDAT, a Special Interest Group of the ACL, held in conjunction with ACL 2004, 25-26 July 2004, Barcelona, Spain*, pages 142–149.
- Stefania Spina. 2000. Il corpus di italiano televisivo (cit): struttura e annotazione. In *Atti del Convegno SILFI*.
- Sander Wubben, Antal van den Bosch, and Emiel Krahmer. 2012. Sentence simplification by monolingual machine translation. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers - Volume 1, ACL ’12*, pages 1015–1024, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Krzysztof Żywicki, Przemysław Zawadzki, and Filip Górski. 2018. Virtual reality production training system in the scope of intelligent factory. In Anna Burduk and Dariusz Mazurkiewicz, editors, *Intelligent Systems in Production Engineering and Maintenance – ISPEM 2017*, pages 450–458, Cham. Springer International Publishing.

Italian and English Sentence Simplification: How Many Differences?

Martina Fieromonte[•], Dominique Brunato[◊], Felice Dell’Orletta[◊], Giulia Venturi[◊]

[•] University of Pavia

m.fieromonte@gmail.com

[◊]Istituto di Linguistica Computazionale “Antonio Zampolli” (ILC-CNR)

ItaliaNLP Lab - www.italianlp.it

{dominique.brunato, felice.dellorletta, giulia.venturi}@ilc.cnr.it

Abstract

The paper proposes a cross-linguistic analysis of two parallel monolingual corpora conceived for automatic text simplification in two languages, Italian and English. The aim is to find similarities and differences in the process of simplification in two typologically different languages. To carry out the comparison, 1,000 sentences were extracted from the two corpora and annotated with a scheme previously used to annotate simplification phenomena.¹

1 Introduction

In recent years, the availability of parallel monolingual corpora has boosted the adoption of data-driven techniques for the task of automatic text simplification (ATS). These corpora are in general aligned at sentence level and consist of complex sentences paired with their simple version. However, except for English which can rely on two large parallel corpora, i.e. the Parallel Wikipedia Corpus² (Coster and Kauchak, 2011)(ParWik) and the Newsela corpus³ (Xu et al., 2015), these corpora are scarce or rather small in other languages. To reduce time and effort required for the construction of parallel corpora, some works tried new approaches to automatically or semi-automatically collect such resources, e.g. Coster and Kauchak (2011), Yatskar et al. (2010), Brunato et al. (2016), Tonelli et al. (2016). Moreover to take advantage of empirical data, most of these resources were annotated with rules aimed at identifying the typologies of modifications an original sentence goes through during the process of simplification. The inspection can be considered use-

ful for several reasons: it permits i) to detect and classify a set of necessary transformations in TS, ii) to assess if a given corpus complies with user requirements and simplification tasks and iii) to evaluate the impact of simplification operations on target populations. If the corpus investigation also encompasses a cross-linguistic comparison, it might also shed light on peculiarities and similarities underlying the process of simplification across languages. However, so far this last issue has been rather ignored with the exception of Gonzalez-Dios et al. (2018), who compared how macro-simplification operations derived from different annotation schemes are distributed in Italian, Basque and Spanish parallel corpora. This paper intends to explore this under-investigated perspective and proposes a cross-linguistic analysis of two parallel monolingual corpora, i.e. the Italian corpus PaCCSS-IT (Parallel Corpus of Complex–Simple Aligned Sentences for Italian) (Brunato et al., 2016) and the English Parallel Wikipedia Corpus (Coster and Kauchak, 2011). Through this comparison, the paper tries to answer the following three questions:

1. To what extent can an annotation scheme conceived for the annotation of simplification in one language be used to annotate simplifications in other language?
2. Are there any differences or similarities in the distribution and nature of simplification operations in the two languages?
3. If we find differences, to what extent do they depend on language only, or on the type of corpora?

To answer these questions, 1,000 paired sentences were extracted from the two corpora and annotated with the scheme described in Brunato et al. (2016). This allows us to carry out a quantitative and qualitative analysis focused on understanding the nature of the modifications occurring in the datasets.

¹Copyright ©2019 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

²<http://www.cs.pomona.edu/~dkauchak/simplification/>

³<https://newsela.com/data/>

2 Related work

Given the relevance of parallel monolingual corpora in ATS, many projects have driven their attention on the development of these resources. The main approaches in the literature vary from the manual simplification of original texts carried out by experts (see e.g. Xu et al. (2015) in English, Bott and Saggion (2014) in Spanish, Brunato et al. (2015) in Italian), to the alignment of already existing text collections, containing same-topic documents written in two different styles, a complex and a simple one. It is the case of e.g. Coster and Kauchak (2011) and Tonelli et al. (2016), both relying on the Wikipedia corpus but in a different way. The first is based on the alignment between articles extracted from the standard and the Simple English Wikipedia, a project started in 2001 containing English Wikipedia pages written in basic English; the latter relies on the edits that users had made on the Italian Wikipedia and explicitly marked as instances of simplification. A further strategy was envisaged by Brunato et al. (2016), who first collected a corpus of sentences sharing the same meaning from a large web corpus, and then ranked the most similar pairs according to their linguistic complexity assigned by an automatic readability assessment system.

In many cases, existing ATS corpora were also annotated with rules to make explicit the most frequent operations occurring in the process of sentence simplification and distinguishing different typologies of linguistic phenomena involved in sentence transformation. The classification of simplification operations is typically two-level based, i.e. it contains a few macro-level operations and for some of them a more specific subclass which can depend on the size of the unit affected (e.g. sentence, phrase or word) or the linguistic level at which the operation applies (i.e. lexical, syntactic, discourse). Comparing ParWik with the manually simplified corpus Newsela, Xu et al. (2015) also noticed that the approach adopted to construct ATS resources has an impact on the type of simplification phenomena. For instance, there are more differences between paired sentences before and after simplification in Newsela, suggesting that complex linguistic structures are often retained in ParWik. Simple sentences in ParWik contains also longer words, together with a greater number of function words and punctuation. Similar differences related to the approach under-

lying the construction of parallel corpora were also observed in Italian. For example, the comparison reported in Tonelli et al. (2016) between a corpus of Wikipedia edit stories and two corpora of heterogeneous texts for young readers manually simplified according to different strategies (i.e. a structural and an intuitive one) proved the existence of differences in terms of the linguistic level affected by simplification. They concern for instance the distribution of some simplification operations and the average of operations per sentence. As regards the first aspect, in manually simplified corpora, editors opted for a word-level lexical substitution, while Wikipedia editors for a phrase-level substitution. As regards the second aspect, the Wikipedia edit story corpus contains an average lower distribution of simplification per sentence. Though related to these works, our contribution differs in that it adds a cross-linguistic level of comparison and also tries to provide an overview of possible factors affecting the distribution and the nature of simplification operations in ATS corpora.

3 Corpora and annotation scheme

Corpora. The corpora used in the analysis are the Italian corpus PaCCSS-IT and the English Parallel Wikipedia Corpus (ParWik). PaCCSS-IT is a parallel corpus composed of about 63,000 paired sentences, obtained crawling the web. The corpus is the result of a three-step approach strongly shaped by the level of simplification under investigation, i.e. syntactic simplification, consisting in: i) an unsupervised step in which a great amount of sentences with overlapping lexicon and different syntactic structure was clustered according to a similarity metric and automatically aligned⁴; ii) a supervised step aimed to train a classifier to predict the sentence alignment and iii) a readability assessment step aimed at assigning a readability score to the sentences in each pair. ParWik instead was obtained aligning two already existing text collections: the English Wikipedia and the Simple English Wikipedia. The authors aligned paragraphs whose TF*IDF cosine similarity was over a threshold of 0.5. The final corpus consists of 167,000 aligned sentence pairs.

To summarize, the two corpora differ in the fol-

⁴To be part of a cluster a sentence had to share all lemmas with PoS ‘noun’, ‘verb’, ‘numeral’, ‘personal pronoun’ and ‘negative adverb’.

lowing aspects: i) language; ii) corpus collection approach iii) domain of texts; iv) level of simplification under investigation.

	PaCCSS-IT	ParWik
i)	Italian	English
ii)	Web crawling	Wiki-based alignment
iii)	Web corpus	Encyclopedic
iv)	Mainly syntax	Lexicon+Syntax

Table 1: Corpora design criteria.

Annotation of simplification operations. The comparison was conducted on 1,000 sentence pairs randomly extracted from the two corpora. To make possible the comparison, the sentences were annotated with the scheme in Table 2, previously conceived to annotate PaCCSS-IT.

Simplification operations
Deletion
Insertion
Verbal Features
Lexical Substitution
Reordering
Sentence Type
Residual

Table 2: Annotated simplification operations.

The manual annotation was carried out by one of the authors using the web-based annotation tool *Brat*⁵. As reported in the next section, the results of the manual annotation process provide an answer to the first question. The adopted schema originally designed to identify simplification operations within different typologies of parallel corpora in another language is able to cover almost all transformations in ParWik. The main limit is that the scheme does not take into account one of the more typical simplification operations, that is splitting long and complex sentences into one or more shorter ones (Narayan et al., 2017). This is because it was conceived to make explicit the transformations occurring in the PaCCSS-IT corpus, which only includes 1:1 pairs, i.e. for each ‘complex’ sentence only one ‘simple’ version exists. To annotate this operation in ParWik, we used the tag residual.

4 Corpora analysis

4.1 Distribution of simplification operations

Figure 1 reports the average distribution of simplification operations in the two corpora. As we

⁵<https://brat.nlplab.org/>

can see, the first three most frequent operations in PaCCSS-IT are: ‘deletion’, ‘verbal features’ and ‘insertion’ and in ParWik ‘deletion’, ‘lexical substitution’ and ‘insertion’. Excluding deletion, the differences resulted to be statistically significant for all operations, according to the Chi-squared test (p value <0.05).

At first glance, these results seem to suggest that language-specific factors affect the process of simplification. However, it is interesting to note that a qualitative analysis of these findings partially rules out this hypothesis, suggesting instead to interpret the differences also in view of the other criteria reported in Table 1. Specifically, the impact of language is limited to the different distribution of the ‘verbal feature’ operation. In PaCCSS-IT, it represents 29% of the total number of annotated operations while it is much less frequent in ParWik ($<5\%$). In particular, the distribution of this operation in the Italian corpus is mainly due the higher number of verbs at the conditional mood, which are transformed into indicative in the simplified sentence. As expected, verbs in ParWik are mostly at the indicative in both versions of the sentence. However, this different distribution has to be read also in view of another factor, i.e. the domain of texts in the corpora. Since it has been crawled from the web, PaCCSS-IT contains heterogeneous domains and many complex sentences belong to a ‘written to be spoken’ style, which implies the use of polite forms, expressed in Italian with the conditional mood. As a consequence of the different domain of texts contained in the two corpora, we can also observe a gap concerning the frequency of ‘insertion’. Specifically, the encyclopedic nature of texts in ParWik may require the insertion of glosses and explanations to improve the understanding of complex terms. The lower frequency of lexical substitution operations in the Italian corpus (8.9% vs 23.9%) is easily explained if one considers the main purpose for which the corpus was designed, i.e. the investigation of syntactic simplification. On the contrary, editors of Simple Wikipedia are explicitly recommended “to write using Basic English words”⁶.

4.2 Linguistic analysis

The diversity between the two corpora affects also the nature of the linguistic phenomena subjected

⁶https://simple.wikipedia.org/wiki/Wikipedia:How_to_write.Simple.English_pages

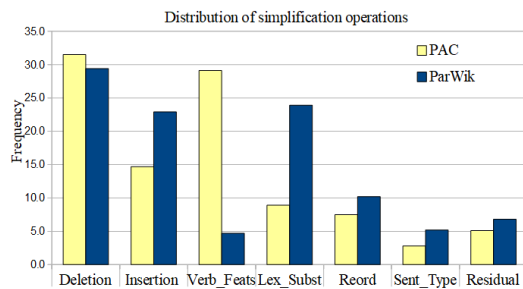


Figure 1: Distribution of simplification operations.

to simplification. This means that the type of linguistic elements which are, for example, deleted, inserted or substituted might be different. Again this variance is poorly attributable to the proprieties of the languages at play. In the following, we will try to outline a categorization of the linguistic elements subjected to modifications in the two corpora, providing an example for each case.

Deletion. This operation involves the deletion of single words or clauses. In particular, we observe a similar trend in the two corpora with the deletion of functional words, modal adverbs and adjectives alone or entire clauses containing these parts of speech.

- C: The main bar at King's is far older **and is the site of more informal meeting between students**. [ParWik]
- S: The main bar at King's is far older. [ParWik]
- C: **Probabilmente sospetto che** non sarebbe comunque una buona idea. (Probably, I suspect that it would not be however a good idea.) [PaCCSS-IT]
- S: Non fu una buona idea. (It was not a good idea.) [PaCCSS-IT]

Insertion. In both corpora auxiliaries and full verbs are inserted. Moreover in ParWik also nouns and pronouns are inserted as subjects of the new sentence, typically as a consequence of a split. As said, this does not occur in PaCCSS-IT, where however, implicit-explicit clause transformation implies the insertion of explicit elements, such as articles and verbs.

- C: Spese del presente grado di giudizio compensate tra le parti costituite. (Expense of the present level of justice compensated among the parts) [PaCCSS-IT]
- S: **Le** spese del presente grado di giudizio **possono essere** compensate tra le parti. (The expense of the present level of justice can be compensated among the parts). [PaCCSS-IT]

As said before, ParWik editors tend to insert explanations of complex terms and concepts. The

contribute to simplicity of this type of insertion is quite clear in:

- C: According to the Armenian tradition, Saint Jude suffered martyrdom about 65 AD in Beirut, in the Roman province of Syria, together with the apostle Simon the Zealot, with whom he is usually connected.
- S: St. Jude was martyred, **killed for his beliefs**, with another apostle, Simon the Zealot in Beirut, Lebanon, around AD 65.

Instead, it is debatable in:

- C: Velvet Revolver is an American hard rock supergroup consisting of former Guns N' Roses members Slash, Duff McKagan, and Matt Sorum, alongside Dave Kushner formerly of punk band Wasted Youth.
- S: Velvet Revolver, **VR**, is a Grammy Award-winning rock supergroup. The members of the band are Slash **guitarist**, Duff McKagan **bassist**, **backing vocals**, Matt Sorum **drums of Guns N' Roses**, **Scott Weiland lead vocals of Stone Temple Pilots** and Dave Kushner **guitarist** of Wasted Youth.

Lexical substitution the more striking difference between the two corpora concerns this operation, not only in terms of frequency but also in respect of the type of substitution. In PaCCSS-IT, the operation affects only the substitution of words whose PoS was not considered in the clustering step, e.g. adjectives, adverbs and articles, etc. Moreover the substitution does not always contribute to the simplification of the sentence: this means that in some cases the complex term may be not replaced with a simpler synonym. In ParWik instead the operation affects phrase and sentence level, yielding to real paraphrases.

- C: Il concorrente è preventivamente stato avvertito **per** assistere all'operazione (The concurrent had been informed in advance to assist to the operation [PaCCSS-IT])
- S: Il concorrente è stato avvertito preventivamente, **affinché** possa assistere all'operazione. (The concurrent had been informed in advance in order to assist to the operation) [PaCCSS-IT]
- C: Sporting venues in the city include the Millennium Stadium the national stadium for the Wales national rugby union team and the Wales national football team, SWALEC Stadium the home of Glamorgan County Cricket Club, Cardiff City Stadium the home of Cardiff City football team and Cardiff Blues rugby union team, Cardiff International Sports Stadium the home of Cardiff Amateur Athletic Club and Cardiff Arms Park the home of Cardiff Rugby Club. [ParWik]
- S: Cardiff has one of the largest stadiums in the United Kingdom, the Millennium Stadium, where important world sports matches and concerts happen. Other big stadiums in the city are the Cardiff City Stadium, where the main football and rugby teams play, and the SWALEC Stadium where cricket is played. [ParWik]

Verbal features As said before, the Italian ‘conditional→indicative’ transformation does not occur in the English corpus, where instead the tag ‘verbal features’ was assigned to mark voice modification and ‘indefinite→finite’ mood transformations.

- C: Salve, **avrei bisogno** di una informazione piuttosto urgente. (Good morning, I would need a rather urgent information.) [PaCCSS-IT]
- S: Ho bisogno di una informazione urgente. (I need a urgent information.) [PaCCSS-IT]
- C: It is most often black but can come in a variety of colors including clear, **allowing** the top of the deck to be decorated. [ParWik]
- S: However, it can come in many different colors like clear. Clear allows the top of the deck to be decorated. [ParWik]

Reordering In general, in PaCCSS-IT, reordering implies the resetting of the canonical word order, while in ParWik there is a tendency to transform noun pre-modifiers in appositive phrases. As regards the position of subordinate clauses, neither of the two corpora assign to them a fixed position, i.e. before or after the main clause, although in ParWik embeddings are often extracted to form a new sentence.

- C: **Un’unica cosa** vorrei aggiungere. (Only a thing I would like to add.) [PaCCSS-IT]
- S: Volevo aggiungere solo una cosa. (I wanted to add only a thing.) [PaCCSS-IT]
- C: The United States presidential election of 1992 had three major candidates: Incumbent **Republican** President George H. W. Bush; **Democratic Arkansas Governor** Bill Clinton, and **independent** Texas businessman Ross Perot. [ParWik]
- S: The United States presidential election of 1992 was on November 3, 1992 in the United States. The three main people running were: George H. W. Bush, a Republican from Texas and the President; Bill Clinton, who was a Democrat and Governor of Arkansas; and Ross Perot an Independent candidate. [ParWik]

Sentence type. Three main phenomena fall under this tag: i) passive-active modification, ii) implicit-explicit clause modification and iii) verbalization-nominalization modification. While the first two modifications occur in both corpora, the third was found only in ParWik. Again, this difference is partly affected by language-dependent factors but it also depends on specific corpus-dependent constraints.

- C: Il presidente, **ricordato che nella seduta di ieri si è svolta la relazione**, dichiara aperta la discussione generale. (The president, reminded that the reporting was held in the yesterday part-session, declares open the general discussion.) [PaCCSS-IT]
- S: Il presidente ricorda che nella seduta di ieri è stata svolta la relazione introduttiva e dichiara quindi aperta la discussione generale. (The president reminds that in the yesterday part-session was held the introductory reporting and declares open the general discussion.) [PaCCSS-IT]
- C: **Findings of** coins indicate that the Romans were in Buxton throughout their occupation. [ParWik]
- S: Roman coins have been found in Buxton. [ParWik]

5 Conclusions and future works

The paper proposed a cross-linguistic comparison between two monolingual parallel corpora for ATS. The comparison tried to answer three main questions. As regards question 1, the annotation stage proved the possibility to use, except few modifications, a language-specific annotation scheme for another language. More than language-specific factors, an in-depth analysis of the annotated pairs of sentences highlighted that the observed differences are due to linguistic phenomena characterizing different textual genres. This is the case for example of modifications due to the insertion of glosses, which is driven by the encyclopedic nature of Wikipedia pages rather than to the specific language. Similarly, textual genre has an impact on the linguistic level involved in the lexical substitution. The higher occurrence of substitutions at phrase level, rather than at word-level, reflects the attempt of Wikipedia editors to make scientific contents clearer and simpler for a wide target population. Corpus-design differences, especially those occurring between manually and automatically derived corpora, may affect the distribution of the simplification operations also within the same genre. This is one of the possible directions that we want to explore in the near future.

Acknowledgments

This work was partially supported by the 2-year project ADA, Automatic Data and documents Analysis to enhance human-based processes, funded by Regione Toscana (BANDO POR FESR 2014-2020).

References

- Stefan Bott and Horacio Saggion. 2014. Text Simplification Resources for Spanish. *Language Resources and Evaluation*. *Language Resources and Evaluation*, 48(1): 93–120.
- Dominique Brunato, Felice Dell’Orletta, Giulia Venturi and Simonetta Montemagni. 2015. *Design and Annotation of the First Italian Corpus for Text Simplification*. Proceedings of the 9th Linguistic Annotation Workshop (LAW15), Denver, Colorado, USA.
- Dominique Brunato, Andrea Cimino, Felice Dell’Orletta and Giulia Venturi. 2016. PaCCSS-IT: A Parallel Corpus of ComplexSimple Sentences for Automatic Text Simplification. *Methods in Natural Language Processing (EMNLP 2016)*, pages 1018.
- William Coster and David Kauchak. 2011. Simple English Wikipedia: a new text simplification task. *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.
- I. Gonzalez-Dios, M. J. Aranzabe, and A. Díaz de Ilaraza. 2018. The corpus of Basque simplified texts (CBST). *Language Resources and Evaluation*, 52 (1) 217–47.
- Shashi Narayan and Claire Gardent and Shay B. Cohen and Anastasia Shimorina. 2017. Split and Rephrase. *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics*.
- Wei Xu, Chris Callison-Burch, and Courtney Napoles. 2015. Problems in Current Text Simplification Research: New Data can Help. *Transactions of the Association for Computational Linguistics*, 3:283–29.
- Sara Tonelli, Alessio Palmero Aprosio, Francesca Saltori. 2016. *SIMPITIKI: a Simplification corpus for Italian*. Proceedings of the Third Italian Conference on Computational Linguistics, Naples, Italy.
- Mark Yatskar, Bo Pang, Cristian Danescu-NiculescuMizil, and Lillian Lee. 2010. For the Sake of Simplicity: Unsupervised Extraction of Lexical Simplifications from Wikipedia. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT ’10, pages 365–368, Stroudsburg, PA, USA. Association for Computational Linguistic.

Error Analysis in a Hate Speech Detection Task: the Case of HaSpeeDe-TW at EVALITA 2018

Chiara Francesconi

Dipartimento di Lingue e Letterature
Straniere e Culture Moderne
University of Turin

`chiara.francesconi@edu.unito.it`

Cristina Bosco

Fabio Poletto

Manuela Sanguinetti

Dipartimento di Informatica

University of Turin

`{bosco,poletto,msanguin}@di.unito.it`

Abstract

Taking as a case study the Hate Speech Detection task at EVALITA 2018, the paper discusses the distribution and typology of the errors made by the five best-scoring systems. The focus is on the sub-task where Twitter data was used both for training and testing (HaSpeeDe-TW). In order to highlight the complexity of hate speech and the reasons beyond the failures in its automatic detection, the annotation provided for the task is enriched with orthogonal categories annotated in the original reference corpus, such as aggressiveness, offensiveness, irony and the presence of stereotypes.

1 Introduction

The field of Natural Language Processing witnesses an ever-growing number of automated systems trained on annotated data and built to solve, with remarkable results, the most diverse tasks. As performances increase, resources, settings and features that contributed to the improvement are (understandably) emphasized, but sometimes little or no room is given to an analysis of the factors that caused the system to misclassify some items.

This paper wants to draw attention to the importance of a thorough error analysis on the performance of supervised systems, as a means to produce advancement in the field. Errors made by a system may entail not only the poorness of the system itself but also the sparseness of the data used in training, the failure of the annotation scheme in describing the observed phenomena or a cue of the data inherent ambiguity. The presence of the same errors in the results of several systems involved in

a shared task may result in also more interesting hints about the directions to be followed in the improvement of both data and systems.

As a case study to carry out error analysis, data from a shared task have been used in this paper. Shared tasks offer clean, high-quality annotated datasets on which different systems are trained and tested. Although often researchers omit to reflect on what caused to system to collect some failures (Nissim et al., 2017), they are an ideal ground for sharing negative results and encourage reflections on "what did not work", an excellent opportunity to carry out a comparative error analysis and search for patterns that may, in turn, suggest improvements in both the dataset and the systems.

Here we analyze the case of the Hate Speech Detection (HaSpeeDe) task (Bosco et al., 2018) presented at EVALITA 2018, the Evaluation Campaign for NLP and Speech Tools for Italian (Caselli et al., 2018). HS detection is a really complex task, starting from the definition of the notion on which it is centered. Considering the growing attention it is gaining, see e.g. the variety of resources and tasks for HS developed in the last few years, we believe that error analysis could be especially interesting and useful for this case, as well as in other tasks where the outcome of systems meaningfully depends on resources exploited for training and testing.

The paper outlines the background and motivations behind this research (Section 2), describes the sub-task on which the study is based (Section 3), reports on the error analysis process (Section 4) and discusses its results (Section 5), and presents some conclusive remarks (Section 6).

2 Background and Motivations

There are several issues connected to the identification of HS: its juridical definition, the subjectivity of its perception, the need to remove potentially illegal content from the web without unjustly re-

Copyright © 2019 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

moving legal content, and a list of linguistic phenomena that partly overlap to HS but need to be kept apart.

Many works have recently contributed to the field by releasing novel annotated resources or presenting automated classifiers. Two reviews on HS detection were recently published by Schmidt and Wiegand (2017) and Fortuna and Nunes (2018). Since 2016, shared tasks on the detection of HS or related phenomena (such as abusive language or misogyny) have been organized, effectively enhancing advancements in resource building and system development. These include HatEval at SemEval 2019 (Basile et al., 2019), AMI at IberEval 2018 (Fersini et al., 2018), HaSpeede at EVALITA 2018 (Bosco et al., 2018) and more. Nevertheless, the growing interest in HS detection suggests that the task is far from being solved: to improve quality and interoperability of resources, to design suitable annotation schemes and to reduce biases in the annotation is still as needed as it is to work on system engineering. Establishing standards and good practices in error analysis can enhance these processes and push towards the development of effective classifiers for HS.

While academic literature is rich with works on human annotation and evaluation metrics, it is not as easy to find works dedicated to error analysis of automated classification systems. This is rather more often found as a section of papers describing a system (see, e.g., (Mohammad et al., 2018)). This section, however, is not always present. To examine the errors made by a system, classify them and search for linguistic patterns appear to be a somewhat undervalued job, especially when the system had an overall good performance. Yet, it is crucial to understand why a system proved to be a weak solution to certain instances of a problem, even while being excellent for other instances.

In the context of COLING 2018, error analysis emerged as one of the most relevant features to be addressed in NLP research¹. This attention to error analysis encouraged authors to submit papers with a dedicated section, with Yang et al. (2018) winning the award for the best error analysis, and is a step towards establishing good practices in the NLP community.

In the wake of this awareness, we apply linguistic insights to one of the annotated corpora

used within the HaSpeede shared task, namely the HaSpeede-TW sub-task dataset (described in Section 3). Characteristics of this dataset make it ideal for our purpose: each tweet is connected to a target and is annotated not only for the presence of HS but for four other parameters. If a comparative analysis of two corpora presenting different textual genres (HaSpeede-TW and HaSpeede-FB) might have offered interesting perspectives, the lack of such characteristic in the FB dataset prevents a thorough comparison. Furthermore, among the in-domain HaSpeede sub-tasks, HaSpeede-TW is the one where systems achieved the lower F_1 -scores, providing thus more material for our analysis.

3 HaSpeede-TW at EVALITA 2018: A Brief Overview

While a description of the HaSpeede task as a whole has been provided in the organizers' overview (Bosco et al., 2018), here we focus on HaSpeede-TW, one of the three sub-tasks into which the competition was structured². The sub-task consisted in a binary classification of hateful vs non-hateful tweets. Training set and test set contain 3,000 and 1,000 tweets respectively, labeled with 1 or 0 for the presence of HS, and with a distribution, in both sets, of around 1/3 hateful against 2/3 non-hateful tweets. Data are drawn from an already existing HS corpus (Poletto et al., 2017), whose original annotation scheme was simplified for the purposes of the task (see Section 4).

Nine teams participated in the task, submitting fifteen runs. The five best scores, submitted by the teams ItaliaNLP (whose runs ranked 1st and 2nd) (Cimino and De Mattei, 2018), RuG (Bai et al., 2018), InriaFBK (Corazza et al., 2018) and sbMMP (von Grünigen et al., 2018), ranged from 0.7993 to 0.7809 in terms of macro-averaged F_1 -score³. They applied both classical machine learning approaches, Linear Support Vector Machine in particular (ItaliaNLP, RuG) and more recent deep learning algorithms, such as Convolutional Neural Networks (sbMMP) or Bi-LSTMs (ItaliaNLP, who adopted a multi-task learning approach ex-

²The other two being HaSpeede-FB, where Facebook data were used both for training and testing the systems, and Cross-HaSpeede, further subdivided into Cross-HaSpeede-FB and Cross-HaSpeede-TW, where systems were trained using Facebook data and tested against Twitter data in the former, and the opposite in the latter.

³All official ranks are available here: <https://goo.gl/xPyPRW>.

¹<https://coling2018.org/error-analysis-in-research-and-writing/>.

ploiting the SENTIPOLC 2016 (Barbieri et al., 2016) dataset as well). Learning architectures resorted to both surface features such as word and character n-grams (RuG) and linguistic information such as Part of Speech (ItaliaNLP).

In the next section, we provide a description of the errors collected from these best five runs as put in relation with the specific factors we chose to analyze in this study, encompassing and merging qualitative and quantitative observations. Our analysis is strictly based on the results provided by those systems. An analysis focused on the features of the systems that determined the errors is unfortunately beyond the scope of this work, as in HaSpeede participants were only requested to provide the results after training their systems.

4 Error Analysis

Error analysis can be used in between runs to improve results or test different feature settings. With the aim of weaving a broader reflection on the especially hard linguistic patterns within a HS detection task, here it is performed *a posteriori* and on the aggregated results of five systems on the HaSpeede-TW test set (1,000 tweets). We focus on the answers given by the majority of the five best systems because we believe they provide a faithful representation of the errors without the noise due to the presence of the worst runs.

The test set was composed of 32.4% of hateful tweets and 67.6% non-hateful tweets. As the first step of our analysis, we compared the gold label assigned to each tweet in the test set with the one attributed by the majority of the five runs considered for the task. An error was considered to occur when the label assigned by the majority of the systems was different from the gold label. If we extend our analysis to all the fifteen submitted runs, 156 out of 1,000 tweets have been misclassified by the majority of them. However, this number increases to 172 if only the five best runs are taken into account.

Regardless of the correct label, agreement among the five best runs is higher than that among all runs and among any other set of runs: those systems which have best modeled the phenomenon on the data provided appear to have made similar mistakes. This supports our hypothesis that errors mostly depend on data-dependent features rather than on systems, which are all different in approach and feature setting.

Even though only the annotation concerning the presence of HS was distributed to the teams, the corpus from which the training and test set of HaSpeede-TW were extracted was provided with additional labels (Poletto et al., 2017; Sanguinetti et al., 2018). These labels (see Table 1) were meant to mark the user’s intention to be aggressive (*aggressiveness*), the potentially hurtful effect of a tweet (*offensiveness*), the use of ironic devices to possibly mitigate a hateful message (*irony*), and whether the tweet contains any implicit or explicit reference to negative beliefs about the targeted group (*stereotype*).

label	values
aggressiveness	no, weak, strong
offensiveness	no, weak, strong
irony	yes, no
stereotype	yes, no

Table 1: The original annotation scheme of the HS corpus that was (partially) used in HaSpeede-TW.

These labels were conceived with the aim of identifying some particular aspects that may intersect HS but occur independently. As a matter of fact, hateful contents towards a given target might be expressed using aggressive tones or offensive/stereotypical slurs, but also in much subtler forms. At the same time, aggressive or offensive content, though addressed to a potential HS target, does not necessarily imply the presence of HS. Our assumption while carrying out this study was that such close, but at times misleading, relation between HS on one side and these phenomena on the other could be considered a source of error for the automatic systems.

In addition, other aspects of both linguistic and extra-linguistic nature were taken into account, so as to complement the analysis. We thus considered the tweets *targets*, i.e. Roma, immigrants and Muslims (also an information available from the original HS corpus). Finally, we selected three features that are typical of computer-mediated communication and social platforms such as Twitter, in particular, the presence of *links*, *multi-word hashtags*, and the use of *capitalized words*.

As for the method adopted, the percentage of errors for the gold positives and the gold negatives in the whole test set was calculated. First, the rates were calculated considering the two labels - hateful and non-hateful - separately, in order to bal-

ance their different distribution in the test set; then the results were halved to represent the whole corpus in percentage and to maintain the proportion between the results of the tags. All the percentages correlating two different tags were calculated this way, so that the results could be easily compared. The percentages of mistakes for each label of the categories were determined and compared to the general result to understand whether they influenced it positively or negatively. Table 2 summarizes the results for each label showing the distribution of the false negatives (FN), false positives (FP), true positives (TP) and true negatives (TN). The error percentages higher than the general result are in bold font.

5 Results and Discussion

In order to find some answers to our research questions and evidence of the influence of the annotated features on the systems' results, we provide in this section an analysis driven by the categories we described in the previous section.

Aggressiveness and Offensiveness. The different degrees of aggressiveness did not affect the systems recall, but we measured more FPs when weak or strong aggressiveness is involved (more than thrice as many as in the overall results when strong aggressiveness is present).

Offensiveness seems to hold a similar but heavier influence on performance, causing better recall but worse precision: FPs are more than doubled when strong offensiveness is present.

The presence of offensiveness is often associated to slurs or vulgar terms: these are not a consistent presence in the dataset (the most vulgar tweets are probably quickly removed by the platform), and mostly appear in tweets classified as HS. However, about half of the non-hateful tweets containing offensive words were wrongly classified as hateful, proving that offensiveness can be misleading for systems. In these cases, a lexicon-based approach can fail, while attention to the context could be crucial: in the most common instances of false positives, in fact, offensive words did not refer to the targets.

HS Targets. Analyzing the three targets of HS allowed us understanding how the systems reacted to different ways of expressing hate.

Most of the errors were caused by the target *Roma*: few hateful tweets were recognized, and

FNs are more than 30%. Results for the target *Immigrants* are similar to the overall performance, only with a slightly higher number of FPs. The target *Muslims* caused a low number of FNs but almost twice as many FPs as in the general performance.

The systems seem to struggle to recognize hateful content against *Roma*: this may be caused by an imbalance in the test set (only 6.3% of tweets with the target *Roma* are labelled as HS, while the targets *Immigrants* and *Muslims* have 12.6% and 13.4% of hateful tweets respectively) or by biases in the annotation.

The poor results achieved in classifying messages with target *Roma* can also be explained by the subtler ways of expressing HS when this target is involved, more heavily based on stereotypes than it happens with the other targets. The hate against the other two targets, in particular Muslims, was instead very explicit. See the following examples extracted from the test set.

2235. *Roma, colpisce una pecora con il pallone: bambino rom accecato da un pastore* <https://t.co/KsSAS3fUx9> @ilmessaggeroit HA DIFESO I SUOI AVERI!⁴ [FN, strong aggressiveness, target: Roma]

4749. @Corriere Uccidere gli islamici, prima di tutto.⁵ [TP, strong aggressiveness, target: religion]

Other features. Some other features were considered in our analysis. The presence of *stereotype* was more frequent in hateful tweets, which caused a slight increase in FPs; conversely, cases of HS without stereotype posed no issues to the systems. Moreover, as expected, the presence of *irony* slightly increased the errors rate both in hateful and non-hateful tweets.

The presence of Twitter's linguistic devices also negatively influenced the results, probably because of the difficulty encountered by systems when some semantic content assumes non-standard forms, e.g. *links*, *multi-word hashtags* and *capitalized words*.

URLs frequently occur in the data, but mostly in non-hateful tweets (although this may be a peculiarity of this dataset). Systems appear to have

⁴"Rome, Roma child hits a sheep with a ball: blinded by a shepherd <https://t.co/KsSAS3fUx9> @ilmessaggeroit HE DEFENDED HIS PROPERTY!"

⁵"@Corriere Kill the Muslims, first of all."

	FN	FP	TP	TN	Gold HS	Gold Not-HS
general	15%	6%	35%	44%	32.3%	67.7%
no aggressiveness	15%	4%	35%	46%	13.5%	56.8%
weak aggressiveness	15%	10%	35%	40%	11.2%	10.1%
strong aggressiveness	15%	19%	35%	31%	7.6%	0.8%
no offensiveness	20%	5%	30%	45%	10.9%	60%
weak offensiveness	13%	11%	37%	39%	14.6%	4.9%
strong offensiveness	12%	16%	38%	34%	6.8%	2.8%
no irony	15%	5%	35%	45%	27.8%	59%
yes irony	18%	9%	32%	41%	4.5%	8.7%
no stereotype	15%	5%	35%	45%	11.6%	49.7%
yes stereotype	15%	8%	35%	42%	20.7%	18%
Immigrants	15%	9%	35%	41%	12.6%	22.4%
Muslims	8%	11%	42%	39%	13.4%	12.2%
Roma	31%	1%	19%	49%	6.3%	33.1%
no link	11%	13%	37%	39%	25.4%	24.4%
yes link	29%	1%	21%	49%	7%	43.2%
multi hashtags	23%	8%	27%	42%	3%	1.9%
no capitalized words	15%	5%	35%	45%	29.1%	64.1%
yes capitalized words	14%	9%	36%	41%	3.3%	3.5%

Table 2: Percentage of correct (TPs and TNs) and erroneous (FPs and FNs) results in relation to the features considered in the analysis, along with the actual distribution of these features in the test set.

troubles recognizing hateful tweets that contain URLs (errors increased by 14%). Conversely, the absence of URLs caused an increase in FPs. This feature is unlikely to be directly connected to hateful language: we rather believe that it could somehow affect predictions regardless of the actual content.

Also multi-word hashtags influenced results, especially for hateful content: their presence increased FNs by 8%. The reason for this kind of error might lie in the fact that our dataset contains some cases where the crucial element in a hateful tweet is precisely the hashtag, as in the example below:

2149. *Quando vedremo lo stessa tema portato in piazza con la stessa forza e determinazione? Mai credo. #stopislam*
⁶ <https://t.co/dDYLZB1BIJ> [multi-word hashtag, FN]

The text in this tweet is not hateful, but an element of hatred is conveyed by the hashtag “#stopislam”.

The ability to separate the multi-word hashtags into the words composing them would improve the

performances of the systems. The tweets with a multi-word hashtag clarifying the text would have a better chance of being correctly identified.

Finally, some capitalized words have been found in the data set, mostly in hateful tweets, which again caused an increase in FPs. Despite their small number, we noticed that, in non-hateful tweets, a higher percentage of capitalized words are named entities (nouns of places, people, newspapers, etc.), while in hateful tweets capitalized words are more often used to intensify opinions or feelings.

Among all the features taken into account, offensiveness seems to have affected the performance in various ways: its absence led systems to classify as non-hateful tweets that are indeed hateful, while its presence caused the inverse error. A possible explanation for this is that, as shown in Sanguinetti et al. (2018), offensiveness does not correlate with HS even though it can be one of its features. The systems might have taken offensive terms as indicators for HS, as also humans tend to do (see for example Bohra et al. (2018)), but this is a false assumption that systems should be trained to avoid. Aggressiveness also caused a certain degree of errors, but only affecting precision.

⁶“When will we see people fighting for the same issue with the same strength and determination? Never, I believe.”

6 Lessons Learned and Conclusion

This paper presents a detailed error analysis of the results obtained within the context of a shared task for HS detection. In our study, we took into account two types of data: content information, provided by gold standard labels assigned to each tweet; and metadata information, namely the presence of URLs, hashtags and capitalized words. Results prove the importance of considering other categories related to that on which the task was centered.

The analysis of performances in relation to URLs poses a controversial result. There are two reasons why tweets collected via Twitter's API may contain a URL: the tweet may have been cut off and a URL automatically generated as a link to the complete tweet, or the URL may be part of the original tweet and lead to an external page. In both cases, unless the URL is followed, the tweet is likely to be harder to understand compared to a tweet that contains no URL. This may cause lower agreement among human judges, and it is a very complicated issue for automated systems to deal with, especially when the meaning of the tweet is unintelligible without first opening the URL. Tweets containing URLs are, for the time being, less reliable as training data and pose a tougher challenge for Sentiment Analysis tasks at large; we encourage an effort towards solving this issue.

As for capitalized words, future work may include investigating how they affect human annotation, as some judges may show a bias towards associating capitalized words to HS or other categories. Furthermore, improvements may come from considering the PoS tags of such words, or the number of consecutive capitalized words.

Multi-word hashtags as well need to be treated with care, as they may affect and even overturn the meaning of the whole tweet. Yet, it happens that a hashtag might require syntactic, semantic and world-knowledge processing in order to be fully understood: for example, by comparing the phrase "stop Islam" with, e.g., "stop harassment", we can see that the word "stop" is not necessarily negative, and it becomes so only because it is followed by the name of a religion whose members are, nowadays and in Western society, particularly subject to discrimination.

Overall, our analysis suggests that systems failures are motivated by the difficulty in dealing with cases where HS is less directly expressed and pave

the way for future work on, e.g., the development of tools that perform a more careful analysis of the text.

Acknowledgments

The work of C. Bosco and M. Sanguinetti is partially funded by Progetto di Ateneo/CSP 2016 (*Immigrants, Hate and Prejudice in Social Media*, S1618_L2_BOSC_01), while that of F. Poletto is funded by Fondazione Giovanni Gorla and Fondazione CRT (*Talenti della Società Civile* 2018).

References

- Xiaoyu Bai, Flavio Merenda, Claudia Zaghi, Tommaso Caselli, and Malvina Nissim. 2018. RuG @ EVALITA 2018: Hate Speech Detection In Italian Social Media. In *Proceedings of Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2018)*. CEUR.org.
- Francesco Barbieri, Valerio Basile, Danilo Croce, Malvina Nissim, Nicole Novielli, and Viviana Patti. 2016. Overview of the Evalita 2016 SENTiment POLarity Classification Task. In *Proceedings of the Fifth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2016)*. CEUR.org.
- Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. 2019. Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in Twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 54–63.
- Aditya Bohra, Deepanshu Vijay, Vinay Singh, Syed Sarfaraz Akhtar, and Manish Shrivastava. 2018. A dataset of Hindi-English code-mixed social media text for hate speech detection. In *Proceedings of the Second Workshop on Computational Modeling of Peoples Opinions, Personality, and Emotions in Social Media*, pages 36–41.
- Cristina Bosco, Dell'Orletta Felice, Fabio Poletto, Manuela Sanguinetti, and Tesconi Maurizio. 2018. Overview of the EVALITA 2018 hate speech detection task. In *Proceedings of Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2018)*. CEUR.org.
- Tommaso Caselli, Nicole Novielli, Viviana Patti, and Paolo Rosso. 2018. EVALITA 2018: Overview of the 6th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. In *Proceedings of Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2018)*. CEUR.org.

- Andrea Cimino and Lorenzo De Mattei. 2018. Multi-task Learning in Deep Neural Networks for Hate Speech Detection in Facebook and Twitter. In *Proceedings of Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2018)*. CEUR.org.
- Michele Corazza, Stefano Menini, Pinar Arslan, Rachele Sprugnoli, Elena Cabrio, Sara Tonelli, and Serena Villata. 2018. Comparing Different Supervised Approaches to Hate Speech Detection. In *Proceedings of Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2018)*. CEUR.org.
- Elisabetta Fersini, Paolo Rosso, and Maria Anzovino. 2018. Overview of the Task on Automatic Misogyny Identification at IberEval 2018. In *Proceedings of the Third Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018), co-located with 34th Conference of the Spanish Society for Natural Language Processing (SEPLN 2018)*, pages 214–228. CEUR-WS.org.
- Paula Fortuna and Sérgio Nunes. 2018. A survey on automatic detection of hate speech in text. *ACM Computing Surveys (CSUR)*, 51(4):85.
- Saif Mohammad, Felipe Bravo-Marquez, Mohammad Salameh, and Svetlana Kiritchenko. 2018. Semeval-2018 task 1: Affect in tweets. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 1–17.
- Malvina Nissim, Lasha Abzianidze, Kilian Evang, Rob van der Goot, Hessel Haagsma, Barbara Plank, and Martijn Wiering. 2017. Sharing is caring: The future of shared tasks. *Computational Linguistics*, 43(4):897–904.
- Fabio Poletto, Marco Stranisci, Manuela Sanguinetti, Viviana Patti, and Cristina Bosco. 2017. Hate Speech Annotation: Analysis of an Italian Twitter Corpus. In *Proceedings of the Fourth Italian Conference on Computational Linguistics (CLiC-it 2017)*. CEUR.org.
- Manuela Sanguinetti, Fabio Poletto, Cristina Bosco, Viviana Patti, and Marco Stranisci. 2018. An Italian Twitter Corpus of Hate Speech against Immigrants. In *Proceedings of the 11th Language Resources and Evaluation Conference (LREC 2018)*.
- Anna Schmidt and Michael Wiegand. 2017. A Survey on Hate Speech Detection using Natural Language Processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*. Association for Computational Linguistics.
- Dirk von Grünigen, Ralf Grubenmann, Fernando Benites, Pius Von Däniken, and Mark Cieliebak. 2018. spMMMP at GermEval 2018 Shared Task: Classification of Offensive Content in Tweets using Convolutional Neural Networks and Gated Recurrent Units. In *Proceedings of GermEval 2018, 14th Conference on Natural Language Processing (KONVENS 2018)*.
- Pengcheng Yang, Xu Sun, Wei Li, Shuming Ma, Wei Wu, and Houfeng Wang. 2018. Sgm: sequence generation model for multi-label classification. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3915–3926.

Nunc Est Aestimandum

Towards an Evaluation of the Latin WordNet

Greta Franzini*, Andrea Peverelli*, Paolo Ruffolo*, Marco Passarotti*,
Helena Sanna°, Edoardo Signoroni°, Viviana Ventura°, Federica Zampedri°

*CIRCSE Research Centre, Università Cattolica del Sacro Cuore, Milan, Italy

°Università degli Studi di Pavia, Pavia, Italy

greta.franzini@unicatt.it

Abstract

English. This paper describes a preliminary expansion and assessment of the Latin WordNet for the purposes of the *LiLa: Linking Latin* project. The objective of this study is to better understand the implications of expanding and evaluating the sense coverage of the Latin WordNet, with a view to identifying the most effective method for its refinement and inclusion in the LiLa Knowledge Base of Latin resources. Our test empirically demonstrates the inadequacy for Latin of a common semi-automated approach of expansion and informs potential lines of improvement for the resource.¹

1 Introduction

WordNets are among the most used lexico-semantic resources in Natural Language Processing (NLP). Indeed, their value is such as to warrant the annual *Global WordNet Conference*, which is now in its tenth edition.² In the words of Fellbaum (1998, p. 52):

WordNet [...] is perhaps the most widely used electronic dictionary of English and serves as the lexicon for a variety [sic] of different NLP applications including Information Retrieval (IR), Word Sense Disambiguation (WSD), and Machine Translation (MT).

Since the release of the *Princeton WordNet* (hereafter PWN) in the mid 1980s (Miller et al., 1990), interest in providing WordNets for modern

languages has far exceeded that for historical languages. With the exception of the *Historical Thesaurus of English*, whose purpose is not dissimilar to that of a WordNet but whose distinct structure sets it apart from this type of resource,³ the only two historical language WordNets in existence today are the Latin (Minozzi, 2017) and the Ancient Greek WordNets (Bizzoni et al., 2014): both have limited lexical coverage and the Latin WordNet (hereafter LWN) is particularly noisy (see Section 3). Their incompleteness poses significant challenges to a number of computational analyses, thus restricting the scope for lexico-semantic research.⁴

The study described here falls within the scope of the *LiLa: Linking Latin* project (Passarotti et al., 2019).⁵ In its wider effort to connect linguistic resources and NLP tools for Latin in a Linked Data Knowledge Base, LiLa is conducting a first assessment of the LWN. Besides being structurally compatible with LiLa, a refined LWN is essential to the Knowledge Base as a connector between Latin and resources in other languages, thus meeting a growing need in the field of Linguistic Linked Open Data (Chiarcos et al., 2013).

This paper describes a preliminary assessment of the LWN with a view to better understanding how to approach its expansion and evaluation: Sections 2 and 3 briefly outline existing research in WordNet evaluation and the structure of the LWN, respectively; Section 4 details our evaluation method; Section 5 discusses our preliminary results; finally, Section 6 summarises our contribution and focusses on directions for future research.

2 Related Work

Evaluation. To evaluate a WordNet is to evaluate its *coverage* of a specific linguistic domain or of

¹Copyright 2019 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

²<http://globalwordnet.org/>

³<https://ht.ac.uk/>

⁴Most recently Franzini et al. (2018).

⁵<https://lila-erc.eu> (2018-2023).

an entire language (period), be that qualitative (accuracy) or quantitative (inclusivity). Among others, Bodenreider et al. (2003) conducted a quantitative evaluation of the bio-genetic domain in the PWN by mapping a list of relevant terms against manually-established semantic classes of nominal synsets, and proved PWN’s coverage to be satisfactory. A study by Hajič et al. (2004) sought to manually evaluate and improve the Czech WordNet using the lexico-semantic annotation of the Prague Dependency Treebank. In spite of achieving poor inter-annotator agreement, their outcome can inform future improvements of the resource.

The first automated, qualitative evaluation of a WordNet was performed by Nadig et al. (2008) on the PWN. Using dictionary definitions, the authors applied different extraction and matching algorithms to automatically validate 38,840 nominal synsets (corresponding to 103,620 lemmas) and 56,203 hypernym-hyponym noun pairs, reaching accuracy rates of 70% and 70.88%, respectively. These high rates are hardly surprising, given that the PWN is a handmade resource; nevertheless, they give us an indication as to what might be expected from a similar evaluation performed on automatically-generated WordNets.

Extension. Researchers looking to extend WordNets in languages other than English typically do so by semi-automatically comparing lemmas and synsets in their target language against the contents of the PWN with the help of bilingual dictionaries and linguistic resources. This is the case of the Arabic WordNet (AWN), extended through semi-automated comparison with a lexicon of modern standard Arabic and the PWN (Abouenour et al., 2013). As far as Latin is concerned, a parallel evaluation effort to the one described here is being conducted by the University of Exeter.⁶ In Exeter, the lexical coverage of the LWN has been automatically extended to 70,000 lemmas using *Freedict.com* as well as the Lewis and Short (1879) and Whitaker’s Words Latin dictionaries (hereafter L&S and WW) as sources, and synsets assigned through a ranking system of glosses.⁷

⁶<https://latinwordnet.exeter.ac.uk/>

⁷**L&S:** https://github.com/PerseusDL/lexica/tree/master/CTS_XML_TEI/perseus/pdlex/lat/ls; **WW:** <https://github.com/mk270/whitakers-words>; **Freedict.com:** <https://www.freedict.com/onldict/lat.html>

3 The Latin WordNet

The LWN was first created in 2004 following the *Expand Method* (Vossen, 2002, p. 52), that is, by automatically translating portions of the aligned Italian and English (PWN) data contained in MultiWordNet (hereafter MWN_I and MWN_E) into Latin with the help of bilingual dictionaries (Latin to English mostly from Glare (1982) via WW; Latin to Italian mostly from Pianezzola et al. (2001)). The LWN comprises 9,378 lemmas distributed across 8,973 synsets (Minozzi, 2017): 5,621 synsets are nominal (denoted by the initial n# in the ID), 2,283 verbal (v#), 775 adjectival (a#) and 294 adverbial (r#). Additionally, it provides two files of synset relations: one containing 13,771 language-dependent lemma-to-lemma relations, the other 4,588 synset-to-synset relations common to MWN (see Table 1).

<i>latin_relation.sql</i> (lemma-to-lemma)		
type	n	%
Antonymy	4,538	32.95%
Pertainymy/Derivation*	9,233	67.04%
<i>common_relation.sql</i> (synset-to-synset)		
type	n	%
Hyper/hyponymy	3,900	85.00%
Meronymy, part of	292	6.36%
Entailment (v)	90	1.96%
Attribute (n)	80	1.74%
Value of (a)	80	1.74%
Similar to (a)	54	1.17%
Cause (v)	34	0.74%
Meronymy, substance of	32	0.69%
Meronymy, member of	26	0.56%

Table 1: The distribution of lemma and synset relations across the LWN. *The *Pertainymy/Derivation* relation between lemmas is not well defined in the LWN documentation.

The criteria behind the selection of LWN lemmas remain unclear, and there are some noticeable gaps, both lexical (*amo*, *amare* ‘to love’) and relational (the adjectives *inaequabilis* ‘unequal’ and *aequabilis* ‘equal’ are placed in a relation of derivation only but could also count as antonyms). Examples of erroneous, modern senses inherited by the LWN from MWN_E are shown in Table 2. In point of fact, in his most recent publication, the creator of LWN states that the lexical coverage and the results of his automatic assignments need further evaluation and verification (Minozzi, 2017, p. 130).

lemma	synset_id	definition
ager	n#W0021124	in un database, ogni area in cui vengono registrate le singole informazioni che compongono il record [...]
capitolium	n#06188340	the federal government of the United States
genetrix	n#W0021113	titolo e appellativo che si dà alle suore professe o a quelle che hanno cariche particolari; sono venuta a fare atto d'obbedienza alla madre badessa di questo convento
voco	v#00720710	send a message or attempt to reach someone by radio, phone, etc; make a signal to in order to transmit a message [...]

Table 2: Synsets to be removed from LWN.

4 Evaluation method

For a close understanding of the implications of evaluating a WordNet, we formulated a first experiment combining a small, automated extension of the sense coverage of the LWN with a follow-up manual revision of their corresponding synsets. The purpose of this experiment was to measure the reliability and feasibility of these two approaches in order to identify the most effective compromise for LiLa.

Data. Firstly, we formatted LWN and all necessary Machine Readable Dictionaries for the task as relational SQL tables: these included WW, L&S, MWNE and MWNI.

Machine-recommended senses. Next, inspired by the work of Abouenour et al. (2013), we formalised a rudimentary algorithm in bash script to automatically extend the sense coverage of the LWN by proposing new synsets taken from the MWNE. While aware that this method would introduce some noise, the neither exact nor approximate amounts could not be quantified *a priori*. Figure 1 exemplifies the algorithmic process: for the LWN adverb *velociter* ‘swiftly, quickly’, the algorithm 1) searched for joint lemma and PoS overlaps between LWN and WW; 2) where there was a match, it then looked for overlaps between the single-word WW glosses and MWNE lemmas; 3) where these also matched, it checked the lemma’s corresponding synset(s) in MWNE for that PoS against existing LWN synsets to 4) label machine recommendations as NEW (machine-suggested and not already present in LWN) or COM

(for “common”, i.e., machine-suggested but already present in LWN). Table 3 lists the results of the recommender system for *velociter*.

synset_id	definition	label
r#00051957	in a swift manner; she moved swiftly	NEW
r#00082992	with rapid movements; he works quickly	COM
r#00102338	with little or no delay; [...]	COM
r#00285860	without taking pains; [...]	COM

Table 3: Synset assignments for the adverb *velociter* to be evaluated by human raters.

The recommender system produced 121,098 lemma-synset entries for the whole LWN: 93,479 synset assignments (77.19%) were classified as NEW, 25,613 (21.15%) as COM and 2,006 (1.65%) as OLD (synsets present in the LWN only). Given the algorithm’s optimisation on recall, we expected these large numbers to include many false positives and homography, e.g., the verbs *edo*, *edere* ‘to eat’ (3rd conjugation) and *edo*, *edare* ‘to publish’ (1st conjugation) or *volo*, *velle* ‘to want’ (irregular conjugation) and *volo*, *volare* ‘to fly’ (1st conjugation).

Lemma selection. Next, for our test evaluation, we randomly selected 100 LWN-WW matched lemmas, 25 per PoS, featuring both NEW and COM synset assignments. This selection resulted in 3,746 lemma-synset entries to be evaluated.⁸

Manual evaluation. Of the five raters recruited for the task, four were in possession of intermediate Latin proficiency and one had expert (including spoken) knowledge of the language.⁹ Using a custom web annotation environment designed to facilitate the task and with Latin dictionaries at hand (Campanini and Carboni, 1993; Castiglioni and Mariotti, 1966 1979 1996 2007; Bianchi et al., 1972), raters were instructed to approve or reject synset assignments.

Unsurprisingly, our synset recommender generated irrelevant assignments, as shown in Table 4.

The evaluation was performed over a period of approximately two months and informed the formulation of guidelines to enforce consistency. Among other directives, the guidelines demanded that raters accept an assignment even if specific

⁸Of the 100 selected lemmas, 36 had multiple homographic entries with the same PoS.

⁹Those with intermediate Latin knowledge were pursuing a Master’s degree in Theoretical and Applied Linguistics, while the expert rater completed a Master’s in Modern Philology (“Lettere” and Semantics).

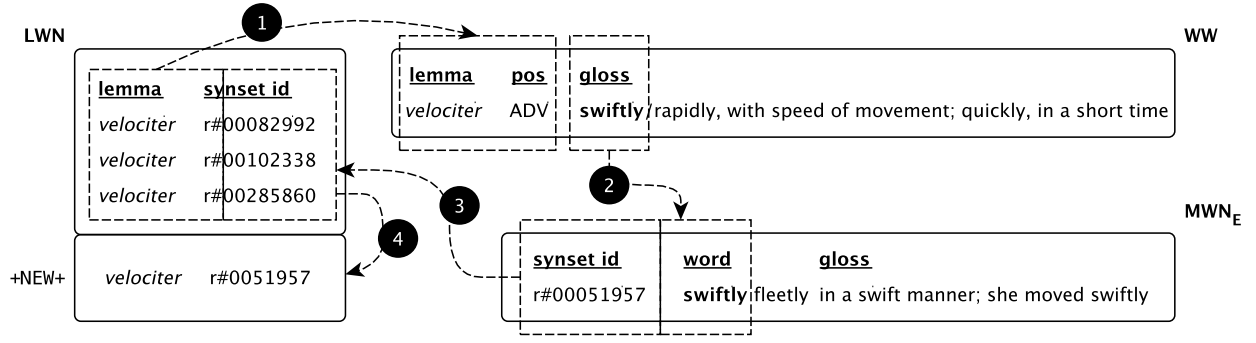


Figure 1: The algorithmic process of synset assignment. Here, a new MWNE synset is added to *velociter*.

lemma	synset_id	definition
albus	a#01549077	used to signify the Confederate forces in the Civil War (who wore gray uniforms); a stalwart gray figure
caput	n#02805750	a toilet on board a boat of ship
contentus	a#00760259	slang for ‘drunk’
deprehendo	v#00733757	be the catcher, in baseball; Who is catching?
tonus	n#00319371	an all-fours game in which the first card led is a trump

Table 4: Machine-proposed synsets to be discarded from LWN.

to an idiomatic use of the lemma (e.g., *edo*, *edere* ‘to eat/consume/devour’ but *edere voces* ‘utter’); accept an assignment even if its specificity is not mirrored in the reference dictionary (e.g., while the specific sense ‘to sodomize’ for *caedo* is not explicitly mentioned in Castiglioni and Mariotti (2007), the verb is said to have sexual connotations as well);¹⁰ reject an assignment if the corresponding sense is not included in their reference dictionary; and reject an assignment should there be any other strong uncertainty not covered by the guidelines. The assessment of the relations, if any, *between* OLD synset assignments in our evaluation set was ignored at this stage.

Missing senses. Where applicable, raters were also instructed to make a note of missing senses, be those from the Classical, Medieval or Late periods of Latin.¹¹ Inclusion of these missing senses in the LWN is not described here but is planned future work (see Section 6). Examples are:

¹⁰IV ed., s.v., “*caedo*,” Def. fig. “in senso osceno, *sbatte*re, Catull. 56, 7 e a.”

¹¹We do not consider contemporary Latin (19th and 20th centuries).

prudenter (r): skillfully;

puto, *putare* (v): to clean; to prune, trim

radix (n): radish; liquorice

tener (a): erotic, amorous; adaptable (style); soft (soil)

Inter-rater reliability agreement. Next, we measured inter-rater reliability (IRR) using percentage agreement without chance correction (McHugh, 2012). Percentage agreement was chosen over Fleiss Kappa (Fleiss, 1971) because the evaluation was performed in a controlled setting with low chances of guessing on a binary yes/no rating. We thus applied the following formula:

$$A_o(r) = \frac{abs(N_A(r) - N_R(r))}{N_V(r)}$$

where the observed agreement A_o on each lemma-synset relation (r) is calculated by dividing the absolute difference of accepted N_A and rejected N_R assignments by the total number of evaluations N_V . Agreement values range between 0.0 and 1.0, where 0.0 means no inter-rater agreement and 1.0 means perfect inter-rater agreement.

5 Results and discussion

In this section we assess IRR agreement rates against the table proposed by McHugh (2012, p. 279). As previously observed in related studies, lower agreements are not a reflection of raters’ inability to distinguish word meanings but, rather, of their difficulty in selecting the synsets that best fit their subjective opinion (Hajič et al., 2004, p. 28). Table 5 provides minimum (m_v), maximum (M_v) and average values of agreement (A_v) per type of synset assignment as well as standard deviations (S_v). The A_v values all fall within the *strong* tier of McHugh agreement (64-81%, corresponding to a square k agreement of .80-.90), but reveal that almost $\frac{1}{3}$ of all synsets was not reliably

rated.

type	n	m_v	M_v	A_v	S_v
OLD	35	0.200	1.000	0.691	0.345
COM	876	0.200	1.000	0.654	0.320
NEW	2,835	0.200	1.000	0.702	0.329

Table 5: Inter-rater agreement values grouped by type of synset assignment.

IRR agreement is a measure of both actual agreement but also of disagreement among raters. So, for a better understanding of the *quality* of both Minozzi’s and our own synset assignment, we calculated the acceptance rates of OLD, COM and NEW assignments. As Table 6 shows, the acceptance rates on all three types of assignment is very low, with an average 77% of all assignments being rejected by all raters and a tenuous average of 0.02% of unanimous acceptance. These results are particularly worrying for OLD and COM assignments, as they give us a first indication of the quality, and hence usability, of LWN.

type	n	Acceptance in %					
		0_r	1_r	2_r	3_r	4_r	5_r
OLD	35	65.7	14.2	5.7	2.8	11.4	0.0
COM	876	79.4	8.6	5.0	3.4	2.2	0.6
NEW	2,835	87.2	6.7	3.4	1.5	1.0	0.0

Table 6: Acceptance rates of synset types per number of raters (N_r).

As far as Part of Speech (PoS) is concerned, the most prolific syntactic category in terms of machine-proposed synset assignments were verbs, followed by nouns, adjectives and adverbs. Table 7 shows IRR agreement rates per PoS relative to the number of synset assignments; \bar{x} indicates the average or arithmetic mean of synsets per lemma (25 in total) per category. Nouns and verbs fared the best, with *strong* agreement on large percentages of assignments (84% and 60%, respectively); adjectives, on the other hand, appear to have been more challenging, as the percentages of assignments on which the raters moderately and strongly agreed are roughly the same (44% and 48%, respectively). Finally, against our expectations, despite the comparatively lower number of synsets, 48% of adverbial assignments were met with moderate agreement. Low agreement values might be caused by incorrect assignments (as was the case of the NEW assignment ‘with honesty; he was rightly considered the greatest singer of his time’

to *proprie*, approved by only one rater) or, more problematically, differences of opinion on subtle semantic differences. A close examination of the data, and, specifically, of the adverbs with agreement values below 60% (6 out of 25), points to the latter. A clear example is given by the adverb *brevisiter*, whose lowest rated assignment ‘with rapid movements; he works quickly’ (COM) was approved by two raters only. The adverb’s primary sense is ‘shortly, in a brief space of time’, and while ‘rapidly’ might, in some cases, reasonably be equated to ‘shortly’, three raters discarded the assignment as the senses conveyed by the terms ‘rapid’ and ‘quickly’ are better expressed by the Latin adverb *celeriter*. Similarly, in the case of *subtus* ‘below, underneath, in a lower position, beneath’, the NEW assignment ‘at a later place; see below’ was also rejected by three raters, despite it being a potentially valid sense. It is worth noting that in these and other arguable cases, synsets carrying temporal meanings tend to show lower agreement rates than those associated with space (i.e., ‘rapidly’ and ‘later’ are temporal equivalents of ‘short’ and ‘below’). The higher agreement rate on the spatial dimension resonates with cognitive linguistic theories on spatial semantics, according to which “Space is at the heart of all conceptualization” (Pütz and Dirven, 1996, xi), as its concreteness over temporal or more abstract meanings induces us to map its structure onto other semantic domains (Lakoff and Johnson, 1980; Lakoff, 1987). The validity of these theories in the context of LWN evaluation remains to be verified.

type	\bar{x} syn/lemma	IRR agreement in %		
		moderate	strong	\approx perfect
VERB	51.32	32	60	8
NOUN	46.56	0	84	16
ADJ	42.04	44	48	8
ADV	8.84	48	28	20

Table 7: IRR agreement rates per PoS relative to the number of synsets.

6 Conclusion and future work

This paper describes a preliminary assessment of the implications of evaluating the LWN carried out in the context of the *LiLa: Linking Latin* project. The objective of LiLa is to connect linguistic resources and NLP tools for Latin with a view to supporting different lines of linguistic and corpus-based research and to connecting Latin to other

languages. Owing to its automatic process of creation, the LWN is lexically and semantically limited, as well as noisy, subjecting its inclusion in LiLa to qualitative revision. For a close understanding of the implications of evaluating the LWN, we formulated a first experiment combining a small, automated extension of the sense coverage on the basis of 100 selected LWN lemmas with a follow-up manual revision of their corresponding synset assignments. The purpose of this experiment was to measure the reliability and feasibility of these two approaches in order to identify the most effective evaluation compromise.

Our synset recommender system produced many false positives, with only 0.18% (7) machine suggestions approved by all five raters. Even if the precision of the synset-recommendation algorithm were to be improved, recall would likely still be high due to the unavoidable assignment of modern senses to a historical resource. If applied to the entire LWN, the evaluation method described here, coupled with the additional evaluation of the relations between synsets, would turn this process of revision into an unsustainable effort or, at the very least, one that is not achievable within the scope and duration of LiLa.¹²

Moving forward, our plan for the improvement of the LWN will develop into various tasks. The first, ongoing effort is the manual removal of the modern senses originally inherited by the LWN. Next, once cleaned, we will extend the sense coverage of the LWN by manually adding the missing senses recorded by the raters for the 100 evaluated lemmas, careful not to introduce too much granularity (i.e., too many senses with only subtle semantic differences); extract hypernyms, synonyms and bags of words from dictionary definitions (Nadig et al., 2008), as well as lemma groups from three Latin synonym dictionaries: the Latin-English *Hand-book of Latin Synonymes* (Döderlein et al., 1875), the Latin-English *The synonymes of the Latin language* (Hill, 1804) and the Latin-Czech *Latinská synonymika pro školu i dům* (Skřivan, 1890).¹³ These are all freely available online in XML dictionary format (XDXF) and, combined, can supply the LWN with some

1,050 additional lemmas.

Thirdly, connect a graph version of the LWN to textual resources in LiLa to acquire lexical knowledge, and explore the possibility of extracting hypernym/hyponym pairs using syntactic patterns (Snow et al., 2004). Finally, extend the LWN with Named Entities extracted from the morphological analyser LEMLAT (Budassi and Passarotti, 2016).

The data and code repository for this paper are available at: <https://github.com/CIRCSE/latinWordnet-evaluation>

Acknowledgments

This project has received funding from the European Research Council (ERC) under the European Union’s Horizon 2020 research and innovation programme Grant Agreement No. 769994. The authors also wish to thank William Michael Short of the University of Exeter for the productive discussions leading up to this article, and the anonymous reviewers for their helpful suggestions.

¹²In an unlikely scenario of uninterrupted evaluation, our method applied to the entire LWN would indicatively require 64.65 months to complete.

¹³Available from: <https://nikita-moor.github.io/dictionaries/dictionaries.html>

References

- Lahsen Abouenour, Karim Bouzoubaa, and Paolo Rosso. 2013. On the evaluation and improvement of Arabic WordNet coverage and usability. *Language Resources and Evaluation*, 47(3):891–917. <https://doi.org/10.1007/s10579-013-9237-0>.
- Enrico Bianchi, Raffaello Bianchi, and Onorio Lelli. 1972. *Dizionario illustrato della lingua latina*. Le Monnier, Firenze.
- Yuri Bizzoni, Federico Boschetti, Riccardo Del Gratta, Harry Diakoff, Monica Monachini, and Gregory Crane. 2014. The making of Ancient Greek WordNet. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the 9th edition of the Language Resources and Evaluation Conference, 26-31 May, Reykjavik, Iceland*, pages 318–325. http://www.lrec-conf.org/proceedings/lrec2014/pdf/1071_Paper.pdf.
- Olivier Bodenreider, Anita Burgun, and Joyce A. Mitchell. 2003. Evaluation of wordnet as a source of lay knowledge for molecular biology and genetic diseases: a feasibility study. *Studies in health technology and informatics*, 95:379–384. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1893008/>.
- Marco Budassi and Marco Passarotti. 2016. Nomen Omen. Enhancing the Latin Morphological Analyser Lemlat with an Onomasticon. In *Proceedings of the 10th SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pages 90–94. ACL. <http://www.aclweb.org/anthology/W16-2110>.
- Giuseppe Campanini and Giuseppe Carboni. 1993. *Nuovo Campanini-Carboni vocabolario latino-italiano italiano-latino, con appendice suddivisa in 11 glossari*. Paravia, Torino.
- Luigi Castiglioni and Scevola Mariotti. 1966, 1979, 1996, 2007. *IL Vocabolario della Lingua Latina,, Latino - Italiano . Italiano - Latino*. Loescher, Torino.
- Christian Chiarcos, John McCrae, Philipp Cimiano, and Christiane Fellbaum, 2013. *Towards Open Data for Linguistics: Linguistic Linked Data*, pages 7–25. Springer Berlin Heidelberg, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-31782-8_2.
- L. Von Döderlein, S. H. Taylor, and H. H. Arnold. 1875. *Döderlein's Hand-book of Latin Synonymes*. Warren F. Draper. <https://archive.org/details/dderleinshandbo00arnogooq>.
- Christiane Fellbaum. 1998. Towards a Representation of Idioms in WordNet. In *Proceedings of the workshop on the Use of WordNet in Natural Language Processing Systems (Coling-ACL)*, pages 52–57. <http://ai2-s2-pdfs.s3.amazonaws.com/280b/cd6c4f1e3b9f9abb32a0c510614e5128d9df.pdf>.
- Joseph L. Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76:378–382. <http://dx.doi.org/10.1037/h0031619>.
- Greta Franzini, Marco Passarotti, Maria Moritz, and Marco Böhler. 2018. Using and evaluating TRACER for an Index fontium computatus of the Summa contra Gentiles of Thomas Aquinas. In *Proceedings of the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018), 10-12 December, Torino, Italy*. <http://ceur-ws.org/Vol-2253/paper22.pdf>.
- P. G. W. Glare. 1982. *Oxford Latin Dictionary*. Oxford University Press, Oxford.
- Jan Hajič, Martin Holub, Marie Hučínová, Martin Pavlík, Pavel Pecina, Pavel Straňák, and Pavel Šidák. 2004. Validating and Improving the Czech WordNet via Lexico-Semantic Annotation of the Prague Dependency Treebank. In *Proceedings of LREC 2004*, pages 25–30. https://www.researchgate.net/publication/237100339_Validating_and_Improving_the_Czech_WordNet_via_Lexico-Semantic_Annotation_of_the_Prague_Dependency_Treebank.
- F. R. S. E. John Hill. 1804. *The synonyms of the Latin language*. Edinburgh. <https://archive.org/details/synonymesoflatin00hilluoft>.
- George Lakoff and Mark Johnson. 1980. *Metaphors We Live By*. The University of Chicago Press, Chicago.
- George Lakoff. 1987. *Women, fire, and dangerous things: What categories reveal about the mind*. The University of Chicago Press, Chicago.
- Charlton T. Lewis and Charles Short. 1879. *Harpers' Latin Dictionary: A New Latin Dictionary Founded on the Translation of Freund's Latin-German Lexicon Edited by E. A. Andrews*. Harper and Brothers, New York.
- Mary L. McHugh. 2012. Interrater reliability: the kappa statistic. *Biochemia Medica*, 22:276–282. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3900052/>.
- George A. Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine J. Miller. 1990. Introduction to WordNet: An On-line Lexical Database. *International Journal of Lexicography*, 3(4):235–244. <https://academic.oup.com/ijl/article/3/4/235/923280>.

- Stefano Minozzi. 2017. Latin WordNet, una rete di conoscenza semantica per il latino e alcune ipotesi di utilizzo nel campo dell'Information Retrieval. In Paolo Mastandrea, editor, *Strumenti digitali e collaborativi per le Scienze dell'Antichità*, number 14 in *Antichistica*, pages 123–134. <http://doi.org/10.14277/6969-182-9/ANT-14-10>.
- Raghuvar Nadig, J. Ramanand, and Pushpak Bhattacharyya. 2008. Automatic Evaluation of Wordnet Synonyms and Hypernyms. In *Proceedings of the Sixth International Conference on Natural Language Processing (ICON-2008)*. <https://www.cse.iitb.ac.in/~pb/papers/icon08-wn-validation.pdf>.
- Marco Passarotti, Flavio M. Cecchini, Greta Franzini, Eleonora Litta, Francesco Mambrini, and Paolo Ruffolo. 2019. LiLa: Linking Latin A Knowledge Base of Linguistic Resources and NLP Tools. In Thierry Declerck, editor, *Proceedings of the 2nd Conference on Language, Data and Knowledge (LDK 2019)*, 20-23 May, Leipzig, Germany. <https://doi.org/10.5281/zenodo.3358550>.
- Emilio Pianezzola, Giuliano Ranucci, and Gian Biagio Conte. 2001. *Il dizionario della lingua latina*. Edmund Le Monnier.
- Martin Pütz and Renè Dirven, editors. 1996. *The construal of space in language and thought*. Mouton de Gruyter, Berlin.
- A. Skřivan. 1890. *Latinská synonymika pro školu i dům*. Chrudim. <https://archive.org/details/SkivanLatinskSynonymika>.
- Rion Snow, Daniel Jurafsky, and Andrew Y. Ng. 2004. Learning syntactic patterns for automatic hypernym discovery. *Advances in neural information processing systems*, 17:1297–1304. http://ai.stanford.edu/~rion/papers/hypernym_nips05.pdf.
- Piek Vossen. 2002. EuroWordNet General Document. <http://dare.ubvu.vu.nl/bitstream/handle/1871/11116/EWNG?sequence=1>.

Iride[®]: an Industrial Perspective on Production Grade End To End Dialog System

Cristina Giannone, Valentina Bellomaria, Andrea Favalli and Raniero Romagnoli

Language Technology Lab
Almawave srl
[first name initial].[last name]@almawave.it

Abstract

This paper aims at describing, from an industrial perspective, the experience in delivering conversational agents via the development of Iride, a platform able to deploy multi-language task-oriented dialog systems. It has been implemented a set of functionalities that can be aggregated in different ways, in order to build domain independent conversational systems, which are able to satisfy needs of real business cases. Along with algorithms and techniques for end to end Dialog management, such as Natural Language Understanding (NLU), Question Answering (QA) and Dialog State tracking and policy management, the technical insights leveraged into the platform are described by outlining the requirements and constraints emerging from these on the field experiences.¹

1 Introduction

Over the last years the human computer conversation has been gathering increasing attention due to its promising potentials by opening up a new profits-making market segment.² The benefits of using dialog systems are manifold, these systems can answer to complex questions and also handle hundreds, thousands of conversations at the same time, reducing response times and probability of error in repetitive tasks. In General, developing conversational agents at industrial level requires to manage several issues: (i) The lack of real data: in the majority of the real business

cases, in our experience, not enough data are available for training pure learning methods, moreover, the research datasets do not fit the industrial purposes; (ii) Domain updates and system maintenance: The domain requires continuous updates (e.g. the introduction of a new product or service) and the delivered system needs the maintenance, update or changes to correct faults and to improve performance; (iii) User Experience: the conversational agent is the front end of the company, multi-modality (i.e. different user experiences depending on different devices) and what the company aims at communicating must be taken into account; (iv) Runtime latency: is required to add no more than few mini seconds to the entire serving stack; (v) Scale and quality of the text collection: in a voice interaction the system cannot answer with a long text document, but needs to answer with a clear short document passage; (vi) Certified Answers: Being the virtual assistant the voice of the company, it must be controlled (i.e. usually the answers and the messages communicated by the assistant have to be certified by the company); (vii) Human in the loop: Although virtual assistants are becoming more and more intelligent, they are not able to satisfy every user need. In this scenario, it would be better a mixed management, combining the use of virtual agent and human operator.

In this paper, we describe the Almawave's developed solution that allows us to quickly design, write and deploy interactive conversational systems without coding, enabling non-technical users (i.e. conversational designers or domain experts) to design conversational agents, and it leverages Natural Learning Processing (NLP) and Machine Learning (ML) to develop a human-like experience for users. This framework is designed to build multi-turn task-oriented dialog able to solve defined tasks and answer to domain questions.

Following, in section 2 related works will be

¹Copyright © 2019 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

²<https://www.gartner.com/smarterwithgartner/4-trends-gartner-hype-cycle-customer-service-customer-engagement/>

discussed: in section 3, the various goals that have leaded the described solution will be discussed; in section 4 the various modules of the architecture will be fully described; finally, in section 5, we formulate some considerations and lessons learned in the conversational agent field.

2 Background

Due to the complexity of task, most studies on Human Machine conversation have addressed individual components such as Intent/Slots detection (Coucke et al., 2018) or Dialog State tracking (Mrksic et al., 2015) about frameworks for building an effective dialog system. Recent works in the end-to-end frameworks are focused on the pure learning approaches, where the sequence of dialog interactions, between the user and the agent, is acquired from large datasets (Wu et al., 2017), (Wen et al., 2017), as well as in the dialog task oriented field (Bordes and Weston, 2016). Although Neural Networks provided a significant improvement in the NLP field, in the conversational agent field, NN end-to-end systems have some limitations, all their components are directly trained on past dialogues, with no assumption on the domain or dialog state structure, thus training with large scale human-human dialog data is required. However, these resources are generally not so easily available for building an end-to-end system. Some works based on NN address on limit the amount of training data: the framework proposed in (Bocklisch, 2017) focused on quickly helping implement machine learning-based dialog management and natural language understanding, the work implements a function to generate, from the input dataset, new data and provided a special function called a story graph that visualize the flow of dialog scenarios in advance. In (Lipton et al., 2017) a deep reinforcement learning algorithm is proposed to tackle a domain extension setting, where new slots can gradually be introduced. On the other hand, in (Lison, 2015), the authors proposed a framework for expressing dialog behaviors as probabilistic rules. The probabilistic rules used in this study consist of conditional statements and actions with probability; these can be made manually or automatically generated by supervised learning or reinforcement learning. Following (Yan et al., 2017), our proposal is toward a platform for the development of a conversational agent able to perform a cold-start with no dialog

training data. Other close works address on the building of frameworks in order to allow the development of conversational agents in several scenarios and domains, in (Crook et al., 2016) is proposed a task configuration language, i.e *TaskForm*, which allows to decouple the conversation management issues with the definitions of the target task, and moreover make available a large set of ML algorithms for the NLU tasks. In a recently proposed platform (Sungjin Lee and Gao, 2019), the issue of evaluating the end-to-end conversational agent is approached.

3 Goals

From our experience, the main objectives of a dialog system for business needs are the usability and the robustness. The system must always be functioning in time and satisfy user needs by operating as few interactions as possible. The conversational platform here proposed was developed with some characteristics concerning those objectives:

3.1 Usability

The usability principles for this kind of framework look to user designers. The main issues to pursue the usability goal are described in the following:

Focus on conversation design

Designing a dialog conversation must take into account both what has already been said and what will happen next; it is much more complex than one-off activities, like answering a search query, playing a song and so on. In relation to this, new professions are emerging, such as the Voice User Interface (VUI) designer who curates the conversation, defining the flow and its underlying logic in a detailed design specification that represents the complete user experience, playing an important role from the conceptual phases of the project until its launch (Urban and Mailey, 2019).

However, these profiles are not necessarily developers or data scientist, so it is very important that tools offer to them all the available technology but are easy to use, so that the designer can focus on aspects more related to domain and policies of dialog management. A solution we delivered to solve this problem is a Visual Dialog Editor, hiding the complexity of programming AI components, allowing the user to construct a dialog agent with a visual building block approach, the drawn flow is thus compiled producing the dialog agent

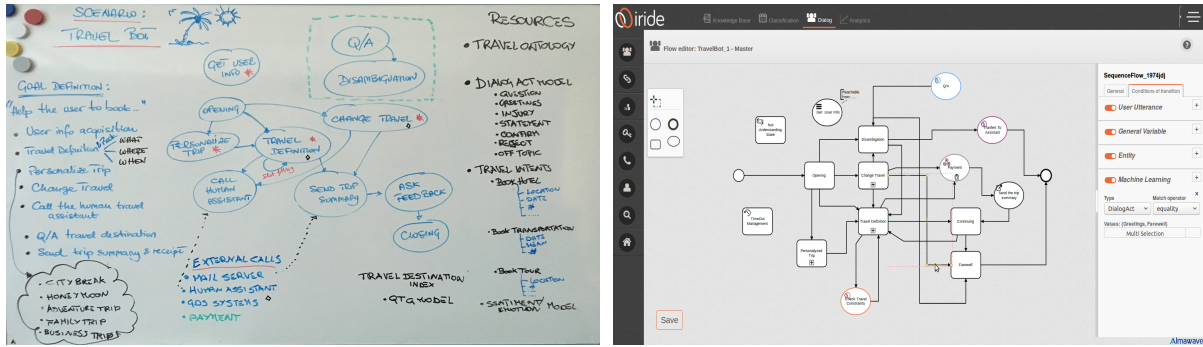


Figure 1: A conversation design process, from conversational map to dialog model in the Iride Conversational Platform

software. In Figure 1 an example of a conversational design process is shown, from a conversational map that highlights the important items to the dialog model drawn with the Dialog Editor.

Component Based

A Component Based approach in a SW architecture lead to quality products, rapid development and an increased ability to adapt to change. In contrast to use of end-to-end conversational model that concentrates all the interaction features and capabilities within a monolithic model as black-box, a modular approach allows the potential of system engineering to be exploited for complexity management. An important aspect we focused was to maximize the re-usability of the platform components, such as algorithms or trained models as well as the dialog flows, within conversational agents for different domains, tasks and languages, maintaining a domain and task independent environment. To pursue it, the framework makes various components and algorithms available, in order to have a different level offer views. There are components dedicated to knowledge management, others that realize language understanding, dialog management and multi-modality connection. Even a single module can be seen as the set of sub-modules that realize more specific functionalities.

3.2 Robustness

In a commercial solution the robustness of a system must be guaranteed, and it can be achieved by a combination of different strategies. A significant effort was made in the system to detect and handle a wide range of errors, ranging from the language understanding, the discourse processing and the domain reasoning. But, whatever input understanding strategy is adopted, managing every

possible user input is difficult, therefore the platform provides different solutions to improve the reliability managing both not understood and mis-understood inputs.

4 The Conversational Platform Overview

This section describes the overall structure of the platform. In order to pursue the main goals we defined this architecture. It is the result of collaborative effort between working on the different technologies and where the different components can be assembled to produce multiple applications.

The components are described dividing them in 3 logical views: The Design Tools for the conversational agent design, the Dialog Core Modules that implements the underline engine dialog components and, in order to provide analysis over the conversations, an Analytics Module.

4.1 Design Tools

Visual Dialog Editor

Modeling a dialog means defining the flow of the conversation and its underlying logic.

Designers define the behaviours of the agent, defining the dialog script in terms of States, Transitions and Actions. The visual editor facilitates the modeling of the flow of dialog, drawing the transitions between the dialog states and actions using graphical approach, and enable the use of the various types of resources.

Moreover, the editor, provides a graphical interface to the resource management (e.g. ontologies, models, indexes).

Simulator

A conversation simulation environment is provided within the editor for the dialog assessment. This tool enables the testing by the designer and

confirms the correctness of the dialog before deployment.

Through the simulator it is possible to verify some relevant aspects in the realization of virtual assistant. Observing the flow of conversation makes it possible to assess the smoothness and naturalness of the discourse, in relation to the management of waiting times and turn-taking. The simulator also helps to evaluate and balance the use of graphic components such as images, buttons and quick replies, usually used to make the interaction easier. It is also important to explore the error management to put in way out and recovery policies.

4.2 Dialog Core Modules

Knowledge Representation

Designers use knowledge representation to build the operational structure of the dialog agent. The concepts of the domain and their relationships are represented by ontologies, taxonomies and dictionaries. If we could develop a dialog agent in a new domain with a rich ontological structure, re-using the knowledge of the existing domain becomes fundamental. The separation of domain knowledge also reduces the complexity of the linguistic components, using both general purpose resources and domain specific ones. Within the conversational platform different types, i.e., dictionaries, ontologies, inference rules, indexes and machine learning models, of knowledge representation are used in combination in order to obtain flexible dialog and dialog agent configurable.

Language Understanding

The platform makes available a proprietary multi-lingual NLP pipeline, composed by several modules that enable language comprehension, providing the language analysis at several levels ranging from morphological to pragmatic and task-dependent analysis.

This pipeline allows an hybrid approach, rule-based and machine learning, depending on needs, that can be both used and combined together, exploiting, for example, the outcomes of DL classification into ontological reasoning. Among the several modules, the following Deep Learning models are leveraged:

- A sentence classification model built over pre-trained language models (Devlin et al., 2018) used for several tasks such as Dialog

Act Classification (Stolcke et al., 2000) or Question Classification (Li and Roth, 2002);

- A Sequence classification models, for NER task (Chen et al., 2018);
- Intent detection and slot filling jointly classification (Castellucci et al., 2019);
- A sentiment analysis NN model, described in (Bonadiman et al., 2017)

The chosen models benefit from the advantages of the transfer learning techniques (Tan et al., 2018) in order to reduce the amount of required training data. Although this approach provides a relevant advantage in reducing the annotation effort, it might be useful to choose, according to the scenario, the right approach between "good old-fashioned techniques" and deep learning approaches.

The framework allows the use of domain dictionaries, ontologies and inferential rules that enable the extraction and inference of semantic concepts. Our framework gains the benefits of each approach by simultaneously applying the rule-based and machine learning approaches combining both techniques to infer complex knowledge structures. It is worth mentioning that with the platform is released a tool that allows, in a simple way, even to non-technical users the training of specific models to customize a system on a given domain.

Dialog Management

The dialog manager (DM) is the core component of the platform. At each turn in the conversation, the dialog management component takes the current dialog state and the user utterance as its inputs, performs different actions based on context, and outputs corresponding results as responses. DM includes two stages: dialog state tracking and dialog policy. The dialog state comprises all that is used when the system makes its decision about what is the next agent action; in this scenario, **the dialog state tracker** updates the context based on the result of the analysis of the last received input, e.g. NLU analysis over the user utterance or the query response of an external knowledge base.

In the proposed approach, the dialog tracking is implemented over hand written probabilistic rules in line with (Lison, 2015), (Wang and Lemon, 2013). The designer draws the flow of interaction as edge transitions between dialog elements

(actions and states) and adding weights for each transition. The resulting transition edges from two states cannot be not mutually exclusive, hence, at time t the tracked state of the dialog, consisting of a representation of the conversation history, the input analysis and the more "weighted" state connected to the previous one. **The dialog policy** generates dialog actions based on the current dialog information state. The system utterances depend on the current action/state, i.e. answers can be randomly selected from a defined list (in a state) or obtained as result of the selected action, as in the QA module. This approach enables a 'cold-start' when past conversation data are not available and the dialog has to be designed from scratch. The tracked state is passed on to the dialog policy module to select the best next action to perform the objective task. A set of predefined and easily customizable actions are available for the dialog design, the platform uses a plug-in mechanism, for each agent the required elements are plugged into the solution. Some of them are:

- **Question Answering:** The Question Answering action follows two steps: it performs a retrieval process over a domain dependent index. The retrieved answers are re-ranked applying NN for learning to Re-Rank process as in the CQA task (Nakov et al., 2016) in line with (Nassif et al., 2016). Moreover the QA action implements clarification strategies in case of ambiguous results.
- **External System Call:** Rest APIs are available for integration with external systems. The conversational designer can graphically draw this action fulfill few input data (e.g. endpoint, authentication and request data)
- **Slot Filling complexio:** The agent engages with the user a set of interactions to fulfill the values of a specified list of entities, e.g. the slot list of an intent or the properties of an ontological concept.
- **Route to Operator:** Under specific conditions, the dialog session can be redirect to a human operator giving to him the visibility of the information acquired up to that time. This action manages specific business cases ensuring robustness and service continuity.

Multimodality

There are different ways of communication and the choice of the users depends on various factors. The platform makes available connectors to different communication channels ranging from social network to legacy systems. The conversational agents can be delivered both through voice and written chat. Moreover, the change of channel is available (e.g. route the chat to operator or vice versa) in order to respond to specific business cases managing the change transparently to the user. Moreover, the conversations based on the different channels, can be equipped with UI components such as images, buttons and quick replies.

4.3 Analytics Module

The analysis of the conversations provides a constant view of how the conversational agent plays the "voice of the company" role. The analytics module allows to extract several insights from the dialog: interaction satisfaction, dialog errors as well as analytics for CX analysis. This one, in addition to provide market information, collects data for the agent maintenance and updates.

5 Industrial Consideration and Conclusion

In this paper we described the experience in building the Iride conversational platform for the design and deployment of task-oriented conversational agents in enterprise environment. The platform has been built taking into account needs and constraints required by an industrial scenario. We focused on a component based architecture able to maximize the re-usability of the components, enforcing a clear separation between the domain-specific aspects of the dialog and domain-independent ones across the several dialog layers (language understanding, dialog management and knowledge management). Moreover, in order to enable the work of conversational designer, the platform offers a suite of tools for conversational designers. Such architectural choices have been verified testing "on the field" the effectiveness and usability of the described solution.

Several conversational agents have been developed with this framework, in different business cases and in different domains and languages; these experiences demonstrate that the platform is efficient and easy-to-use and meets the needs of various types of use cases.

References

- J.; Pawlowski N.; Nichol A. Bocklisch, T.; Faulker. 2017. Rasa: Open source language understanding and dialogue management. *arXiv*.
- Daniele Bonadiman, Giuseppe Castellucci, Andrea Favalli, Raniero Romagnoli, and Alessandro Moschitti. 2017. Neural sentiment analysis for a real-world application. *CLiC-it 2017 11-12 December 2017, Rome*, page 42.
- Antoine Bordes and Jason Weston. 2016. Learning end-to-end goal-oriented dialog. *CoRR*.
- Giuseppe Castellucci, Valentina Bellomaria, Andrea Favalli, and Raniero Romagnoli. 2019. Multilingual intent detection and slot filling in a joint bert-based model. *CoRR (To Appear)*, abs/1907.XXX.
- Lingzhen Chen, Alessandro Moschitti, Giuseppe Castellucci, Andrea Favalli, and Raniero Romagnoli. 2018. Transfer learning for industrial applications of named entity recognition. 12.
- Alice Coucke, Alaa Saade, Adrien Ball, Théodore Bluche, Alexandre Caulier, David Leroy, Clément Doumouro, Thibault Gisselbrecht, Francesco Caltagirone, Thibaut Lavril, Maël Primet, and Joseph Dureau. 2018. Snips voice platform: an embedded spoken language understanding system for private-by-design voice interfaces. *CoRR*, abs/1805.10190.
- Paul A. Crook, Alex Marin, V. Agarwal, K. Agarwal, T. Anastasakos, R. Bikkula, D. Boies, Asli Celikyilmaz, S. Chandramohan, Z. Feizollahi, R. Holenstein, M. Jeong, Omar Zia Khan, Young-Bum Kim, E. Krawczyk, X. Liu, D. Panic, V. Radostev, N. Ramesh, J.-P. Robichaud, A. Rochette, L. Stromberg, and Ruhi Sarikaya. 2016. Task completion platform: A self-serve multi-domain goal oriented dialogue platform. *ACL - Association for Computational Linguistics*, June.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
- Xin Li and Dan Roth. 2002. Learning question classifiers. In *Proceedings of the 19th International Conference on Computational Linguistics - Volume 1, COLING '02*, pages 1–7, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Zachary Lipton, Xiujuan Li, Jianfeng Gao, Lihong Li, Faisal Ahmed, and li Deng. 2017. Bbq-networks: Efficient exploration in deep reinforcement learning for task-oriented dialogue systems. 11.
- Pierre Lison. 2015. A hybrid approach to dialogue management based on probabilistic rules. *Comput. Speech Lang.*, 34(1):232–255, November.
- Nikola Mrksic, Diarmuid Ó Séaghdha, Blaise Thomson, Milica Gasic, Pei-hao Su, David Vandyke, Tsung-Hsien Wen, and Steve J. Young. 2015. Multi-domain dialog state tracking using recurrent neural networks. *CoRR*, abs/1506.07190.
- Preslav Nakov, Lluís Màrquez, Alessandro Moschitti, Walid Magdy, Hamdy Mubarak, Abed Alhakim Freihat, Jim Glass, and Bilal Randeree. 2016. SemEval-2016 task 3: Community question answering. In *Proceedings of the 10th International Workshop on Semantic Evaluation, SemEval '16*, San Diego, California, June. Association for Computational Linguistics.
- Henry Nassif, Mitra Mohtarami, and James Glass. 2016. Learning semantic relatedness in community question answering using neural models. In *Proceedings of the 1st Workshop on Representation Learning for NLP*, pages 137–147, Berlin, Germany, August. Association for Computational Linguistics.
- Andreas Stolcke, Noah Coccaro, Rebecca Bates, Paul Taylor, Carol Van Ess-Dykema, Klaus Ries, Elizabeth Shriberg, Daniel Jurafsky, Rachel Martin, and Marie Meteer. 2000. Dialogue act modeling for automatic tagging and recognition of conversational speech. *Comput. Linguist.*, 26(3):339–373, September.
- Ryuichi Takanobu Xiang Li Yaoqin Zhang Zheng Zhang Jinchao Li Baolin Peng Xiujuan Li Minlie Huang Sungjin Lee, Qi Zhu and Jianfeng Gao. 2019. Convlab: Multi-domain end-to-end dialog system platform.
- Chuanqi Tan, Fuchun Sun, Tao Kong, Wenchang Zhang, Chao Yang, and Chunfang Liu. 2018. A survey on deep transfer learning. *CoRR*, abs/1808.01974.
- Margaret Urban and Stephen Mailey. 2019. Conversation design: Principles, strategies, and practical application. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems, CHI EA '19*, pages C26:1–C26:3, New York, NY, USA. ACM.
- Zhuoran Wang and Oliver Lemon. 2013. A simple and generic belief tracking mechanism for the dialogue state tracking challenge: On the believability of observed information. In *Proceedings of SIGDIAL 2013*. Association for Computational Linguistics, 8.
- Tsung-Hsien Wen, David Vandyke, Nikola Mrkšić, Milica Gasic, Lina M. Rojas Barahona, Pei-Hao Su, Stefan Ultes, and Steve Young. 2017. A network-based end-to-end trainable task-oriented dialogue system. In *EACL*, pages 438–449, Valencia, Spain, April. Association for Computational Linguistics.
- Yu Wu, Wei Wu, Chen Xing, Ming Zhou, and Zhoujun Li. 2017. Sequential matching network: A new architecture for multi-turn response selection in retrieval-based chatbots. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages

496–505, Vancouver, Canada, July. Association for Computational Linguistics.

Zhao Yan, Nan Duan, Peng Chen, Ming Zhou, Jian-she Zhou, and Zhoujun Li. 2017. Building task-oriented dialogue systems for online shopping.

When Lexicon-Grammar Meets Open Information Extraction: a Computational Experiment for Italian Sentences

Raffaele Guarasci, Emanuele Damiano, Aniello Minutolo, Massimo Esposito

National Research Council of Italy

Institute for High Performance Computing and Networking (ICAR), Naples, Italy

{name.surname}@icar.cnr.it

Abstract

In this work we show an experiment on building an Open Information Extraction system (OIE) for Italian language. We propose a system wholly reliant on linguistic structures and on a small set of verbal behavior patterns defined putting together theoretical linguistic knowledge and corpus-based statistical information¹. Starting from elementary one-verb sentences, the system identifies elementary tuples and then, all their permutations, preserving the overall well-formedness (grammaticality) and trying to preserve semantic coherence (acceptability). Although the work focuses only on the Italian language, it can be proficiently extended also to other languages, since it is essentially based only on linguistic resources and on a representative corpus for the language under consideration².

1 Introduction

One of the most interesting approach to handle the rapid growth of textual data emerged in the last decade is Open Information Extraction (OIE). Starting from natural language sentences, it allows to extract one or more domain-independent propositions, scaling to the diversity and size of the corpus considered (Banko et al., 2007). Each extracted proposition is represented by a verb and its arguments, i.e. “Maria goes to the party” is a proposition with a relation (the verb *goes*) that links together two arguments (*Maria*, *the party*). Arguments (nouns or noun groups) can have different roles (subject, direct object...) and they can

be mandatory or optional. In this sentence, both arguments *Maria* (subject) and *the party* (direct object) are mandatory, so it is impossible to remove one of them or the sentence becomes unacceptable from a grammatical point of view. Due to the high field of Natural Language Processing (NLP) tasks in which OIE outputs can be used (Christensen et al., 2013; Fader et al., 2014; Stanovsky et al., 2015; 2016; Khot et al., 2017; Rahat et al., 2017), numerous OIE approaches for English have been developed. However, being a language-dependent task, OIE systems cannot be shifted from one language to another, i.e. a system created for English is not compatible with Italian. Moreover, many of the proposed OIE approaches rest on unstable grounds. Some of them use heuristics to manage large quantities of textual data, others lack the support of a theoretical basis, outlining the natural language in a reductive way. Differently from the vast majority of existing OIE approaches, we propose a linguistic-based unsupervised system designed to extract n-ary propositions (not only “relation-argument” triples) from natural language sentences in Italian, ensuring domain independence and scalability.

Our system aims to identify the elementary tuple(s) from the input sentence, then all its (their) permutations, by adding progressively arguments composing the sentence. After that – according the behavior patterns of the verb – it generates every possible syntactically valid n-ary proposition, granting grammaticality.

To reach this result we have combined two types of resources. To gather information about verb behavior in sentences, we grounded our work on the linguistic basis provided by Lexicon Grammar (LG) (Gross, 1994). In order to obtain a fine-grained characterization of arguments, we

¹ An online demo showing some features of the system is freely available at the address <https://nlpit.na.icar.cnr.it/>

² Copyright © 2019 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

combine this theoretical knowledge with distributional corpus-based information extracted from it-WaC (Baroni et al., 2009). From LG tables we extract patterns of verbs behaviors, and from it-WaC we enrich these patterns with statistical information. Using complex linguistic structures and dependency parse trees (DPT) we can detect verbal behavior patterns occurring in one-verb sentences and generate from them all the possible well-formed propositions, by adding complements and adverbials. The use of formal patterns derived from a theoretical framework allows to better distinguish between necessary verbal arguments and optional removable adjuncts and to verify syntactic restrictions in verb possible structures.

Arguments optionality and syntactic constraints are critical features to grant the grammaticality of the propositions generated, also trying to approximate a first level of semantic acceptability.

2 Related Work

In the last years, several approaches to OIE has been developed (Banko et al., 2007; Zhu et al., 2009; Wu et al., 2010; Fader et al., 2011; Schmitz et al., 2012; Del Corro et al., 2013), all of them with the characteristic of utilizing a set of patterns in order to obtain propositions, granting scalability and portability across different domains.

They differ in many aspects such as performances (precision, recall, speed); linguistic structures used (Part-of-Speech tags, chunks, DPT); patterns to extract information (hand-crafted based on heuristics or learned from a training corpus); type of generated output (binary extractions, n-ary extractions, nested extractions).

However, most of these existing approaches so far has been focused on English, with only some recent attempts that have appeared for other languages, such as Spanish (Zhila et al, 2013), Chinese (Wang et al, 2014), Vietnamese (Truong et al., 2017), German (Falke et al., 2016; Bassa et al., 2018) and Romance languages (Gamallo et al., 2012; Gamallo et al., 2015). As far as we know only one approach has been attempted for the Italian (Damiano et al., 2018). It is a preliminary experiment based on a limited set of patterns and heuristics, and experimented on a hand-crafted dataset of reduced size.

3 Lexicon-Grammar

As the theoretical basis for our system we decided to use LG since it regards the systematic formalization of a very broad quantity of data for the Italian language (Elia et al., 1981; D’Agostino, 1992). Other resources describing a subset of Italian verbs have been developed, such as LexIt (Lenci et al. 2012), MultiWordNet (Pianta et al. 2002), SensoComune (Oltramari et al. 2013) and T-PAS (Jezek et al., 2014). However, none of them provides a formal classification of verbs in classes or clusters. Conversely, LG groups verbs in classes according to their behavior, specifying for each verb its essential arguments and possible syntactic structures in order to create well-formed sentences (Leclère, 2002).

3.1 How data are structured in LG

LG classes are represented in the form of tables. Each row of the table corresponds to a verb of the class, each column lists all properties that may be valid or not for the different members of the class. At the intersection of a row and a column, the symbol + or – may indicate that the property corresponding to the column is valid or not for the verb corresponding to the row, as shown in Table 1³, which reports some Italian verbs and their properties as encoded in a LG. Properties can be of different types. They can refer to the syntactic structure and the prepositions admitted by that specific verb, semantic restrictions (e.g. human/non-human argument) or possible transformations (e.g. passive form). For the purpose of this work, only syntactic properties will be considered. This choice reflects the syntactic nature of OIE, which focuses on shapes and structures of verbs.

Verb	N ₀ VN ₁	N ₀ V	N ₀ VprepN ₁	N ₀ VN ₁ prepN ₂
Mangiare (to eat)	+	+	-	-
Muovere (to move)	+	-	-	+
Girare (to turn)	+	+	+	+

Table 1 Example of an LG table

The first column contains the defining property, which corresponds to the basic syntactic structure

³ The formal notation used in LG is summarized as follows: N indicates a nominal group and is followed by a progressive subscript indicating its nature (N₀ is the subject, N₁ is the first complement, N₂ is the second complement, etc.), V represents the verb, prep indicates prepositions.

of the elementary sentence. The property expressed in the second column is a syntactic property called deletion (Harris, 1982), labeled as N_0V , which allows the cancellation of the element N_1 from the basic syntactic structure specified with the defining property. Deleting the element N_1 on the right of the verb is valid for the verb “mangiare” (“Max mangia”, *Max eats*), while it produces ungrammatical unacceptable sentences for the verb “muovere” (“*Max muove”, **Max moves*). Prep represents a set of every possible adjuncts placed before every argument N_i .

3.2 From tables to patterns

Despite the richness of this fine-grained information, LG tables suffer from some limitations that have made them useless in real NLP applications: they are verbose and properties is neither uniform nor standardized. Therefore, many changes were necessary to be able to use these resources in the OIE system:

Grouping. We divided verbs into classes: direct (D) without preposition, indirect with a preposition (I), and locative (L). This distinction is preferred to the classical distinction between transitive and intransitive verbs, since locative verbs can accept both transitive and intransitive construction. Verbs assuming a copulative function (support verbs) form a further class (S). For the purpose of this work, we do not consider complement-clause verbs, because of the variability of the structures possible for the definition of unique patterns.

Enrichment: Prep element is too coarse. We need to specify which kind of preposition the selected verb admits. To overcome this limit, we add a syntactic profile to each verb, containing the most frequent prepositions associated to it. We extract this information from itWaC corpus.

Formal representation. To reduce redundant information of the original tables we formalize a grammar to compactly represent verbs behavior, indicating selection preferences on the possible arguments of a verb. Square brackets [] represent the possibility of deleting arguments, round brackets () indicates there are many possible arguments separated by a vertical bar, and XOR symbol \oplus represents the exclusive alternativity of patterns.

As it is shown table 2, the notation $N_0V[N_1]$ indicates that the verb “mangiare” (*to eat*) can accept both the structures N_0VN_1 or N_0V , and the notation $N_0V(in|a)N_1$ denotes that the verb can accept

alternatively and also simultaneously both the patterns N_0VinN_1 and N_0VaN_1 . On the other hand, a notation like $N_0VN_1 \oplus N_0VinN_1$ denotes that the verb can accept exclusively only one between the patterns N_0VN_1 and N_0VinN_1 , even if they are both valid from a grammatical perspective. This is due to the fact that their selection preferences are representative of different verb usages and, thus, are alternative and exclusive from a semantic perspective. Note that in the table 2 possible prepositions are reduced for a better readability of the pattern.

Verbs	Patterns
mangiare (<i>to eat</i>)	$N_0V[N_1]$
muovere (<i>to move</i>)	$N_0VN_1 \oplus N_0V(in<in> da<from> verso<toward>)N_1$
girare (<i>to turn</i>)	$N_0V(a<to> intorno<around>)N_1 \oplus N_0VN_1[(a<to> da<from> verso<toward>)N_2]$

Table 2 Patterns derived from LG tables

4 Proposed Approach

Our approach for OIE is arranged in the form of a multi-step pipeline and it consists into 4 steps:

Sentence Processing: every input sentence is checked to verify that it is suitable for the approach.

Arguments Identification: arguments of the verb are identified (i.e. subjects, direct complements, indirect complements...).

Pattern Recognition: verbal structures that match the patterns are identified and elementary tuples made by the combination of arguments are generated.

Proposition Generation: n-ary propositions depending on the elementary tuples and the remaining arguments (i.e. adverbs, complements and modifiers) are generated.

As an example, for the sentence “Da domani Anna andrà da Roma a Milano” (*From tomorrow Anna will go from Rome to Milan*), both the tuples and corresponding propositions that are generated are reported in Table 3.

The verb “andare” (*to go*) belongs to locative group loc, and its complete pattern is the following $N_0V[daN_1](a|in|verso|su|sopra)N_2$. In the first column of the table identified patterns for the verb are reported, the second column lists tuples and propositions generated from every single pattern.

Pattern	Generations
N ₀ VaN ₁	1. ("Anna"<Anna>, "andrà"<will go>, "Milano"<Milan>) Anna andare a Milano (<i>Anna to go to Milan</i>)
	2. ("Domani"<tomorrow>, "Anna"<Anna>, "andrà"<will go>, "Milano"<Milan>) Da domani Anna andare a Milano (<i>From Tomorrow Anna to go to Milan</i>)
	3. ("Anna"<Anna>, "andrà"<will go>, "Roma"<Rome>, "Milano"<Milan>) Anna andare da Roma a Milano (<i>Anna to go from Rome to Milan</i>)
N ₀ daVaN ₁	4. ("Domani"<tomorrow>, "Anna"<Anna>, "andrà"<will go>, "Roma"<Rome>, "Milano"<Milan>) Da domani Anna andare da Roma a Milano (<i>From Tomorrow Anna to go from Rome to Milan</i>)

Table 3 tuples and propositions generated from an input sentence

5 Experiment and validation

We carried out the evaluation using quantitative metrics well known in NLP literature: precision and recall. Precision measures the average on all the sentences of the percentage of extractions obtained by the proposed approach that are correct, whereas recall measures the average on all the sentences of the percentage of extractions manually annotated in the dataset that are correctly identified by the proposed approach. Performances was evaluated on a dataset of sentences containing verbs belonging to different classes, and the validation took place with respect to grammaticality and acceptability (i.e. syntactic well-formedness of the sentences and its meaningfulness in the context) using the gold standard proposed in (Guarasci et al. *in press*). Notice that grammaticality and acceptability judgements is a much debated topic in theoretical and computational linguistics in the past (Phillips, 2009; Phillips, 2011; Gibson et al., 2010) and still today it is considered a controversial subject (Lau et al., 2017; Sprouse et al., 2018). Even if OIE is a syntactic task, so it focus on the structure of the sentence, but not its meaning (Lau et al., 2017), we aim to generate sentences not only well-formed but also respecting some syntactic constraints and selection preferences, trying to approximate the first level of semantic acceptability.

Sentences		Grammaticality		Acceptability	
		P	R	P	R
Total verbs	195	0.91	0.78	0.79	0.84
Locative	62	0.93	0.73	0.77	0.83
Direct	30	0.90	0.93	0.79	0.93
Indirect	65	0.88	0.81	0.78	0.83
Support	38	0.98	0.66	0.86	0.78

Table 4 results for different verb classes

Table 4 shows precision (P) and recall (R) scores with respect to the two criteria on the verbs divide by classes.

Precision and recall achieve high values with respect to both grammaticality and acceptability. More precisely, with respect to the different structures of verbs considered, precision has resulted sensibly higher for sentences containing support verbs with respect to grammaticality and acceptability. This behavior is reversed for recall, which has resulted for sentences containing direct, indirect or locative verbs.

5.1 Comparison with other OIE systems

Globally, generations per sentences and performances achieved are comparable with state-of-the-art OIE systems in other languages, respectively ClausIE (English) and GerIE (German). Moreover, we compare our results with the only other experiment conducted on Italian presented by the authors and named ItalIE (Damiano et al, 2018).

Sentences		Grammaticality		Acceptability	
		P	R	P	R
Total verbs	195	0.84	0.40	0.73	0.43
Locative	62	0.91	0.46	0.74	0.51
Direct	30	0.82	0.56	0.74	0.57
Indirect	65	0.72	0.27	0.68	0.57
Support	38	0.91	0.36	0.86	0.45

Table 5 Performances of ItalIE

As shown in Tables 5, our approach has reached the best overall performances in terms precision and recall for both grammaticality and acceptability. ItalIE highlighted a sensibly lower number of generations (511 vs 918 of our approach) with a moderate decrease in precision but a significant reduction in recall. This behavior can be explained by the fact that ItalIE is based on a fixed set of clause patterns not considering the extreme variability of verb behaviors and also the selection preferences on their possible arguments. Furthermore, its algorithm based on DPT to identify constituents through dependency relations has shown some weaknesses. It fails in detecting and properly handling named entities, multi-word expressions, adjectives, numerals, dates and some patterns related to support verbs.

5.2 Error Analysis

The number of both false positives and negatives generated in the experiments is shown in Table 6

with respect to grammaticality (G) and acceptability (A).

	False positives					False negatives		
	DP	NE	SC	MC	Tot	DP	VU	Tot
G	78	3	0	0	81	145	86	231
A	78	3	76	38	195	114	21	135

Table 6 Summary of the errors generating false positives and negatives with respect to grammaticality and acceptability.

Various types of errors are divided as follows:

DP: errors caused by incorrect dependency parsing due to wrong and/or missing dependencies between element occurring in the input sentence. They represent the vast majority of the errors affecting overall performances of the proposed approach. With respect to grammaticality and acceptability, false positives have been generated by DP errors in 96% and 40% of cases, whereas false negatives are due to DP errors in 63% and 84% of cases, respectively.

NE: error in the identification of named-entities. NE errors have occurred in a not significant number of cases, only 3, generating false positives with respect to both grammaticality and acceptability.

VU: behavior patterns not associated to the verb usage selected for the input sentence. It represents the second source of errors causing false negatives with respect to grammaticality and acceptability (in 37% and 16% of cases, respectively).

MC: missing morpho-syntactic concordance among different parts-of-speech or missing contractions or combinations between prepositions and articles. It causes 19% of false positives in acceptability.

SC: violated semantic constraints. It affects only acceptability, causing 39% of false positives. Notice that this error is referred only to the semantic perspective, while others are related to grammatical aspects.

6 Conclusions and Future Work

In this work we have shown an experiment to perform OIE for Italian language, extracting n-ary propositions from natural language sentences, granting well-formedness of the generations. The system relies on a linguistic resource (LG) and on a representative corpus for Italian (itWaC). While these resources are specific to Italian, they also exist for other languages, so the system can be easily extended. In particular, LG tables exist in digital format also for French (Tolone, 2012),

English (Garcia-Vega, 2010; Machonis, 2010), Portuguese (Baptista, 2001), Romanian (Ciocanea, 2011). Likewise, the itWaC corpus used in this work is part of the WaCky Wide Web corpora collection (Baroni et al., 2009), which includes corpora of English (ukWaC), German (deWaC), French (frWaC). Concerning performances of the system, although the results are encouraging, we are looking forward to further developments.

With regard to methodological progress, we plan to integrate novel methods based on deep learning to increase the performance of the system, trying to reduce DP errors and better handle named entities, frozen and semi-frozen bigrams and multi-word expressions. From an applicative perspective, this work will be experimented in Italian Question Answering system, with the goal to improve the ability in reading complex texts and extracting the correct answers to users' questions. Other possible outcomes can include text summarization or other NLP tasks.

References

- Michele Banko, Michael J Cafarella, Stephen Soderland, Matthew Broadhead, and Oren Etzioni. 2007. Open information extraction from the web. In *Proceeding of IJCAI*, vol. 7, pp. 2670-2676.
- Jorge Baptista. 2012. Viper: A lexicon-grammar of european portuguese verbs. In *31e Colloque International sur le Lexique et la Grammaire*.
- Marco Baroni, Silvia Bernardini, Adriano Ferraresi, and Eros Zanchetta. 2009. The WaCky wide web: a collection of very large linguistically processed web-crawled corpora. *Language Resources & Evaluation*, 43(3):209–226, September.
- Akim Bassa, Mark Kröll, and Roman Kern. 2018. GerIE-An Open Information Extraction System for the German Language. *Journal of Universal Computer Science*, 24(1):2–24.
- Janara Christensen, Stephen Soderland, and Oren Etzioni. 2013. Towards Coherent Multi-Document Summarization. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp 1163–1173
- Cristiana Ciocanea. 2011. *Lexique-grammaire des constructions converses en a da/ a primi en roumain. (Lexicon-grammar of converse constructions in a da/ a primi in Romanian)*. PhD Thesis, University of Paris-Est, France.
- Emilio D’Agostino. 1992. *Analisi del discorso: metodi descrittivi dell’italiano d’uso*. Loffredo.
- Emanuele Damiano, Aniello Minutolo, and Massimo Esposito. 2018. Open Information Extraction for

- Italian Sentences. In *Proceedings of 2018 32nd International Conference on Advanced Information Networking and Applications Workshops*, pp. 668-673.
- Luciano Del Corro and Rainer Gemulla. 2013. ClauseIE: clause-based open information extraction. In *Proceedings of the 22nd International Conference on World Wide Web*, pp. 355-366.
- Annibale Elia, Maurizio Martinelli, and Emilio d'Agostino. 1981. *Lessico e strutture sintattiche: introduzione alla sintassi del verbo italiano*. Liguori, Napoli.
- Oren Etzioni, Anthony Fader, Janara Christensen and Stephen Soderland. 2011. Open Information Extraction: The Second Generation. In *Proceeding of IJCAI*, vol. 11, pp. 3-10.
- Anthony Fader, Stephen Soderland, and Oren Etzioni. 2011. Identifying Relations for Open Information Extraction. In *EMNLP '11*, pages 1535-1545, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Anthony Fader, Luke Zettlemoyer, and Oren Etzioni. 2014. Open question answering over curated and extracted knowledge bases. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 1156-1165.
- Pablo Gamallo, Marcos Garcia, and Santiago Fernández-Lanza. 2012. Dependency-based Open Information Extraction. In *ROBUS-UNSUP '12*, pages 10-18, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Pablo Gamallo, Marcos Garcia. 2015. Multilingual open information extraction. In *Portuguese Conference on Artificial Intelligence*, pp. 711-722.
- Edward Gibson and Evelina Fedorenko. 2013. The need for quantitative methods in syntax and semantics research. *Language and Cognitive Processes*, 28(1-2):88-124.
- Maurice Gross. 1994. *Constructing lexicon-grammars*. Centre national de la recherche scientifique, Universités de Paris 7 et 8.
- Zellig Sabbettai Harris. 1982. *A grammar of English on mathematical principles*. John Wiley & Sons Incorporated.
- Elisabetta Jezek, Bernardo Magnini, Anna Feltracco, Alessia Bianchini, and Octavian Popescu. T-PAS: A resource of corpus-derived Typed Predicate Argument Structures for linguistic analysis and semantic processing. In *Proceedings of LREC*, pp. 890-895.
- Tushar Khot, Ashish Sabharwal, and Peter Clark. 2017. Answering complex questions using open information extraction. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, (2)pp. 311-316.
- Jey Han Lau, Alexander Clark, and Shalom Lappin. 2017. Grammaticality, Acceptability, and Probability: A Probabilistic View of Linguistic Knowledge. *Cognitive Science*, (41): 5, pp. 1202-1241.
- Christian Leclère. 2005. The Lexicon-Grammar of French Verbs. In *Linguistic Informatics State of the Art and the Future: The first international conference on Linguistic Informatics*, (1)pp. 29-45.
- Christian Leclère. 2002. Organization of the lexicon-grammar of French verbs. *Linguisticae Investigationes*, 25(1):29-48, January.
- Alessandro Lenci, Gabriella Lapesa, and Giulia Bonansinga. LexIt: A Computational Resource on Italian Argument Structure. In *LREC*, pp. 3712-3718.
- Alessandro Oltramari, Guido Vetere, Maurizio Lenzerini, Aldo Gangemi, and Nicola Guarino. 2010. Senso Comune. In *LREC* pp. 3873-3877.
- Colin Phillips. 2009. Should we impeach armchair linguists. *Japanese/Korean Linguistics*, 17:49-64.
- Colin Phillips. 2013. Some arguments and nonarguments for reductionist accounts of syntactic phenomena. *Language and Cognitive Processes*, 28(1-2):156-187.
- Emanuele Pianta, Luisa Bentivogli, and Christian Girardi 2002. Developing an aligned multilingual database. In *Proceedings of Global WordNet Conference*.
- Michael Schmitz, Robert Bart, Stephen Soderland, and Oren Etzioni. 2012. Open Language Learning for Information Extraction. In *EMNLP-CoNLL '12*, pages 523-534, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Jon Sprouse, Beracah Yankama, Sagar Indurkha, Sandiway Fong, and Robert C Berwick. 2018. Colorless green ideas do sleep furiously: gradient acceptability and the nature of the grammar. *The Linguistic Review*, 35(3):575-599.
- Gabriel Stanovsky and Ido Dagan. 2015. Open IE as an Intermediate Structure for Semantic Tasks. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 2300-2305.
- Gabriel Stanovsky and Ido Dagan. 2016. Creating a large benchmark for open information extraction. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2300-2305.
- Elsa Tolone. 2012. Analyse syntaxique à l'aide des tables du Lexique-Grammaire du français. *Linguisticae Investigationes*, 35(1):147-151.

- Diem Truong, Duc-Then Vo, Uyen Trang Nguyen. 2017. Vietnamese Open Information Extraction. In *Proceedings of the Eighth International Symposium on Information and Communication Technology*, pp. 135-142.
- Mingyin Wang, Lei Li, and Fang Huang. 2014. Semi-supervised chinese open entity relation extraction. In *2014 IEEE 3rd International Conference on Cloud Computing and Intelligence Systems*, pages 415–420.
- Fei Wu and Daniel S Weld. 2010. Open Information Extraction using Wikipedia. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, pp. 118–127.
- Alisa Zhila and Alexander Gelbukh. 2013. Comparison of open information extraction for English and Spanish. *Computational Linguistics and Intelligent Technologies*, 12(19):714–722.
- Jun Zhu, Zaiqing Nie, Xiaojiang Liu, Bo Zhang, and Ji-Rong Wen. 2009. StatSnowball: a statistical approach to extracting entity relationships. In *Proceedings of the 18th international conference on World wide web*, pages 101–110. ACM.

Are Subtitling Corpora really Subtitle-like?

Alina Karakanta^{1,2}, Matteo Negri¹, Marco Turchi¹

¹ Fondazione Bruno Kessler, Via Sommarive 18, Povo, Trento - Italy

² University of Trento, Italy

{akarakanta, negri, turchi}@fbk.eu

Abstract

Growing needs in translating multimedia content have resulted in Neural Machine Translation (NMT) gradually becoming an established practice in the field of subtitling. Contrary to text translation, subtitling is subject to spatial and temporal constraints, which greatly increase the post-processing effort required to restore the NMT output to a proper subtitle format. In this work, we explore whether existing subtitling corpora conform to the constraints of: 1) length and reading speed; and 2) proper line breaks. We show that the process of creating parallel sentence alignments removes important time and line break information and propose practices for creating resources for subtitling-oriented NMT faithful to the subtitle format.

1 Introduction

Machine Translation (MT) of subtitles is a growing need for various applications, given the amounts of online multimedia content becoming available daily. Subtitling translation is a complex process consisting of several stages (transcription, translation, timing), and manual approaches to the task are laborious and costly. Subtitling has to conform to spatial constraints such as length, and temporal constraints such as reading speed. While length and reading speed can be modelled as a post-processing step in an MT workflow using simple rules, subtitle segmentation, i.e. where and if to insert a line break, depends on semantic and syntactic properties. Subtitle segmentation is particularly important, since it has been shown that a

proper segmentation by phrase or sentence significantly reduces reading time and improves comprehension (Perego, 2008; Rajendran et al., 2013).

Hence, there is ample room for developing fully or at least partially automated solutions for subtitle-oriented NMT, which would contribute in reducing post-processing effort and speeding-up turn-around times. Automated approaches though, especially NMT, are data-hungry. Performance greatly depends on the availability of large amounts of high-quality data (up to tens of millions of parallel sentences), specifically tailored for the task. In the case of subtitle-oriented NMT, this implies having access to large subtitle training corpora. This leads to the following question: **What should data specifically tailored for subtitling-oriented NMT look like?**

There are large amounts of available parallel data extracted from subtitles (Lison and Tiedemann, 2016; Pryzant et al., 2018; Di Gangi et al., 2019). These corpora are usually obtained by collecting files in a subtitle specific format (.srt) in several languages and then parsing and aligning them at sentence level. MT training at sentence level generally increases performance as the system receives longer context (useful, for instance, to disambiguate words). As shown in Table 1, this process compromises the subtitle format by converting the subtitle blocks into full sentences. With this “merging”, information about subtitle segmentation (line breaks) is often lost. Therefore, recovery of the MT output to a proper subtitle format has to be performed subsequently, either as a post-editing process or by using hand-crafted rules and boundary predictions. Integrating the subtitle constraints in the model can help reduce the post-processing effort, especially in cases where the input is a stream of data, such as in end-to-end Speech Neural Machine Translation. To date, there has been no study examining the consequences of obtaining parallel sentences from sub-

Copyright © 2019 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

1	00:00:14,820 -- > 00:00:18,820
	Grazie mille, Chris.
	É un grande onore venire
2	00:00:18,820 -- > 00:00:22,820
	su questo palco due volte.
	Vi sono estremamente grato.
<hr/>	
	Grazie mille, Chris.
	É un grande onore venire su questo palco due volte.
	Vi sono estremamente grato.

Table 1: Subtitle blocks (top, 1-2) as they appear in an .srt file and the processed output for obtaining aligned sentences (bottom).

titles on preserving the subtitling constraints.

In this work, we explore whether the large, publicly available parallel data compiled from subtitles conform to the temporal and spatial constraints necessary for achieving quality subtitles. We compare the existing resources to an adaptation of MuST-C (Di Gangi et al., 2019), where the data is kept as subtitles. For evaluating length and reading speed, we employ character counts, while for proper line breaks we use the Chink-Chunk algorithm (Lieberman and Church, 1992). Based on the analysis, we discuss limitations of the existing data and present a preliminary road-map towards creating resources for training subtitling-oriented NMT faithful to the subtitling format.

2 Related work

2.1 Subtitling corpora

Building an end-to-end subtitle-oriented translation system poses several challenges, mainly related to the fact that NMT training needs large amounts of high-quality data representative of the target application scenario (subtitling in our case). Human subtitlers translate either directly from the audio/video or they are provided with a template with the source text already in the format of subtitles containing time codes and line breaks, which they have to adhere to when translating.

Several projects have attempted to collect parallel subtitling corpora. The most well-known one is the OpenSubtitles¹ corpus (Lison and Tiedemann, 2016), extracted from 3.7 million subtitles across 60 languages. Since subtitle blocks do not always correspond to sentences (see Table 1), the blocks are merged and then segmented into sentences us-

ing heuristics based on time codes and punctuation. Then, the extracted sentences are aligned to create parallel corpora with the time-overlap algorithm (Tiedemann, 2008) and bilingual dictionaries. The 2018 version of OpenSubtitles has high-quality sentence alignments, however, it does not resemble the realistic subtitling scenario described above, since time and line break information are lost in the merging process. The same methodology was used for compiling MontenegrinSubs (Božović et al., 2018), an English – Montenegrin parallel corpus of subtitles, which contains only 68k sentences.

The Japanese-English Subtitle Corpus JESC (Pryzant et al., 2018) is a large parallel subtitling corpus consisting of 2.8 million sentences. It was created by crawling the internet for film and TV subtitles and aligning their captions with improved document and caption alignment algorithms. This corpus is aligned at caption level, therefore its format is closer to our scenario. On the other hand, non-matching alignments are discarded, which might hurt the integrity of the subtitling documents. As we will show, this is particularly important for learning proper line breaks between subtitle blocks.

A corpus preserving both subtitle segmentation and order of lines is SubCo (Martínez and Vela, 2016), a corpus of machine and human translated subtitles for English–German. However, it only consists of 2 source texts (~150 captions each) with multiple student and machine translations. Therefore, it is not sufficient for training MT systems, although it could be useful for evaluation because of the multiple reference translations.

Slightly deviating from the domain of films and TV series, corpora for Spoken Language Translation (SLT) have been created based on TED talks. The Web Inventory of Transcribed and Translated Talks (Cettolo et al., 2012) is a multilingual collection of transcriptions and translations of TED talks. The talks are aligned at sentence level without audio information. Based on WIT, the IWSLT campaigns (Niehues et al., 2018) are annually releasing parallel data and their corresponding audio for the task of SLT, which are extracted based on time codes but again with merging operations to create segments. MuST-C (Di Gangi et al., 2019) is to date the largest multilingual corpus for end-to-end speech translation. It contains (audio-source language transcription-target

¹<http://www.opensubtitles.org/>

language translation) triplets, aligned at segment level. The process of creation is the opposite from IWSLT; the authors first align the written parts and then match the audio. This is a promising corpus for an end-to-end system which translates from audio directly into subtitles. However, the translations are merged to create sentences, therefore they are far from the suitable subtitle format. Given the challenges discussed above, there exists no systematic study of the suitability of the existing corpora for subtitling-oriented NMT.

2.2 Subtitle segmentation

Subtitle segmentation techniques have so far focused on monolingual subtitle data. Álvarez et al. (2014) trained Support Vector Machine and Logistic Regression classifiers on correctly/incorrectly segmented subtitles to predict line breaks. Extending this work, Álvarez et al. (2017) used a Conditional Random Field (CRF) classifier for the same task, also differentiating between line breaks (next subtitle line) and subtitle breaks (next subtitle block). Recently, Song et al. (2019) employed a Long-Short Term Memory Network (LSTM) to predict the position of the period in order to improve the readability of automatically generated Youtube captions. To our knowledge to date, there is no approach attempting to learn bilingual subtitle segmentation or incorporating subtitle segmentation in an end-to-end NMT system.

3 Criteria for assessing subtitle quality

3.1 Background

The quality of the translated subtitles is not evaluated only in terms of fluency and adequacy, but also based on their format. We assess whether the available subtitle corpora conform to the constraints of length, reading speed (for the corpora where time information is available) and proper line breaks on the basis of the criteria for subtitle segmentation mentioned in the literature of Audiovisual Translation (AVT) (Cintas and Remael, 2007) and the TED talk subtitling guidelines²:

1. **Characters per line.** The space available for a subtitle is limited. The length of a subtitle depends on different factors, such as size of screen, font, age of the audience and country. For our analysis, we consider max. 42 chars for Latin alphabets, 14 for Japanese (including spaces).

2. **Lines per subtitle.** Subtitles should not take up too much space on screen. The space allowed for a subtitle is about 20% of screen space. Therefore, a subtitle block should not exceed 2 lines.

3. **Reading speed.** The on-air time of a subtitle should be sufficient for the audience to read and process its content. The subtitle should match as much as possible the start and the end of an utterance. The duration of the utterance (measured either in seconds or in feet/frames) is directly equivalent to the space a subtitle should occupy. As a general rule, we consider max. 21 chars/second.

4. **Preserve ‘linguistic wholes’.** This criterion is related to subtitle segmentation. Subtitle segmentation does not rely only on the allowed length, but should respect linguistic norms. To facilitate readability, subtitle splits should not “break” semantic and syntactic units. In an ideal case, every subtitle line (or at least subtitle block) should represent a coherent linguistic chunk (*i.e.* a sentence or a phrase). For example, a noun should not be separated from its article. Lastly, subtitles should respect natural pauses.

5. **Equal length of lines.** Another criterion for splitting subtitles relates to aesthetics. There is no consensus about whether the top line should be longer or shorter, however, it has been shown that subtitle lines of equal length are easier to read, because the viewer’s eyes return to the same point on the screen when reading the second line.

While subtitle length and reading speed are factors that can be controlled directly by the subtitle software used by translators, subtitle segmentation is left to the decision of the translator. Translators often have to either compromise the aesthetics in favour of the linguistic wholes or resort to omissions and substitutions. Therefore, modelling the segmentation decisions based on the large available corpora is of great importance for a high-quality subtitle-oriented NMT system.

3.2 Quality criteria filters

In order to assess the conformity of the existing subtitle corpora to the constraints mentioned above, we implement the following filters.

Characters per line (CPL): As mentioned above, the information about line breaks inside

²<https://www.ted.com/participate/translate/guidelines>

subtitle blocks is discarded in the process of creating parallel data. Therefore, we can only assume that a subtitle fulfils the criteria 1 and 2 above by calculating the maximum possible length for a subtitle block; $2 * 42 = 84$ characters for Latin scripts and $2 * 14 = 28$ for Japanese. If $CPL > max_length$ then the subtitle doesn't conform to the length constraints.

Characters per second (CPS): This metric relates to reading speed. For the corpora where time codes and duration are preserved, we calculate CPS as follows: $CPS = \frac{\#chars}{duration}$

Chink-Chunk: Chink-Chunk is a low-level parsing algorithm which can be used as a rule-based method to insert line breaks between subtitles. It is a simple but efficient way to detect syntactic boundaries. It relates to preserving linguistic wholes, since it uses POS information to split units only at punctuation marks (logical completion) or when an open-class or content word (chunk) is followed by a closed-class or function word (chink). Here, we use this algorithm to compute statistics about the type of subtitle block breaks in the data (punctuation break, content-function break or other). The algorithm is described in Algorithm 1.

Algorithm 1: Chink-Chunk algorithm

```

1 if POSlast in ['PUNCT', 'SYM', 'X'] then
2   punc_break += 1;
3 else
4   if POSlast in content_words and POSnext in
      function_words then
5     cf_break += 1;
6   else
7     other_split += 1;
8   end
9 end
10 return punc_break, cf_break, other_split

```

4 Experiments

For our experiments we consider the corpora which are large enough to train NMT systems; OpenSubtitles, JESC and MuST-C. We focus on 3 language pairs, Japanese, Italian and German, paired with English, as languages coming from different families and having a large portion of sentences in all corpora. We tokenise and then tag the data with Universal Dependencies³ to obtain POS tags for the Chink-Chunk algorithm.

To observe the effect of merging processes on preserving the subtitling constraints, we create a version of MuST-C at a subtitle level. We obtain

³<https://universaldependencies.org/>

LP	Total	Extracted	MuST-C
EN-IT	671K	452K / 3.4M	253K / 4.8M
EN-DE	575K	361K / 2.7M	229K / 4.2M
EN-JA	669K	399K / 3M	-

Table 2: Total number of subtitles vs. number of extracted subtitles (in lines) from TED talks .srt files vs. the original MuST-C corpus. The first number shows lines (or sentences respectively), while the second words on the English side.

the same .srt files used to create MuST-C. We extract only the subtitles with matching timestamps from the common talks in the language pair without any merging operations. Table 2 shows the statistics of the extracted corpus. We randomly sample 1,000 sentence pairs and manually inspect their alignments. 94% were correctly aligned, 3% partially aligned and 3% misaligned.

We apply each of the criteria filters in Section 3.2 to the corpora both on the source and the target side independently. Then, we take the intersection of the outputs of all the filters to obtain the lines/sentences which conform to all the criteria.

5 Analysis

Table 3 shows the percentage of preserved lines/sentences after applying each criterion.

Length: The analysis of Characters per line filter shows that both OpenSubtitles and JESC conform to the quality criterion of length in at least 94% of the cases. Despite the merging operations to obtain sentence alignments, OpenSubtitles still preserves a short length of lines, possibly because of the nature of the text of subtitles. A manual inspection shows that the text is mainly short dialogues and the long sentences are parts of descriptions or monologues, which are more rare. On the other hand, the merging operations in MuST-C create long sentences that do not resemble the subtitling format. This could be attributed to the format of TED talks. TED talks mostly contain text written to be spoken, prepared talks usually delivered by one speaker with few dialogue turns. Among all corpora, MuST-C_subs shows the highest conformity to the criterion of length, since indeed no merging operations were performed.

Reading speed: Conformity to the criterion of reading speed is achieved to a lesser degree, as

LP	Corpus	Format	Time	CPL (s/t) %	CPS (s/t) %	Chink-Chunk (s/t) %	Total%
EN-IT	MuST-C	segment	✓	49 / 48	78 / 72	99 / 99	45
	OpenSubtitles	segment	-	95 / 94	-	99 / 99	91
	MuST-C_subs	subtitle	✓	99 / 98	86 / 81	87 / 83	79
EN-DE	MuST-C	segment	✓	51 / 47	77 / 66	99 / 99	42
	OpenSubtitles	segment	-	95 / 95	-	99 / 99	92
	MuST-C_subs	subtitle	✓	99 / 98	84 / 75	87 / 87	74
EN-JA	OpenSubtitles	segment	-	96 / 93	-	99 / 98	91
	JESC	subtitle	-	97 / 94	-	88 / 87	85
	MuST-C_subs	subtitle	✓	99 / 94	85 / 99	92 / 91	83

Table 3: Percentage of data preserved after applying each of the quality criteria filters on the subtitling corpora independently. Percentages are given on source and target side (s/t), except for the *Total* where source and target are combined.

shown by the Characters per second filter. Except for Japanese, where the allowed number of characters per line is lower, all other languages range between 66%-86%. In general, MuST-C_subs, being in subtitling format, seems to conform better to reading speed. Unfortunately, time information is not present in corpora other than the two versions of MuST-C, therefore a full comparison is not possible.

Linguistic wholes: The Chink-Chunk algorithm shows interesting properties of the subtitle breaks for all the corpora. MuST-C and OpenSubtitles conform to the criterion of preserving linguistic wholes in 99% of the sentences, which does not occur in the corpora in subtitle format; JESC and MuST-C_subs. Since these two corpora are compiled by removing captions based on unmatched time codes, the integrity of the documents is possibly broken. Subtitles are removed arbitrarily, so consecutive subtitles are often not kept in order. This shows the importance of preserving the order of subtitles when creating subtitling corpora.

This observation might lead to the assumption that JESC and MuST-C_subs are less subtitle-like. However, a close inspection of the breaks shows that OpenSubtitles and MuST-C end in a punctuation mark in 99.9% of the cases. Even though they preserve logical completion, these corpora do not contain sufficient examples of line breaks preserving linguistic wholes. On the other hand, the subtitle-level corpora contain between 5%-11% subtitle breaks in the form of content-function word. In a realistic subtitling scenario, an NMT system at inference time will often receive unfinished sentences, either from an audio stream or a subtitling template. Therefore, line break information might be valuable for training NMT systems

that learn to translate and segment.

The total retained material shows that OpenSubtitles is the most suitable corpus for producing quality subtitles in all investigated languages, as more than 90% of the sentences passed the filters. However, this is not a fair comparison, given that the data was filtered with only 2 out of the 3 filters. One serious limitation of OpenSubtitles is the lack of time information, which does not allow for modelling reading speed. We showed that corpora in subtitling format (JESC, MuST-C_subs) contain useful information about line breaks not ending in punctuation marks, which are mostly absent from OpenSubtitles. Since no information about subtitle line breaks (inside a subtitle block) is preserved in any of the corpora, the criterion of equal length of lines cannot be explored in this study.

6 Conclusions and discussion

We explored whether the existing parallel subtitling resources conform to the subtitling constraints. We found that subtitling corpora generally conform to length and proper line breaks, despite the merging operations for aligning parallel sentences. We isolated some missing elements: the lack of time information (duration of utterance) and the insufficient representation of line breaks other than at punctuation marks.

This raises several open issues for creating corpora for subtitling-oriented NMT; i) **subtitling constraints:** a subtitling corpus, in order to be representative of the task, should respect the subtitling constraints; ii) **duration of utterance:** since the translation of a subtitle depends on the duration of the utterance, time information is highly relevant; iii) **integrity of documents:** a subtitle often occupies several lines, therefore the order of

subtitles should be preserved whenever possible; iv) **line break information:** while parallel sentence alignments are indispensable, they should not compromise line break and subtitle block information. Break information could be preserved by inserting special symbols.

We intend to use these observations for an adaptation of MuST-C, containing triplets (audio, source language subtitle, target language subtitle), preserving line break information and taking advantage of natural pauses in the audio. In the long run, we would like to train NMT systems which predict line breaks while translating, possibly extending the input context using methods from document level translation.

Acknowledgements

This work is part of a project financially supported by an Amazon AWS ML Grant.

References

- Aitor Álvarez, Haritz Arzelus, and Thierry Etchegoyhen. 2014. Towards customized automatic segmentation of subtitles. In *Advances in Speech and Language Technologies for Iberian Languages*, pages 229–238, Cham. Springer International Publishing.
- Aitor Álvarez, Carlos-D. Martínez-Hinarejos, Haritz Arzelus, Marina Balenciaga, and Arantza del Pozo. 2017. Improving the automatic segmentation of subtitles through conditional random field. In *Speech Communication*, volume 88, pages 83–95. Elsevier BV.
- Petar Božović, Tomaž Erjavec, Jörg Tiedemann, Nikola Ljubešić, and Vojko Gorjanc. 2018. Opus-Montenegrinsubs 1.0: First electronic corpus of the Montenegrin language. In *Conference on Language Technologies & Digital Humanities*.
- Mauro Cettolo, Christian Girardi, and Marcello Federico. 2012. Wit³: Web Inventory of Transcribed and Translated Talks. In *Proceedings of the 16th Conference of the European Association for Machine Translation (EAMT)*, pages 261–268, Trento, Italy, May.
- Jorge Diaz Cintas and Aline Remael. 2007. *Audiovisual Translation: Subtitling*. Translation practices explained. Routledge.
- Mattia Antonino Di Gangi, Roldano Cattoni, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2019. MuST-C: a multilingual speech translation corpus. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, Minneapolis, MN, USA, June.
- Mark Liberman and Kenneth Church. 1992. Text analysis and word pronunciation in text-to-speech synthesis. *Advances in Speech Signal Processing*, pages 791–831.
- Pierre Lison and Jörg Tiedemann. 2016. Opensubtitles2016: Extracting large parallel corpora from Movie and TV subtitles. In *Proceedings of the International Conference on Language Resources and Evaluation, LREC*.
- José Manuel Martínez Martínez and Mihaela Vela. 2016. SubCo: A learner translation corpus of human and machine subtitles. In *Language Resources and Evaluation Conference (LREC)*.
- Jan Niehues, Roldano Cattoni, Mauro Cettolo Sebastian Stuke an, Marco Turchi, and Marcello Federico. 2018. The IWSLT 2018 evaluation campaign. In *Proceedings of IWSLT 2018*.
- Elisa Perego. 2008. Subtitles and line-breaks: Towards improved readability. *Between Text and Image: Updating research in screen translation*, 78(1):211–223.
- Reid Pryzant, Yongjoo Chung, Dan Jurafsky, and Denny Britz. 2018. JESC: Japanese-English Subtitle Corpus. *Language Resources and Evaluation Conference (LREC)*.
- Dhevi J. Rajendran, Andrew T. Duchowski, Pilar Orero, Juan Martinez, and Pablo Romero-Fresco. 2013. Effects of text chunking on subtitling: A quantitative and qualitative examination. *Perspectives*, 21(1):5–21.
- Hye-Jeong Song, Hong-Ki Kim, Jong-Dae Kim, Chan-Young Park, and Yu-Seop Kim. 2019. Inter-sentence segmentation of YouTube subtitles using long-short term memory (LSTM). 9:1504.
- Jörg Tiedemann. 2008. Synchronizing translated movie subtitles. In *Proceedings of the International Conference on Language Resources and Evaluation, LREC 2008*.

Asymmetries in Extraction From Nominal Copular Sentences: a Challenging Case Study for NLP Tools

Paolo Lorusso, Matteo Greco, Cristiano Chesi, Andrea Moro

NEtS at Scuola Universitaria Superiore IUSS.

P.zza Vittoria 15, I-27100 Pavia (Italy)

{paolo.lorusso, matteo.greco,
andrea.moro, cristiano.chesi}@iusspavia.it

Abstract

In this paper we discuss two types of nominal copular sentences (Canonical and Inverse, Moro 1997) and we demonstrate how the peculiarities of these two configurations are hardly considered by standard NLP tools that are currently publicly available. Here we show that example-based MT tools (e.g. Google Translate) as well as other NLP tools (UDpipe, LinguA, Stanford Parser, and Google Cloud AI API) fail in capturing the critical distinctions between the two structures in the end producing both wrong analyses and, possibly as a consequence of a non-coherent (or missing) structural analysis, incorrect translations in the case of MT tools. To support the proposed analysis, we present also an empirical study showing that native speakers are indeed sensitive to the critical distinctions. This poses a sharp challenge for NLP tools that aim at being cognitively plausible or at least descriptively adequate (Chowdhury & Zamparelli 2018).

1. Introduction

The main hypothesis of this paper is that sentence comprehension cannot be achieved independently from a coherent structural analysis. To support this claim, we first present a precise structural analysis that is critical for recovering the relevant dependencies within specific constructions, then we will show that the crucial structural properties captured by the theoretical framework are in fact correctly perceived by native speakers, but not

revealed by some widely used Natural Language Processing (NLP) tools. This leads to poor performance in tasks like Machine Translation (MT).

This argument seems to us especially relevant in those structural configurations in which a non-local dependency must be established: in parsing, for instance, interpreting correctly a *wh*-dependency requires that the *dependent* (the *wh*-phrase) and the *dependee* (the head selecting the *wh*-phrase as its argument/modifier) are identified, and the nature of the dependence disambiguated (e.g. argument vs. modifier). In (1) we exemplify the special case of a non-local dependency between a *wh*-PP and a DP it depends on (a co-indexed underscore signals the possible extraction sites, hence the dependent constituent; the diacritic “*” prefixes, as usual, illegal sites):

- (1) [Di quale segnale]_i [i telescopi * __i] hanno
Of which signal the telescopes have
scoperto * __i [un’interferenza __i]?
discovered an interference?
‘[which signal]_i did the telescopes discover
an interference of __i?’

The second DP *un’interferenza* (an interference) (the internal argument) is the dependee of the *wh*-phrase and neither the subject DP nor the predicate can host this *wh*-dependency instead.

According to Google Translate (as of 12th July 2019), this second option seems indeed a viable one:

- (2) What signal did the telescopes find an interference?

The translation is ill formed being the internal argument of *find* filled both by the *wh*-phrase and

the DP *an interference* (which cannot take a *wh*-DP as its own argument due to the absence of a relevant preposition).

In this work we focus on a similar non-local dependency involving two kinds of copular sentences: Inverse (3.a) and Canonical (3.b). Using these constructions, we will test the availability of *wh*-PP sub-extraction from both the first and the second DP as exemplified in (4).

- (3) a. le foto del muro **sono** la causa della rivolta
the pictures of the wall **are** the cause of the riot
b. la causa della rivolta **sono** le foto del muro
the cause of the riot **are** the pictures of the wall
'the cause of the riot **is** the pictures of the wall'
- (4) a. [Di quale rivolta]_i le foto del muro **sono**
of which riot the pictures of the wall **are**
la causa _i ?
the cause
b. [Di quale muro]_i le foto _i **sono**
of which riot the pictures of the wall **are**
la causa della rivolta?
the cause of the riot

In the first part of this paper (§2), we will briefly present an analysis for these constructions, then we will demonstrate that native speakers are selectively sensitive both to the copular structural configuration (Canonical vs. Inverse) and to the extraction site (subject vs. predicate) (§3). In §4 we will test the insensitivity of some freely available NLP tools (Google Translate, the Natural Language service of Google Cloud AI API, UDpipe, Stanford Parser and Lingua) to the syntactic oppositions previously discussed.

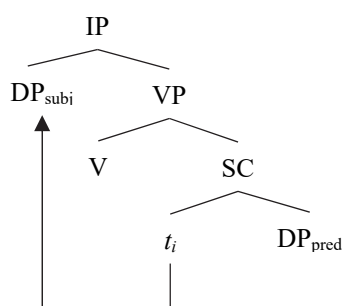
2. The structure of nominal copular sentences

Copular sentences are those sentences whose main verb is *to be* (the copula) and its equivalents across languages. A subset of copular sentences is the one involving two DPs, linearly ordered as DP V DP. Those are dubbed *nominal copular sentences*. In this configuration, a nominal phrase realizes the predicate of the sentence ("the cause..." in (3)) while the other is the subject of the predicate ("the pictures..." in (3)). According to Moro (1997), nominal copular sentences can be distinguished in two subtypes: *Canonical copular sentences* (3.a) – in which the order is subject-copula-predicative expression – and *Inverse copular sentences* (3.b) – in which the order is inverted, i.e. predicative expression-copula-subject.

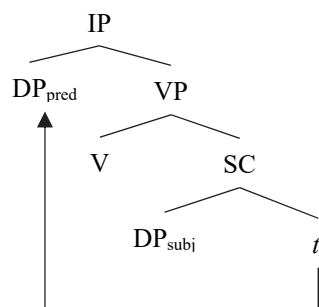
Moro (1991, 1997, 2006) showed that these two types of copular constructions can be distinguished on the basis of different diagnostics like agreement on the verb, grammaticality for the extraction of DPs (*Wh*- or clitic) and pronominal binding.

Traditionally, copular sentences are analyzed as involving the raising of a DP from the same base generated structure (Stowell 1978). Moro (1997, 2018) showed that the predicate DPs (including *there* and its equivalents across languages) can be raised along with the subject DPs to the preverbal position from the so-called *Small Clause* (SC) – a structure resulting from merging two DPs (Moro 2000, 2009 Chomsky 2013, Rizzi 2016). In other words, while in Canonical copular sentences the subject DP raises to the preverbal position and the predicative DP stays *in situ* inside the small clause in the postverbal position (4), in the Inverse copular sentences the predicative DP raises to the preverbal position and the subject DP stays *in situ* inside the small clause in the postverbal position (5).

- (5) Canonical copular sentence structure



- (6) Inverse copular sentence structure



2.1 Asymmetries in copular sentences

These two different representations offer a principled explanation for many asymmetries across languages. Distinguishing between Canonical and Inverse copular sentences is not

always easy or possible (see Jespersen 1924 as cited in Moro 1997). However, agreement and PP/*ne* sub-extraction offer robust diagnostics. For example, verbs invariably agree with the subject DP in Italian (7), regardless of the pre-verbal or post-verbal position, while they invariably agree with the preverbal DP in English (8):

- (7) a. le foto **sono**/*è la causa
the pictures are /*is the cause
b. la causa **sono**/*è le foto
the cause are/*is the pictures

Italian

- (8) a. the pictures **are**/*is the cause.
b. the cause ***are/is** the pictures

English

Extraction is only allowed from the post-verbal DP – the predicate – in Canonical sentences (9), whereas it is not allowed from the post-verbal DP – the subject – in Inverse copular sentences (10).

- (9) a. **which** *rioto*_i do you think a picture of the wall was **the cause of** *__i*?
b. **di quale** *rivolta*_i pensi che una foto del of which riot_i do you think that a picture of *__the* muro sia la causa *__i*?
wall is the cause *__i*?
(10) a. ***which** *wall*_i do you think a cause of the riot was a **picture of** *__i*?
b. ***di quale** *muro*_i pensi che la causa della of which wall_i you think that the cause of *__the* rivolta sia **una foto** *__i*?
riot is a picture *__i*?

3. Experimental evidence supporting the analysis of copular sentences

Before considering the computational side or the proposed structural analysis we investigated whether the human parser is sensitive to the critical distinctions illustrated here. Two experiments are discussed, testing the processing of Canonical vs Inverse copular sentences (first condition) involving the extraction of a wh-element from a DP embedded either under the subject or the predicate (second condition).

Our prediction was that the sensitivity to agreement and to the *argumental* vs. *predicative* role distinction for the two DPs involved would have influenced both the online and the offline performance of native speakers: participants should show an advantage in parsing Canonical copular sentences (vs. Inverse ones), since only

the Canonical configuration allow the extraction from the predicate DP, whereas all the other kinds of extraction – from the subject in Canonical and from both the subject and the predicate in Inverse – should be disallowed (§2.1).

In order to test these hypotheses, we performed (i) a Self-Paced Reading (SPR) experiment with a Sentence Comprehension Task at the end, and (ii) an Acceptability Judgement Task (AJT).

3.1 Material and methods

In both the SPR and AJT the set of stimuli was the same: 128 items (divided in 4 conditions) and 40 fillers, in SPR, and 60 fillers, in AJT per condition (72 items per experiment in SPR, 92 in AJT). The 2x2 design produced four experimental conditions, exemplified in (11):

(11) *Condition 1:*

Canonical + Extraction from the Subject

*[_{PP} Di quale muro]_i ... [_{DP} le **foto** *__i*]_a sono [_{SC} [_a]
Of which wall the pictures are
[_{DP} la **causa** [_{PP} della rivolta]]]?
the cause of *__the* riot?

Condition 2:

Canonical + Extraction from the Predicate

[_{PP} Di quale rivolta]_k ... [_{DP} le **foto** [_{PP} del muro]]_a
Of which riot the pictures of *__the* wall
sono [_{SC} [_a] [_k]]
are the cause?

Condition 3:

Inverse + Extraction from the Subject

*[_{PP} Di quale muro]_i ... [_a la **causa** [_{PP} della rivolta]]_b
Of which wall the cause of *__the* riot
sono [_{SC} [_i le **foto** *__i*] [_b]]?
are (=is) the pictures?

Condition 4:

Inverse + Extraction from the Predicate

*[_{PP} Di quale rivolta]_k ... [_a la **causa** *__k*]_b sono [_{SC}
Of which riot ... the cause are (=is)
[_{DP} le **foto** [_{PP} del muro]] [_b]]?
the pictures of *__the* wall

3.2 Self-Paced Reading

32 native Italian speakers participated in the experiment. Stimuli were composed by questions and by their answers; participants had to read the question word by word and, then, the answer. Finally, they had to judge the appropriateness of the answer.

3.3 Results

Participants showed higher accuracy in answering to comprehension questions when the extraction occurred from the post-verbal DP in Canonical copular sentences – DP *predicate* in Condition 2 – than in Inverse copular sentences – DP *subject* in Condition 3 – while extraction from the Inverse copular constructions induced lower accuracy (-0.41 , $z=-2.054$, $p=0.04$; Fig. 1). This confirms that the structural asymmetry between referential subjects and predicative DPs has a central role in both the processing and the comprehension of nominal copular sentences. Similarly, Inverse vs Canonical opposition seems relevant since extractions from both sites in the Inverse copular constructions produce lower accurate answers compared to the extraction from the predicate in canonical copulars (coherently with Moro 1997, 2006 that predict the DP in both inverse constructions to be illegal extraction sites).

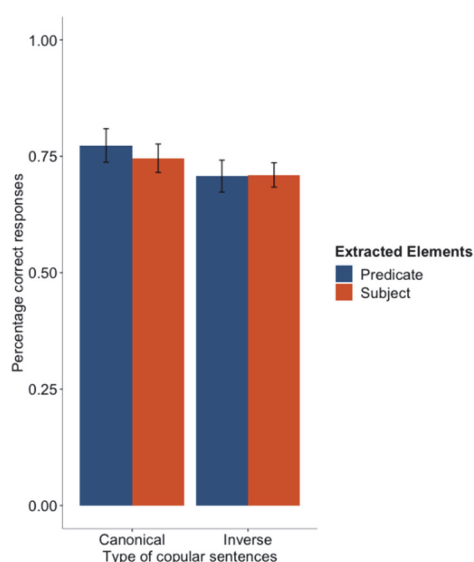


Fig.1 Percentage of correct answers across conditions.

Reading times, on the other hand, revealed a clear difference at the copular region for the two conditions ($t=3.37$ $p=0.002$) suggesting a penalty for the Inverse copular constructions compared to the Canonical one. Also at the first DP region the Predicate vs Subject distinction is productively differentiated ($t>2$ $p=0.008$) indicating the *la causa* (“the cause”) and *“le foto”* (“the pictures”) conditions, respectively predicate and subject condition, are perceived as different.

3.4 Acceptability Judgement Task

40 native Italian speakers participated in the experiment. Stimuli were the same than in SPR.

Participants had to rate the acceptability of questions on a scale from 1 to 7.

3.5 Results

The results (fig.2) confirm the previous on-line findings and show that (i) Canonical constructions were more acceptable than Inverse ones and that (ii) among the different types of copular sentences, the ones with an extraction from predicates have higher rates than the ones with extraction from subjects.

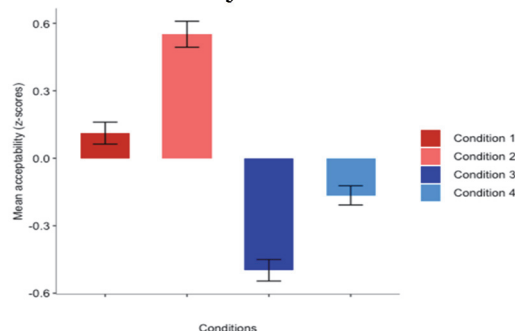


Fig.2 Acceptance rates across conditions.

4. Parsing copular sentences

To evaluate the state-of-the-art of NLP with respect to the contrasts we discussed (Canonical vs Inverse copular sentences) in a configuration where overt agreement disambiguates the critical roles (predicate vs subject), we ran few tests using the following tools:

1. UDPipe (Straka et al 2016)
2. Stanford Parser - English (Chen & Manning 2014)
3. Lingua parser (Attardi, Dell’Orletta 2009)
4. Google Translate (translate.google.com)
5. Google Cloud AI Solutions (cloud.google.com)

We first tested standard Canonical (3.a) and Inverse (3.b) copular constructions, then we tried to assess qualitatively the output analyses provided by these tools with respect to sub-extraction from the predicate in Canonical sentences (9.a-b), here repeated for convenience:

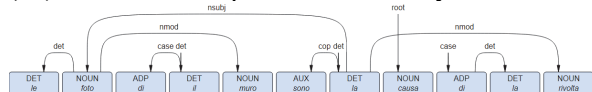
- (3) a. le foto del muro **sono** la causa della rivolta
the pictures of the wall are the cause of the riot
b. la causa della rivolta **sono** le foto del muro
the cause of the riot are the pictures of the wall
the cause of the riot **is** the pictures of the wall

- (9) a. **which riot_i** do you think a picture of the wall was **the cause of** _i?
 b. **di quale rivolta_i** pensi che una foto del muro sia la causa _i?
of which riot_i do you think that a picture of the wall is the cause _i?

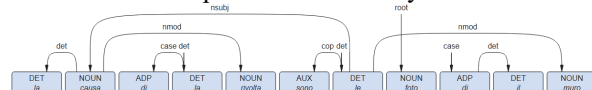
4.1 UDpipe

UDPipe Natural Language Processing - Text Annotation interface (Wijffels 2018, Straka et al 2016) provides a handy tool easily integrated in the R environment. Various pre-trained models are available for many languages. We run our analyses using the pre-trained model *italian-isdt-ud-2.4-190531*. The results of the analysis for both Canonical (10.a) and Inverse (10.b) are simply the same. In fact, not even the basic local dependencies are fully recovered (e.g. det-noun). The analysis of the sub-extraction from predicate in Canonical structures (13.a) is paradoxically less disastrous than the other analyses, but if we try to analyze sub-extraction from the subject of a Canonical construction, we obtain wrong analyses (13.b) (the *wh*- items is considered an extra argument of *cause*):

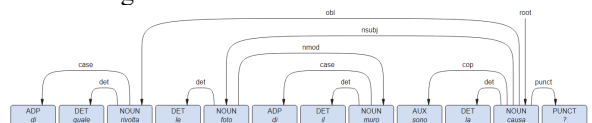
- (12) a. Canonical copular sentence analysis



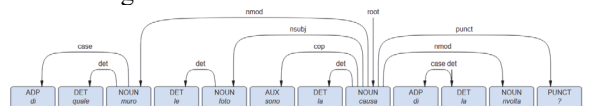
- b. Inverse copular sentence analysis



- (13) a. sub-extraction from predicate in Canonical configuration



- b. sub-extraction from subject in Canonical configuration

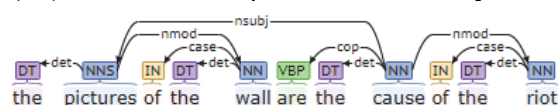


4.2 Stanford Parser

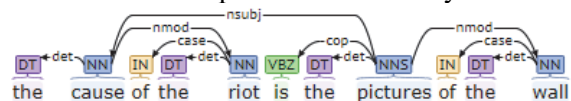
Stanford parser (Chen & Manning 2014) can be considered the state-of-the-art parser for English. Canonical constructions, in fact, gave the opportunity to live up to expectations: the analysis of the canonical copular sentence (14.a) is perfectly in line with the analysis presented in §2-§2.1 (*cause* is identified as predicate and *pictures*

as its subject). Unfortunately, the same analysis is proposed for inverse copular constructions (14.b).

- (14) a. Canonical copular sentence analysis

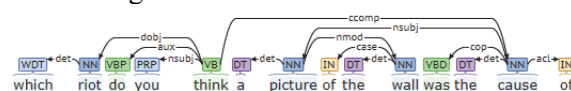


- b. Inverse copular sentence analysis



The quality of the analysis for the sub-extraction case confirms every suspicion: the sub-extracted *wh*-item (*which riot*) is wrongly associated to the matrix predicate (*think*) (15).

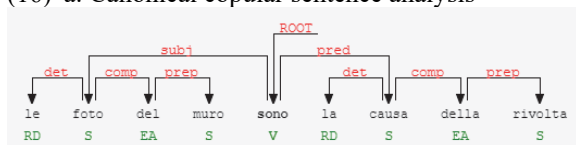
- (15) sub-extraction from predicate in Canonical configuration



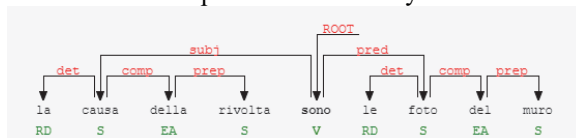
4.3 LinguA

LinguA annotation pipeline (service provided on-line by ItaliaNLP Lab at Istituto di Linguistica Computazionale "Antonio Zampolli" ILC in Pisa) has been used for our tests on Italian, implementing a version of Attardi & Dell'Orletta (2009) parser (currently the state-of-the-art parser for Italian). The analyses of this parser are definitely more precise than the ones proposed by the UDpipe tool, but the symmetric results returned for both Canonical and Inverse copular sentences did not identify either the dependency between the predicate and the subject or their actual role in the structure (16.a-b). The analysis of the extraction, interestingly attempts an interpretation of the *wh*- item as an (extra) argument of the first DP (*le foto [di quale rivolta] (del muro)*). This is a wrong analysis, but it is coherent with the slow-down observed in self-paced reading experiment (§3.3) at the first DP region, though the parser does not make the relevant distinction between subject (17.a) and predicate (17.b) (in this second case, sub-extraction is interpreted as a copula argument).

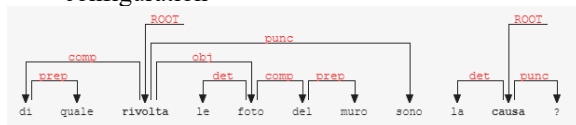
(16) a. Canonical copular sentence analysis



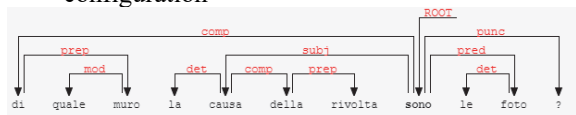
b. Inverse copular sentence analysis



(17) a. sub-extraction from predicate in Canonical configuration



b. sub-extraction from subject in Inverse configuration



4.4 Google AI

We finally investigated the Natural Language service – one of the tools provided by Google Cloud AI Solutions API – which returns syntactic representations of sentences (<https://cloud.google.com/natural-language/>).

While both canonical and inverse copular analyses are equivalent in English to the ones provided by the Stanford Parser (hence partially consistent with our analyses), in Italian, using the Canonical copular sentence ‘*le intercettazioni_k sono_k la documentazione_i*’ (‘the interceptions are the documentation’), the tool incorrectly analyses the predicate DP *the documentation* as an attribute (fig. 4) (this might be a consistent annotation of all nominal predicates Google adopted, but it is clearly misleading here). Moreover, when it is provided with the Inverse form of the sentence ‘*la documentazione sono le intercettazioni*’ (lett. the documentation are the interceptions; ‘The documentation is the interceptions’), the tool incorrectly analyzes the raised predicative DP *the documentation* – singular noun – as the subject, putting it in a wrong agreement relation with the verb (plural form) (Fig. 5). Then, in the end, this parser fails in recognizing the critical difference between Canonical and Inverse copular sentences giving exactly the same analysis for both cases (3.a) and (3.b).

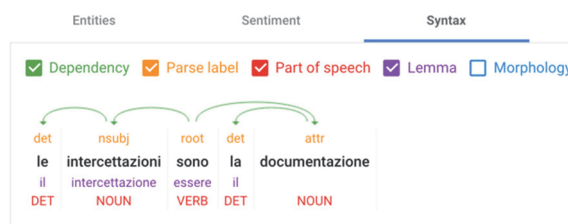


Fig.4 The structural analysis of the Canonical sentence ‘*le intercettazioni sono la documentazione*’ (‘The interceptions are the documentation’) given by Google Natural Language.

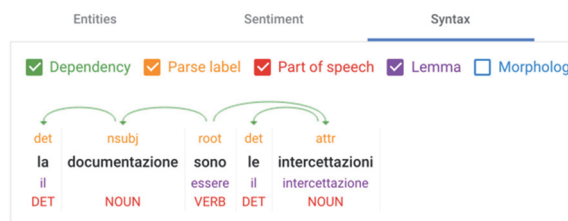


Fig.5 Structural analysis of the Inverse copular sentence ‘*la documentazione sono le intercettazioni*’ (lett. the documentation are the interceptions; ‘The documentation is the interceptions’) given by Google Natural Language.

4.4 Google Translate

In order to evaluate the impact of these wrong analyses on a practical NLP task, we finally carried out our conclusive experiments on one of the most famous and largely exploited machine translation software: *Google Translate*.

Starting with simple examples, we observed that when the tool is provided with the Italian Inverse copular sentence ‘*La causa della rivolta sono le foto del muro*’ (lett. the cause of the riot are the pictures of the wall; ‘The cause of the riot is the pictures of the wall’), it gives the wrong English translation ‘**The cause of the uprising are the photos of the wall*’ (Fig.6), in which the verb does not agree with the pre-verbal DP ‘*the cause of the uprising*’, contrary to what it does in English (as we saw in 7).

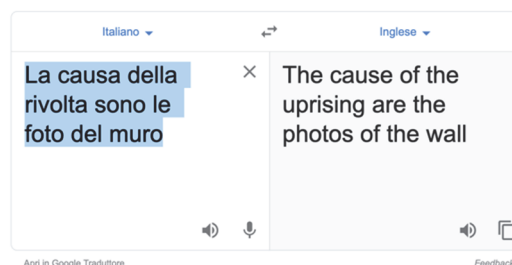


Fig.6 Example from Google translate: <https://translate.google.it/?hl=it#view=home&op=translate&sl=auto&tl=en&text=La%20causa%20della%20rivolta%20sono%20le%20foto%20del%20muro>

Interestingly, reversing the translation from English to Italian *the cause of the riot is the pictures of the wall* the system correctly produces *la causa della rivolta sono le immagini del muro* where proper agreement (with the post-verbal subject) is in place. Since the analysis provided by any tool we tested is theoretically inconsistent with this result, we hypothesized that this translation could have been obtained adopting an example-based approach; it was worth then to test if the correct agreement with the post-verbal subject is just an accident (this is a well known prototypical sentence, widely discussed in literature and it might have been included in the Google Translate training set) or if the analysis is generalized of any possible subject/predicate pair.

A sentence like *la documentazione sono le intercettazioni* (lett. the documentation are the interceptions, that means ‘The documentation is the interceptions’) would suit our purpose nicely. In the English > Italian direction the correct singular copular agreement is produced (“the documentation is the interceptions”) but from Italian to English this time the wrong agreement is obtained, totally ignoring the number of the real post-verbal subject (*the documentation is the interceptions* > *la documentazione è le intercettazioni*). We concluded then that no deep analysis is attempted so as to distinguish between subject and predicate roles and this turns out to be fatal.

5. Conclusion

In this paper we demonstrated that nominal copular sentences constitute a clear challenge for the computational analysis since the same string of elements [DP V DP] can have in principle two different syntactic representations (hence two different meanings), depending on which kind of copular sentence is realized (Canonical or Inverse). In this paper, we spotted various glitches in the automatic analyses which in the end led either to significant failures (Google Translate) or to rough structural hypotheses that bluntly ignore the relevant contrasts here discussed. Our empirical study, testing both online and offline the *wh*-PP sub-extraction possibilities from both subject and predicate DPs, shows that native speakers are sensitive with respect to the different structural roles; in addition, they perceive as expected the underlying structural representation of Canonical vs. Inverse copular construction. None of the NLP tools we tested succeeded in providing a full set of coherent analyses, with the

exception of the Stanford Parser for English that at least succeeded in analyzing correctly the canonical copular sentences. This analysis was however insufficient in the case of inverse constructions and in case of sub-extraction, confirming that non-local dependencies are critical configurations native speakers are able to parse but machine do not, yet.

Reference

- Attardi G., Dell’Orletta F. (2009). Reverse Revision and Linear Tree Combination for Dependency Parsing“. In: *NAACL-HLT 2009 – North American Chapter of the Association for Computational Linguistics – Human Language Technologies (Boulder, Colorado, June 2009). Proceedings*, Association for Computational Linguistics, 2009. pp. 261 – 264.
- Chen D., C. D. Manning. (2014). A Fast and Accurate Dependency Parser using Neural Networks. *Proceedings of EMNLP 2014*. pp. 740-750
- Chomsky, N., (2013). ‘Problems of projection.’ *Lingua* 130:33–49
- Chowdhury, S. A., & Zamparelli, R. (2018, August). ‘RNN simulations of grammaticality judgments on long-distance dependencies.’ In *Proceedings of the 27th International Conference on Computational Linguistics* (pp. 133-144).
- Jespersen, O., (1924) *The Philosophy of Grammar*, Allen & Unwin, London.
- Moro, A., (1991). The raising of predicates: copula, expletives and existence. *MIT Working Papers in Linguistics* 15: 119-181.
- Moro, A., (1997). *The Raising of Predicates*. Cambridge: Cambridge UP
- Moro, A., (2000). *Dynamic Antisymmetry*. *Linguistic Inquiry Monograph, Series*, MIT Press
- Moro, A., (2006). ‘Copular sentences.’ In Everaert, M. & H. van Riemsdijk (eds.), *MA. Blackwell Companion to Syntax II*, Blackwell, Oxford, 1-23.
- Moro, A., (2009). ‘Rethinking Symmetry: A Note on Labelling and the EPP.’ In *La grammatica tra storia e teoria: Scritti in onore di Giorgio Graffi*, edited by P. Cotticelli Kurras and A. Tomaselli, 129–31. *Alessandria: Edizioni dell’Orso*; also at <http://www.ledonline.it/snippets/allegati/snippets19007.pdf>.
- Moro, A., (2018). ‘Copular sentences.’ In Everaert, M. & H. van Riemsdijk (eds.), *MA. Blackwell Companion to Syntax, Revised edition vol. II*, Blackwell, Oxford, 1-23.

- Rizzi, L., (2016). 'Labeling, maximality, and the head-phrase distinction.' *The Linguistic Review* 33, 103–127.
- Straka, M., Hajic, J., & Straková, J. (2016). UDPipe: trainable pipeline for processing CoNLL-U files performing tokenization, morphological analysis, pos tagging and parsing. In *Proceedings of the tenth international conference on language resources and evaluation* (LREC 2016) (pp. 4290-4297).
- Stowell, T., (1978). 'What was there before there was there.' In D. Farkas et al., eds., *Papers from the Fourteenth Regional Meeting, Chicago Linguistic Society*. Chicago Linguistic Society, University of Chicago.
- Wijffels, J. (2018). *udpipe: Tokenization, Parts of Speech Tagging, Lemmatization and Dependency Parsing with the ,UDPipe ',NLP 'Toolkit*. R package version 0.5.

Objective Frequency Values of Canonical and Syntactically Modified Idioms: Preliminary Normative Data

Azzurra Mancuso & Alessandro Laudanna

LaPSUS, Laboratory of Experimental Psychology, University of Salerno

Via Giovanni Paolo II, 132 Fisciano, SA, 84084, Italy

amancuso@unisa.it; alaudanna@unisa.it

Abstract

In this study we collected several objective frequency values for 124 Italian idiomatic expressions, in order to verify the relation among these measures of frequency and a set of subjective variables (e.g., familiarity, meaning knowledge, age of acquisition, etc.) which are relevant from a psycholinguistic perspective, since they are supposed to play a role in idiom processing. Specifically, we calculated the following frequency types: occurrences of content words, (lemma and word-form values), occurrences of canonical idioms (e.g., Paolo broke the ice), occurrences of syntactically manipulated idioms (e.g., The ice was suddenly broken by Paolo). We discuss the results of correlational analyses.

1. Introduction

Several psycholinguistic norms are available for pictures and words (e.g., Barca, Burani, & Arduino, 2002; De Martino, Mancuso and Laudanna, 2017; Janssen, Pajtas, & Caramazza, 2011; Montefinese, Ambrosini, Fairfield, & Mammarella, 2014). However, this is less frequent for longer word-combinations, such as idiomatic expressions. An idiomatic expression comprises several words whose overall figurative meaning is not a direct function of its components (Tabossi, Arduino, & Fanari, 2011). For instance, the Italian idiomatic expression *rompere il ghiaccio* (“break the ice”) means “to take the initiative in an embarrassing situation” and thus its global meaning is far from the meaning of its components.

Some norms are available in English (Abel, 2003; Cronk, Lima, & Schweigert, 1993; Libben

& Titone, 2008; Titone & Connine, 1994b), in French (Caillies, 2009; Bonin, Méot, & Bugaiska, 2013), in Bulgarian (Nordmann & Jambazova, 2017), in German (Citron et al., 2016) and in Italian (Tabossi et al., 2011). These databases collect mean values obtained from subjective ratings for some relevant psycholinguistic variables (such as age of acquisition, familiarity, meaning knowledge, etc.).

The existence of norms for idiomatic expressions has made it possible to account for issues concerning the comprehension, the production and the lexical storage of idioms (e.g., Cutting & Bock, 1997; Konopka & Bock, 2009; Sprenger, Levelt, & Kempen, 2006).

There are different theories on the topic of how idioms are stored in memory. According to some authors, idioms correspond to lexical units (e.g., Swinney & Cutler, 1979), whereas for others, they are stored as configurations of words (Cacciari & Tabossi, 1988; 2014). As claimed by Bonin et al. (2013), “it is therefore obvious that no empirical test of the different views of idiom processing is possible without first collecting norms for idioms”.

2. The present study

In the present research, we computed the frequency of 124 Italian idiomatic expressions in text corpora, in order to verify the relation among objective measures of frequency and a set of subjective variables which are available for Italian (Tabossi et al., 2011).

Some studies have underlined the influence exerted by the frequency values in the processing of these strings (Cronk et al., 1993; Libben & Titone, 2008; Bonin et al., 2013). In these works, the frequency values were obtained by calculating the familiarity of the expressions or the objective frequency (occurrence) of the individual words that compose the strings. Between the two methods, the first proved to be a better predictor of the complexity of processing (Bonin et al.,

2013; Libben & Titone, 2008). The authors attributed this effect to the fact that the idiomatic meaning is often arbitrarily related to that of the individual constituents.

In our study, we pursued three main goals. The first was to collect the objective frequency of the isolated words that make up the Italian idiomatic expressions. Word frequency is certainly one of most important variables to have been considered by studies investigating reading or speaking. For instance, all influential models of word reading (e.g., Coltheart, Rastle, Perry, Langdon, & Ziegler, 2001; Harm & Seidenberg, 2004) are able to account for the finding that high-frequency words are processed faster and more accurately than low-frequency words in experimental tasks such as lexical decision and reading aloud. However, the influence of objective word frequency in idiom processing has received little attention (Cronk et al., 1993; Libben & Titone, 2008; Bonin et al., 2013). In the Italian normative study of idiomatic expressions (Tabossi et al., 2011), this variable was not taken into account.

The second goal was to obtain the objective frequency of idiomatic expressions, intended as the frequency of use of the idiomatic expression considered in its entirety. To our knowledge, all previous studies had not calculated this variable but focused exclusively on the subjective frequency of idioms. We claim that this methodology could offer several advantages to the research on idiom processing. First of all, it provides an objective measure of the degree of exposure to a given idiomatic expression by speakers, without being affected by any distortion or idiosyncrasy coming from subjective evaluations of familiarity. Some studies have shown that subjective frequency is a good index of the frequency of encounter of the words (Balota, Pilotti, & Cortese, 2001). However, the reliability of estimates of other-based familiarity (as considered in Bonin et al., 2013 and Tabossi et al., 2011) can be problematic, since it is more likely that participants can reliably estimate their own frequency of exposure to an idiomatic expression than how well other people know such expressions (Cronk et al., 1993; Libben & Titone, 2008; Titone & Connine, 1994b).

Moreover, the availability of corpus-based frequency values may offer an ideal shortcut to the preparation of psycholinguistic experiments, since familiarity estimates are often difficult to obtain, as they typically require running pre-studies to collect ratings. In this direction, recent studies claimed that subjective frequency ratings

are no longer needed when objective word frequency norms are available (Brysbaert et al., 2011).

The third purpose of our study was to obtain objective frequency values of idioms used in a not canonical form (e.g., passive form, adjective and adverb insertion, etc.). Idioms have been traditionally described as fixed expressions, highly restricted in their realization (Cacciari & Tabossi, 1988; Gibbs, 1980; Swinney & Cutler, 1979; Titone & Connine, 1999). However, more recent corpus and experimental studies have shown that they are more flexible than previously thought (Moon, 1998; Barlow, 2000; Geeraert, Baayen, & Newman, 2017; Langlotz, 2006; Tabossi, Wolf, & Koterle, 2009; Vietri, 2014; Mancuso, Elia, Laudanna, & Vietri, 2019; Kyriacou, Conklin, & Thompson, 2019). The issue of idiom syntactic flexibility has received a renewed interest, since it also addresses the problem of how idioms are mentally stored.

3. Method

Materials. The idiomatic expressions used in the present work were taken from a study by Tabossi and colleagues (2011), who elicited normative judgments for Italian verbal idioms on the following variables:

- meaning knowledge, the proportion of correct meaning definitions given for each idiom;
- familiarity, the subjective frequency with which speakers encounter an idiom in its written or spoken form, regardless of their familiarity with the actual meaning of the phrase;
- age of acquisition, which indicates at what age the subjects thought they had learnt the expressions;
- predictability, the proportion of idiomatic completions given for a certain idiom, which was presented with the final word missing;
- syntactic flexibility, obtained by asking how much the meaning of the idiom in the syntactically modified version¹ was similar to its unmarked meaning, expressed in the form of a paraphrase;

¹Each idiom was inserted in a sentence containing one of the following five syntactic modifications: adverb insertion, adjective insertion, left dislocation, passivization and wh-movement.

- literality, the plausibility of a literal interpretation for an idiom²;
- compositionality, obtained by asking how much the component words of the idioms contribute to their overall meaning.

Each idiom was also associated with a length value calculated in words.

Procedure. In order to assess the frequency of content words that compose the idiomatic expressions we calculated their cumulative frequency, namely, the summed frequencies of the individual words divided by the number of words, as in Cronk et al. (1993) and Bonin et al. (2013). Differently from previous studies, we took into account both word-form and lemma frequencies; values were taken from CoLFIS (Bertinetto et al., 2005) and ItWaC (Baroni, Bernardini, Ferraresi, & Zanchetta, 2009).

Moreover, we calculated the overall objective frequency of the expressions, intended as the frequency of co-occurrence of all words that make up the string, by means of ad-hoc queries within ITWaC.

We extracted the occurrence values of the idiomatic expressions in all the inflected forms of the verb (e.g., 'break/broke/breaks/etc. the ice'), by searching for the lemma (e.g., 'to break') and filtering the query by specifying one or more constituents (e.g., 'ice'). We adopted a context window of 7/10 elements (depending on the length of idioms), both to the right and left of the lemma, in order to obtain not only the frequency values of canonical idioms, but also the frequency of any possible syntactic manipulations where the order of presentation of the elements is modified (as in passive form, e.g., 'the ice was broken') or other lexical elements are inserted (as in adjective/adverb insertion, e.g., 'he has suddenly broke the ice', etc.). The results of each query were manually checked in order to eliminate casual co-occurrences (as instance, the sentence *la macchina si rompe con il ghiaccio*, 'the car broke because of the ice' contains all words adopted as filters but does not correspond to the given idiomatic expression).

An example of a query is reported in Figure 1.

Figure 1. An example of query in ItWaC

(The idiomatic expression *rompere il ghiaccio* ('break the ice') is searched by filtering for the lemma *rompere* (to break) and the word-form *ghiaccio* (ice), within a context window of 7 tokens, both to the right and the left of the lemma)

4. Results

Data are now available for 124 idiomatic expressions with different degrees of length.

For each idiom, we collected several frequency values:

- Total frequency of idioms;
- Frequency of idioms occurring in a canonical form;
- Frequency of idioms occurring in a transformed form;
- Frequency in CoLFIS of word-forms and lemmas related to content-words appearing in idioms;
- Frequency in ItWaC of word-forms and lemmas related to content-words appearing in idioms.

Table 1 shows the means and the range of all frequency values calculated.

	means	range
TotFq	2,4	0-27
CanonFq	1,9	0-19
VariedFq	0,5	0-9
%varied	23%	0-100%
Ff CoLFIS	1.218	17 - 23.322
Fl CoLFIS	6.939	28 - 72.546
Ff ItWAC	281.642	3.741 - 4.512.480
Fl ItWAC	1.813.494	7.618 - 9.700.850

Table 1: Descriptive statistics (means and range) for the set of 124 idioms. **TotFq**=total frequency of idioms; **CanonFq**=frequency of canonical idioms; **VariedFq**=frequency of manipulated idioms; **FfCoLFIS**=word-form frequency in CoLFIS; **FlCoLFIS**=lemma

²For instance, *perdere il treno* "to miss the boat" (lit. "to miss the train") has also a clear literal meaning beside the figurative one, while *andare in rosso* "to go into the red" does not have a plausible literal meaning and can only be idiomatically interpreted.

frequency in CoLFIS; **FfItWaC**=word-form frequency in ItWaC; **FICoLFIS**=lemma frequency in ItWaC

Hereafter, we report some examples of very frequent idioms:

[1] *Cantar vittoria*, ‘to sing victory’

[2] *Guardarsi allo specchio*, ‘to look in a mirror’

and some examples of infrequent idioms:

[3] *Passare la misura* ‘to cross the line’

[4] *Avere ancora i denti da latte*, ‘to still have baby teeth’

For each idiom, all context occurrences are available in an Excel file. For ambiguous idioms (e.g., *break the ice*), we computed the frequency of all uses, both idiomatic and literal. Data about the syntactic flexibility of each idiom (the percentage of manipulations and the types of manipulation) can also be extracted. In this way, it will be possible for future research to obtain detailed information about the syntactic behavior of each idiomatic expression. Moreover, by analyzing context occurrences of expressions, it will be possible to disambiguate the figurative vs. literal use of ambiguous idioms, in order to derive objective frequency dominance values, in addition to subjective literal plausibility estimates, which are already available in Tabossi et al. (2011).

Below we report some examples of idioms which rarely occur in a manipulated form (less than 5%):

[5] *Battere la fiacca*, ‘to loaf about’

[6] *Mettere il carro davanti ai buoi*, ‘to put the cart before the horse’

and some examples of much flexible idioms (more than 30%):

[7] *Ingoiare la pillola*, ‘to swallow a bitter pill’

[8] *Mettersi nei panni di qualcuno*, ‘to put yourself in someone’s shoes’.

We carried out some correlational analyses in order to evaluate the relationship among objective frequency values and subjective variables, which are available for this set of idiomatic expressions (Tabossi et al., 2011). Hereafter, we will discuss most interesting results.

Relationship among subjective and objective frequency. As shown by Table 2, there is not a correlation between the frequency values of idioms and the frequency values of content words that compose the expressions: most used idioms are not necessarily made up by frequent words; rather, it often happens that frequent idiomatic expressions are composed by words that

are used predominantly – if not exclusively – within such expressions (e.g., ‘cuoia’ in ‘tirare le cuoia’, ‘pull the skins’). Nevertheless, there are positive correlations between frequency values of words (both taken by CoLFIS and ItWaC) and subjective variables of familiarity and meaning knowledge: in other words, idiomatic expressions which have been rated more familiar and known by speakers are made up by frequent words. Interestingly, more frequent idioms are also more familiar but there is not a correlation between the frequency of idioms and meaning knowledge. We may interpret this finding as an evidence that speakers do not always know the exact meaning of idioms, independently by the fact that they occur very frequently in their language. As regards the frequency of manipulated idioms, we found a positive correlation with the frequency of lemmas (taken by CoLFIS): idioms which more often occur in corpora in a manipulated form are made up by frequent words. As expected, there are strong positive correlations among frequency values of words (both lemmas and word-forms) collected in CoLFIS and ItWaC.

Correlations between objective and subjective frequency								
	2	3	4	5	6	7	8	9
1.TotFq	.99***	.87***	-.01	-.02	-.03	-.01	-.04	.21***
2.CanonFq		.77***	-.03	-.05	.01	-.06	.01	.23***
3.VariedFq			.06	.19**	.14	.16	.01	.09
4.Ff CoLFIS				.71***	.76***	.62***	.21***	.14
5.FI CoLFIS					.78***	.94***	.24***	.18***
6.Ff ItWAC						.74***	.24***	.18**
7.FI ItWAC							.26***	.20***
8.Know								.45***
9.Famil								1.00

Table 2. TotFq=total frequency of idioms; CanonFq=frequency of canonical idioms; VariedFq=frequency of manipulated idioms; FfCoLFIS=word-form frequency in CoLFIS; FICoLFIS=lemma frequency in CoLFIS; FfItWaC=word-form frequency in ItWaC; FICoLFIS=lemma frequency in ItWaC; Know=meaning knowledge; Famil=familiarity

Relationship among objective frequency values and psycholinguistic variables. As reported in Table 3, there is a negative correlation between the frequency and the age of acquisition of idioms: the idiomatic expressions acquired earlier are also the most frequent in corpora. Also, more frequent idioms are the shorter ones (negative correlation with the length, even in the case of manipulated idioms). Interestingly, all frequency values of words correlate negatively with literality: idioms containing frequent words have been judged less literally plausible by speakers.

Correlations between objective frequency and psycholinguistic variables						
	Length	AoA	Pred	Flex	Lit	Comp
1.TotFq	-.39***	-.21***	-.07	.05	.04	-.06
2.CanonFq	-.39***	-.22***	-.05	.04	.03	-.07
3.VariedFq	-.32***	-.13	-.10	.07	.05	-.02
4.Ff CoLFIS	.21***	-.04	-.04	.12	-.25***	-.05
5.Fl CoLFIS	.10	-.12	-.12	.17	-.29***	-.05
6.Ff ItWAC	.19	-.11	.03	.16	-.19***	-.03
7.Fl ItWAC	.10	-.15	-.13	.17	-.30***	-.06

Table 3. TotFq=total frequency of idioms; CanonFq=frequency of canonical idioms; VariedFq=frequency of manipulated idioms; FfCoLFIS=word-form frequency in CoLFIS; FlCoLFIS=lemma frequency in CoLFIS; FfItWAC=word-form frequency in ItWAC; FlCoLFIS=lemma frequency in ItWAC; Length=number of words; AoA=age of acquisition; Pred=predictability; Flex=syntactic flexibility; Lit=literality

5. Conclusions

In the present study, we pursued the main goal of collecting objective frequency values of idioms and evaluating their relation with a set of subjective variables available for Italian idiomatic (Tabossi et al., 2011). The novelty of our methodology allowed us to obtain corpus-based frequency values not only for content-words composing idioms (as reported in other normative data available for other languages, e.g., Cronk et al., 1993; Libben & Titone, 2008; Bonin et al., 2013), but also for idioms considered in their entirety. Furthermore, frequency values took into account also the occurrences of syntactically manipulated idioms (passive form, left dislocation, etc.).

The possibility of having objective frequency values of idiomatic expression can be an important support for directing future research on idiom processing. Recent psycholinguistic studies (e.g., Tabossi, Fanari, & Wolf, 2009) have questioned the hypothesis that the so-called 'idiom superiority effect' - namely, the established fact that idiomatic expressions are faster to process than literal sentences - is due to the idiomaticity itself of the expressions. According to the authors, the phenomenon could depend, more simply, on the fact that the idiomatic expressions adopted in most of the existing experimental studies were much more familiar than the literal sentences of control to which they were compared, which, in many cases, were completely new expressions, obtained by manipulating in part the idiomatic expressions of origin. A possible continuation of these studies could involve the implementation of experiments, in which idiomatic and literal expressions are matched for the objective frequency of occurrence, as well as

a series of other well-known parameters. Moreover, studies aiming to explore the syntactic behavior of idioms might rely on objective frequency values of idioms occurring in a non-canonical form and explore the type and the percentage of manipulations for each idiomatic expression.

Acknowledgments

The authors would like to thank Simonetta Vietri for her constructive comments and recommendations on an earlier version of the paper and to Annibale Elia for his constant support to our research work.

Reference

- Abel, B. (2003). English idioms in the first language and second language lexicon: A dual representation approach. *Second language research*, 19(4), 329-358.
- Balota, D. A., Pilotti, M., & Cortese, M. J. (2001). Subjective frequency estimates for 2,938 monosyllabic words. *Memory & Cognition*, 29(4), 639-647.
- Barca, L., Burani, C., & Arduino, L. S. (2002). Word naming times and psycholinguistic norms for Italian nouns. *Behavior Research Methods, Instruments, & Computers*, 34(3), 424-434.
- Barlow, M. (2000). Usage, blends and grammar. *Usage-based models of language*, 315-345.
- Baroni, M., Bernardini, S., Ferraresi, A., & Zanchetta, E. (2009). The WaCky wide web: a collection of very large linguistically processed web-crawled corpora. *Language resources and evaluation*, 43(3), 209-226.
- Bertinetto, P. M., Burani, C., Laudanna, A., Marconi, L., Ratti, D., Rolando, C., & Thornton, A. M. (2005). *Corpus e Lessico di Frequenza dell'Italiano Scritto (CoLFIS)*. <http://linguistica.sns.it/CoLFIS/Home.htm>
- Bonin, P., Méot, A., & Bugaiska, A. (2013). Norms and comprehension times for 305 French idiomatic expressions. *Behavior research methods*, 45(4), 1259-1271.
- Brysbaert, M., Buchmeier, M., Conrad, M., Jacobs, A. M., Bölte, J., & Böhl, A. (2011). The word frequency effect. *Experimental psychology*.
- Cacciari, C., & Tabossi, P. (1988). The comprehension of idioms. *Journal of memory and language*, 27(6), 668-683.
- Cacciari, C., & Tabossi, P. (2014). *Idioms: Processing, structure, and interpretation*. Psychology Press.
- Caillies, S. (2009). Descriptions de 300 expressions idiomatiques: Familiarité, connaissance de leur signification, plausibilité littérale, "décomposabilité" et "prédicibilité". *L'Année Psychologique*, 109, 463-508.
- Citron, F. M., Cacciari, C., Kucharski, M., Beck, L., Conrad, M., & Jacobs, A. M. (2016). When emo-

- tions are expressed figuratively: Psycholinguistic and Affective Norms of 619 Idioms for German (PANIG). *Behavior research methods*, 48(1), 91-111.
- Coltheart, M., Rastle, K., Perry, C., Langdon, R., & Ziegler, J. (2001). DRC: a dual route cascaded model of visual word recognition and reading aloud. *Psychological review*, 108(1), 204.
- Cronk, B. C., Lima, S. D., & Schweigert, W. A. (1993). Idioms in sentences: Effects of frequency, literalness, and familiarity. *Journal of Psycholinguistic Research*, 22(1), 59-82.
- Cutting, J. C., & Bock, K. (1997). That's the way the cookie bounces: Syntactic and semantic components of experimentally elicited idiom blends. *Memory & Cognition*, 25(1), 57-71.
- De Martino, M., Mancuso, A., & Laudanna, A. (2017). Variabili Rilevanti nella Rappresentazione delle Parole nel Lessico Mentale: Dati Psicolinguistici da una Banca-Dati di Nomi e Verbi Italiani. In Basili, R., Nissim, M., & Satta, G. (Eds.), *Proceedings of the Fourth Italian Conference on Computational Linguistics CLiC-it 2017: 11-12 December 2017, Rome*. Torino: Accademia University Press.
- Geeraert, K., Baayen, R. H., & Newman, J. (2017). Understanding idiomatic variation. In *Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017)* (pp. 80-90).
- Gibbs, R. W. (1980). Spilling the beans on understanding and memory for idioms in conversation. *Memory & Cognition*, 8(2), 149-156.
- Harm, M. W., & Seidenberg, M. S. (2004). Computing the meanings of words in reading: cooperative division of labor between visual and phonological processes. *Psychological review*, 111(3), 662.
- Janssen, N., Pajtas, P. E., & Caramazza, A. (2011). A set of 150 pictures with morphologically complex English compound names: Norms for name agreement, familiarity, image agreement, and visual complexity. *Behavior Research Methods*, 43(2), 478-490.
- Konopka, A. E., & Bock, K. (2009). Lexical or syntactic control of sentence formulation? Structural generalizations from idiom production. *Cognitive Psychology*, 58(1), 68-101.
- Kyriacou, M., Conklin, K., & Thompson, D. (2019). Passivizability of Idioms: Has the Wrong Tree Been Barked Up?. *Language and speech*. <https://doi.org/10.1177/0023830919847691>
- Langlotz, A. (2006). *Idiomatic creativity: A cognitive-linguistic model of idiom-representation and idiom-variation in English* (Vol. 17). John Benjamins Publishing.
- Libben, M. R., & Titone, D. A. (2008). The multiterminated nature of idiom processing. *Memory & Cognition*, 36(6), 1103-1121.
- Mancuso, A., Elia, A., Laudanna, A., & Vietri, S. (2019). The Role of Syntactic Variability and Literal Interpretation Plausibility in Idiom Comprehension. *Journal of Psycholinguistic Research*, <https://doi.org/10.1007/s10936-019-09673-8>
- Montefinese, M., Ambrosini, E., Fairfield, B., & Mammarella, N. (2014). The adaptation of the affective norms for English words (ANEW) for Italian. *Behavior research methods*, 46(3), 887-903.
- Moon, R. (1998). *Fixed expressions and idioms in English: A corpus-based approach*. Oxford University Press.
- Nordmann, E., & Jambazova, A. A. (2017). Normative data for idiomatic expressions. *Behavior research methods*, 49(1), 198-215.
- Sprenger, S. A., Levelt, W. J., & Kempen, G. (2006). Lexical access during the production of idiomatic phrases. *Journal of memory and language*, 54(2), 161-184.
- Swinney, D. A., & Cutler, A. (1979). The access and processing of idiomatic expressions. *Journal of verbal learning and verbal behavior*, 18(5), 523-534.
- Tabossi, P., Arduino, L., & Fanari, R. (2011). Descriptive norms for 245 Italian idiomatic expressions. *Behavior Research Methods*, 43(1), 110-123.
- Tabossi, P., Fanari, R., & Wolf, K. (2009). Why are idioms recognized fast? *Memory & Cognition*, 37(4), 529-540.
- Tabossi, P., Wolf, K., & Koterle, S. (2009). Idiom syntax: Idiosyncratic or principled?. *Journal of Memory and Language*, 61(1), 77-96.
- Titone, D. A., & Connine, C. M. (1994). Descriptive norms for 171 idiomatic expressions: Familiarity, compositionality, predictability, and literality. *Metaphor and Symbol*, 9(4), 247-270.
- Vietri, S. (2014). *Idiomatic constructions in Italian: a lexicon-grammar approach* (Vol. 31). John Benjamins Publishing Company.

Gender Detection and Stylistic Differences and Similarities between Males and Females in a Dream Tales Blog

Raffaele Manna

UNIOR NLP Research Group
University L'Orientale
Naples, Italy
rmanna@unior.it

Antonio Pascucci

UNIOR NLP Research Group
University L'Orientale
Naples, Italy
apascucci@unior.it

Johanna Monti

UNIOR NLP Research Group
University L'Orientale
Naples, Italy
jmonti@unior.it

Abstract

English. In this paper we present the results of a gender detection experiment carried out on a corpus we built downloading dream tales from a blog. We also highlight stylistic differences and similarities concerning lexical choices between men and women. In order to carry the experiment we built a feed-forward neural network with traditional sparse n-hot encoding using the Keras open source library.

1 Introduction

It is generally accepted that dreams are just an unconscious production, and that represent a type of non-manipulable happening. However, many people believe that dreams are premonitory of future events as well as representations and reworkings of past events. Humans tend to preserve all personal events, some of them in the form of a diary, namely the best method to tell an event and keep its aura of magic.

Until recently, dream reports were relegated to the pages of paper journals or revealed to familiar people. At an earlier time, dreams are gathered from sleep research labs, psycho-therapeutic and in patient settings, personal dream journals and occasionally classroom settings where “most recent dreams” and “most vivid dreams” are collected as in (Domhoff, 2003).

Social media have opened millions of pages where people feel at ease to confess their thoughts, their experience and even their secret fantasies. These platforms such as Twitter, Facebook and web blogs are a good ground for computational text analysis research in social science and mental health assessment via language.

Diary narratives represent a field already investigated by researchers. The recent development of web communities focused on telling dreams allows researchers to access and discover new characteristics related to the language of dreams. Stylistic and linguistic features of dreams in blog reports are essential in order to detect writing style and content differences between men and women, but also enable future researches associated to the different types of personality and styles associated with mental health diagnoses and therapeutic outcomes.

The aim of this paper is to show that despite dreams are just an unconscious production, there are several stylistic differences between the reports of dreams by males and females on online blogs. The model we built is able to represent and classify all stylistic differences.

Moreover, this research represents a preliminary step in the field of dream tales which will be followed by an attempt to find stylistic differences between dream tales and other forms of self narration (i.e. travel tales).

The paper is organized as follows: in Section 2 we introduce Related Work, in Section 3 we describe the corpus we built and the blog. Methodology is described in Section 4 and Results are in Section 5. In Section 6 we present our Conclusions and we introduce Future Work.

2 Related Work

Textual analysis of dream reports is still not a completely investigated field in NLP. One of the purposes of computational dream report analysis lies in understanding how and why a dream narrative differs from a waking narrative (Hendrickx et al., 2016). For example, if a dream description contains more function words than a waking narrative, what is the relationship between the content of dreams and the use of more function words?

Earlier studies were conducted by (Domhoff, 2003

Copyright © 2019 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

and Bulkeley, 2009). In their researches, dream reports are analyzed and a systematic category list of words that can be used for queries and word-frequency counts in the DreamBank.net is provided. The categories are related to the content of dreams and used to retrieve the mentions of emotions, characters, perception, movement and socio-cultural background.

On the basis of this approach (Bulkley, 2014) update the categories list and evaluate it on four datasets of the *DreamBank* corpus. It has been shown that this type of word analysis can be applied to detect the topics of dreams. In addition, this latter contribution provides evidence that it is possible to guess about a person's life and activities, personal concerns and interests based on an individual dream collection.

Other works focus on identifying the emotions in the reports of dreams. In particular (Razav et al., 2014) use a machine learning method to assign emotion labels to dreams on a four-level negative/positive sentiment scale. In their research, dreams are represented as word vectors and dynamic features are included to represent sentiment changes in dream descriptions.

In a more accurate sentiment analysis, (Frantova and Bergler, 2009) train a classifier, based on semi-automatically compiled emotion word dictionaries, in order to assign five fuzzy-emotion categories to dream reports. Then, they compare their results against a sample from the *DreamBank* that is manually labeled with emotion annotations.

In some non-computational studies and aimed at highlighting gender differences (Schredl, 2005; Schredl, 2010), dream reports are used to spot gender differences in dream recall. The first research demonstrates that gender differences in dream recalls and dream contents are stable. Human judges are able to correctly match the dreamer's gender based on a single dream report with a probability better than chance. Based on these findings, in the latter study the stability of gender differences in dream content is analyzed over time. Two dream themes (work-related dreams and dreams of deceased persons) were investigated and gender differences resulted quite stable over time. In (Mathes, 2013) gender differences are associated to personality traits. The analysis indicate that some of the big five personality dimensions might be linked with some dream characteristics such as characters and the occurrence of weapons or

clothes in dreams.

In psychiatric studies, the gender variable is identified as a predictive for psychotic behaviors and disorders. In (Thorup, et al., 2007), the authors showed that, in psychotic patients, the gender-related variable has a role in showing different psycho-pathological characteristics and different social functioning. Although no dream samples were taken as a subject in this study.

Dream diaries refine the research in uncovering connections between dreams and dreamer's socio-cultural background, mental conditions and neuro-physiological factors. The language of online dreams in relation to mental health conditions has yet to be analyzed, however prior laboratory research suggests that dream content may differ according to clinical conditions.

In (Skancke et al., 2014), emotional tone, themes and actor focus in dream report were associated with anxiety disorders, schizophrenia, personality and eating disorders. However, it is not clear whether dream content can be predictive with respect to mental disorders.

In (Scarone, 2008), the hypothesis of the dreaming brain as a neurobiological model for psychosis is tested by focusing on cognitive bizarreness, a distinctive property of the dreaming mental state defined by discontinuities and incongruities in the dream report, thoughts and feelings. Cognitive bizarreness is measured in written reports of dreams and in verbal reports of waking fantasies in thirty schizophrenics and thirty normal controls. The differences between these two groups indicate that, under experimental conditions, the waking cognition of schizophrenic subjects shares a common degree of formal cognitive bizarreness with dream reports of both normal controls and schizophrenics. These results support the hypothesis that dreaming brain could be a useful experimental model for psychosis. Taking advantage of all the above considerations and mixing the psychiatric and neurobiological information of the studies shown, the present research wants first of all to reveal the differences between genders in dreams. And as a future goal, starting from the hypothesis of cognitive similarity between dreams and psychoses and using dreams as an experimental path, to clarify the relationship between gender and psychosis.

3 Dataset Description

The web is full of blogs, where people can share opinions, questions and personal feelings and thoughts about their own life. Furthermore, people also share their dreams, one of the most personal hidden aspects of life.

It is very easy to find a blog in which thousands of people share their “dream experiences”, sometimes discovering that other people have had similar experiences dictated by similar life styles.

We investigated a blog, called *SogniLucidi*, on which every day thousands of people tell their dreams and nightmares, mixing their nightly fantasies with their unconscious writing style choices. *SogniLucidi*, that literally can be translated in *LucidDreams* took its name from a term coined by the Dutch psychiatrist Frederik van Eeden in 1913: it describes the situation in which dreamers are aware that they are dreaming.

There are many techniques that, when correctly applied, allow dreamers to obtain a “Lucid Dream” and that we report for completeness: **CAT** (*Cycle Adjustment Technique*), **MILD** (*Mnemonic Induction of Lucid Dreaming*), **WBTB** (*Wake Back To Bed*), **WILD** (*Wake Initiated Lucid Dreams*), **RCT** (*Reality Control Test*) and **ITES** (*Induction Through External Stimulus*).

The corpus we built for the investigation is balanced with gender and the number of authors analyzed is not randomly selected but represents the precise number of participants to the blog.

3.1 Dataset Statistics

In this paragraph, we present the resulting statistics obtained using the NLTK module together with other statistics formulas for the analysis of the corpus we built on *SogniLucidi* blog. In Table 1 we report two important statistics about words: the number of tokens in texts written by men and women and word types. We can notice that there is a big difference in the number of tokens used by Males (80629) and Females (57673).

	Males	Females
Number of Tokens	80629	57673
Word Types	12254	11158

Table 1: Words’ statistics in the whole corpus in terms of Number of Tokens and Word Types.

In Tables 2 and 3 we present four lists of six exclusive nouns and six exclusive verbs used by men or women. Both exclusive nouns and exclusive verbs are the most relevant for frequency for Males and Females classes. Verbs are reported in their base form. The results indicate, without interpretative effort for a human, that most relevant topics given these high frequency words are associated to activities and events that the dreamers want to happen, in settings and adventurous situations for male dreamers. Meanwhile dreamers belonging to Females class seem to set their dreams in a baleful scenario, where “transizione” (transition) and “trapasso” (transition) mean that they dream about twilight state, beyond death or they fantasize about surreal activities.

Males	Females
destinazione (destination)	balzo (bound)
esplosione (explosion)	luce (light)
foresta (wood)	nuvola (cloud)
lenzuola (linens)	piscina (swimming pool)
spiaggia (beach)	transizione (transition)
terrazze (terraces)	trapasso (transition)

Table 2: Most frequent Exclusive Nouns in the whole corpus.

Males	Females
assomigliare(to resemble)	affrontare(to face)
baciare(to kiss)	cadere(to fall)
funzionare(to function)	ragionare(to reason)
ottenere(to obtain)	stringere(to tighten)
scomparire(to disappear)	succedere(to happen)
superare(to overcome)	volare(to fly)

Table 3: Most frequent Exclusive Verbs in the whole corpus.

Lastly, in Table 4 we report the average of tokens per sentence.

Males Tokens AVG	Females Tokens AVG
18,74 tokens/sentence	10,01 tokens/sentence

Table 4: Average of tokens per sentence in texts written by men and women.

4 Methodology

The training corpus consists in dream text descriptions written by two groups of authors:

- 28 Male authors;
- 28 Female authors.

The corpus is balanced and labelled with gender. Gender annotation has been done manually and based on the name of the users, their profile photos and description. For each author, a total of fifteen texts about dreams are provided. Authors are coded with an alpha-numeric author-ID. For each author, the last fifteen texts about dreams have been retrieved from the personal web diary's timeline. As a result, the time frame of the dream reports might vary from days to months, depending on how frequently users report their dreams on the blog. To train our classification model, we exploited the descriptions of dreams only and not the comments (both comments of the authors and comments of other members of the *SogniLucidi* blog).

4.1 Preprocessing

For preprocessing we used the Python library *BeautifulSoup* along with same regex procedures. We performed the following preprocessing steps:

- Removing the html tags;
- Removing URLs;
- Removing @username mentions;
- Lower-casing the characters;
- Detecting stop-words by document frequency and removing. Only n-grams that occurred in all documents has been considered a stop-word and ignored.

4.2 Features

Feature selection is a very critical step in any model. For feature selection we use the sklearn utilities *SelectKbest*. It selects the n-best feature based on a given criterion. In our experiments, the features are selected on the *f_classif* criteria. This function perform an ANOVA test, a type of hypothesis test, on each feature on its own and assign that feature a p-value. The SelectKbest rank the features by that p-value and keep only the n-best features. The feature set for the dream dataset benefits from word trigrams in addition to other n-grams. In our final model, we use the following n-grams features: Word unigrams, bigrams and trigrams.

Word level n-grams used the following parameters:

- Minimum document frequency = 2. Terms with a document frequency lower than would be ignored;
- Term frequency-inverse document frequency (tf-idf) weighting;
- Maximum document frequency = 1.0 or rather terms that occur in all documents would be ignored.

4.2.1 Classification Model

We built a neural network to perform the gender detection issue. We decided to run a feed-forward neural network with traditional sparse one-hot encoding with the *Keras* open source library. After a parameters selection, the model obtained the best performance with an Adam optimizer and a learning rate of 0.32, feeding it with a batch size of seventy and training for thirty epochs. Moreover, the input layer of sixty-five neurons with an initialization using a norm kernel. Then, a RELU activation function was applied, followed by a dropout layer. During optimization, we found that a relatively big dropout rate of 0.5 outperformed the smaller dropout rates. The output layer is a single neuron, followed by a linear activation function. The feature set provided to the model was an n-hot encoding of the uni-, bi- and trigrams.

5 Results

In this section we describe the results on the training data and the test data. The data we used was split into training and test data. The training set contains a known output and the model learns on this data in order to be generalized to other data later on. We have the test set (or subset) in order to test our model+ prediction on this subset. We calculated accuracy scores on the training data, both on validation set (Dev set) of 0.3 and Test set of 0.2. The performances (both for Dev test and Test set) are shown in Table 5 in terms of Accuracy, Precision and F1 Score. We obtained roughly the same results for Accuracy in Dev set and the Test set, 0.794 and 0.775, respectively.

Finally, in order to compare our approach, we considered two other baseline models namely Multinomial Naive Bayes (MNB) and Linear Support Vector Machine (SVM) besides the feed-forward

	Dev set	Test set
Accuracy	0.796	0.776
Precision	0.937	0.917
F1 Score	0.803	0.786

Table 5: Performances in Dev set and Test set in terms of Accuracy, Precision and F1 Score.

neural network for performance comparisons on Test set.

MNB	SVM
0.411	0.588

Table 6: Baseline Accuracy Comparisons.

To assess the performance of the model, the Root Mean Square Error (RMSE) was computed. RMSE measures the distance of the predicted value to the true value. It is a measure of error, so the lower is the score, the better is the performance. We show RMSE results in Table 7.

Dev set	Test set
0.233	0.224

Table 7: RMSE of the feed-forward model on the Dev set and when using Test set.

Using classification accuracy alone when evaluating the performance of the classification algorithm could be misleading, especially if the dataset - as in our case - is limited in size or is unbalanced or contains more than two classes. Hence, a confusion matrix is used to evaluate the results of the experiments. The confusion matrix M is a N -dimensional matrix, where N is the number of classes, that summarizes the classification performance of a classifier with respect to Test set and Dev set, both as in our case. Each column of the matrix represents predicted classifications and each row represents actual defined classifications. As shown in Table 8, during the validation phase, the classifier made a total of two hundred-sixteen predictions, while during the test phase the classifier made a total of two hundred-fourteen predictions. Out of two hundred-sixteen cases in validation, the classifier predicted “Females” forty-four times and sixty-four “Males”. Actually, sixty people in the sample belong to “Females” class and forty-eight

to “Males” class.

	Males	Females
Males	45	3
Females	19	41

Table 8: Confusion Matrix on Dev set.

After this intermediate phase and after having tuned the parameters in order to optimize the model on the previous results, the classifier made a total of two hundred-fourteen predictions during the test phase. Out of two hundred-fourteen predictions, the model predicted “Females” forty-three times and sixty-four “Males”. Indeed, fifty-nine people belong to “Females” class and, as predicted during the validation phase, forty-eight to “Males” class. We report gender prediction results on test data in the confusion matrix in Table 9.

	Males	Females
Males	44	4
Females	20	39

Table 9: Confusion Matrix on Test set.

6 Conclusions and Future Work

In this paper we have shown our results on gender detection in dream diaries and writing styles differences and similarities between males and females in dream tales. First we explored the vocabulary of dream descriptions for both the genre-class by listing some of the representative words for each genre. Then, we evaluated our gender detection model on the dream reports dataset. The model succeeded in obtaining good results managing to distinguish a good part of dreams made by men or women. This research represents our preliminary step in the field, toward subsequent studies, in which we are trying to detect stylistic differences between dream tales and personal descriptive narratives, such as travel tales and other forms of self-narration.

Acknowledgments

This project has been partially supported by the PON Ricerca e Innovazione 2014/20 and the POR Campania FSE 2014/2020 funds.

References

- Altszyler, E., Sigman, M., Ribeiro, S., Slezak, D. F. (2016). Comparative study of LSA vs Word2vec embeddings in small corpora: a case study in dreams database. *arXiv preprint arXiv:1610.01520*.
- Altszyler, E., Ribeiro, S., Sigman, M., Slezak, D. F. (2017). The interpretation of dream meaning: Resolving ambiguity using Latent Semantic Analysis in a small corpus of text. *Consciousness and cognition*, 56, 178-187.
- Bulkeley, K.. (2009). Seeking patterns in dream content: A systematic approach to word searches. *Consciousness and cognition*, 18(4), 905-916.
- Bulkeley, K.. (2014). Digital dream analysis: A revised method. *Consciousness and cognition*, 29, 159-170.
- Coelho, H.. (2010). Classification of dreams using machine learning. In *ECAI: 19th European Conference on Artificial Intelligence: Including Prestigious Applications of Artificial Intelligence (PAIS-2010): Proceedings (Vol. 215, p. 169)*.
- Domhoff, G. W.. (2003). *The scientific study of dreams: Neural networks, cognitive development, and content analysis*. American Psychological Association.
- Domhoff, G. W., Schneider, A.. (2008). Similarities and differences in dream content at the cross-cultural, gender, and individual levels. *Consciousness and cognition*, 17(4), 1257-1265.
- Frantova, E., Bergler, S.. (2009). Automatic emotion annotation of dream diaries. In *Proceedings of the analyzing social media to represent collective knowledge workshop at K-CAP 2009, The fifth international conference on knowledge capture*.
- Hawkins, I. I., Raymond, C., Boyd, R. L. 2017. *Such stuff as dreams are made on: Dream language, LIWC norms, and personality correlates*, *Dreaming*, 27(2), 102.
- Hendrickx, I., Onrust, L., Kunneman, F., Hriyetolu, A., Bosch, A. V. D., Stoop, W. 2016. Unraveling reported dreams with text analytics. *arXiv preprint arXiv:1612.03659*.
- Koppel, M., Argamon, S., Shimon, A. R., 2002. *Automatically categorizing written texts by author gender. Literary and linguistic computing*, 17(4), 401-412.
- Mathes, J., Schredl, M.. (2013). Gender differences in dream content: Are they related to personality?. *International Journal of Dream Research*.
- Mechti, S., Jaoua, M., Belguith, L. H., Faiz, R., 2013. *Author profiling using style-based features*, *Notebook Papers of CLEF2*.
- McNamara, P., Duffy-Deno, K., Marsh, T.. (2019). Dream content analysis using Artificial Intelligence. *International Journal of Dream Research*, 42-52.
- Mukherjee, A., Liu, B.. (2010, October). Improving gender classification of blog authors. In *In Proceedings of the 2010 conference on Empirical Methods in natural Language Processing (pp. 207-217)*. Association for Computational Linguistics.
- Niederhoffer, K., Schler, J., Crutchley, P., Loveys, K., Coppersmith, G.. (2017, August). In your wildest dreams: the language and psychological features of dreams. In *Proceedings of the Fourth Workshop on Computational Linguistics and Clinical Psychology From Linguistic Signal to Clinical Reality (pp. 13-25)*.
- Nielsen, T. A., Stenstrom, P. (2005). What are the memory sources of dreaming?. *Nature*, 437(7063), 1286.
- Rangel, F., Rosso, P. 2013. *Use of language and author profiling: Identification of gender and age*, *Natural Language Processing and Cognitive Science*, 177.
- Razavi, A. H., Matwin, S., De Koninck, J., Amini, R. R.. (2014). *Dream sentiment analysis using second order soft co-occurrences (SOSCO) and time course representations*. *Journal of Intelligent Information Systems*, 42(3), 393-413.
- Scarone, S., Manzone, M. L., Gambini, O., Kantzas, I., Limosani, I., D'agostino, A., Hobson, J. A.. (2008). The dream as a model for psychosis: an experimental approach using bizarreness as a cognitive marker. *Schizophrenia Bulletin*, 34(3), 515-522.
- Scarpelli, S., Bartolacci, C., D'Atri, A., Gorgoni, M., De Gennaro, L.. (2019). The functional role of dreaming in emotional processes. *Frontiers in Psychology*, 10.
- Schredl, M., Sahin, V., Schfer, G.. (1998). Gender differences in dreams: do they reflect gender differences in waking life?. *Personality and Individual Differences*, 25(3), 433-442.
- Schredl, M., Ciric, P., Gtz, S., Wittmann, L.. (2004). Typical dreams: stability and gender differences. *The journal of psychology*, 138(6), 485-494.
- Schredl, M., Piel, E.. (2005). Gender differences in dreaming: Are they stable over time?. *Personality and Individual Differences*, 39(2), 309-316.
- Schredl, M., Becker, K., Feldmann, E.. (2010). Predicting the dreamers gender from a single dream report: A matching study in a non-student sample. *International Journal of Dream Research*.
- Schredl, M., Noveski, A.. (2018). Lucid Dreaming: A Diary Study. *Imagination, Cognition and Personality*, 38(1), 517. <https://doi.org/10.1177/0276236617742622>

- Siclari, F., et al. (2017). The neural correlates of dreaming. *Nature neuroscience*, 20(6), 872.
- Silberman, Y., Bentin, S., Miikkulainen, R.. (2007). Semantic Boost on Episodic Associations: An EmpiricallyBased Computational Model. *Cognitive Science*, 31(4), 645-671.
- Skancke, J. F., Holsen, I., Schredl, M.. (2014). Continuity between waking life and dreams of psychiatric patients: a review and discussion of the implications for dream research. *International Journal of Dream Research*.
- Thorup, Anne and Petersen, Lone and Jeppesen, Pia and Ohlenschläger, Johan and Christensen, Torben and Krarup, Gertrud and Jorgensen, Per and Nordentoft, Merete. (2007). Gender differences in young adults with first-episode schizophrenia spectrum disorders at baseline in the Danish OPUS study. *The Journal of nervous and mental disease*, 195(5), 396-405
- Van Eeden, F.. (1913, July). A study of dreams. In *Proceedings of the Society for Psychical Research*. Vol. 26, No. Part 47, pp. 431-461.

CROATPAS: A Resource of Corpus-derived Typed Predicate Argument Structures for Croatian

Costanza Marini

University of Pavia
Department of Humanities
costanza.marini01@
universitadipavia.it

Elisabetta Ježek

University of Pavia
Department of Humanities
jezek@unipv.it

Abstract

The goal of this paper is to introduce CROATPAS, the Croatian sister project of the Italian Typed-Predicate Argument Structure resource (TPAS¹, Ježek et al. 2014). CROATPAS is a corpus-based digital collection of verb valency structures with the addition of semantic type specifications (SemTypes) to each argument slot, which is currently being developed at the University of Pavia. Salient verbal patterns are discovered following a lexicographical methodology called Corpus Pattern Analysis (CPA, Hanks 2004 & 2012; Hanks & Pustejovsky 2005; Hanks et al. 2015), whereas SemTypes – such as [HUMAN], [ENTITY] or [ANIMAL] – are taken from a shallow ontology shared by both TPAS and the Pattern Dictionary of English Verbs (PDEV², Hanks & Pustejovsky 2005; El Maarouf et al. 2014). The theoretical framework the resource relies on is Pustejovsky’s Generative Lexicon theory (1995 & 1998; Pustejovsky & Ježek 2008), in light of which verbal polysemy and metonymic argument shifts can be traced back to compositional operations involving the variation of the SemTypes associated to the valency structure of each verb. The corpus used to identify verb patterns in CROATPAS is the Croatian Web as Corpus (hrWac 2.2, RELDI PoS-tagged) (Ljubešić & Erjavec 2011), which contains 1.2 billion types and is available on the Sketch Engine³ (Kilgarrieff et al.

2014). The potential uses and purposes of the resource range from multilingual pattern linking between compatible resources to computer-assisted language learning (CALL).

1 Introduction

Nowadays, we live in a time when digital tools and resources for language technology are constantly mushrooming all around the world. However, we should remind ourselves that some languages need our attention more than others if they are not to face – to put it in Rehm and Hegelesevere’s words – “a steadily increasing and rather severe threat of digital extinction” (2018: 3282).

According to the findings of initiatives such as the META-NET White Paper Series (Tadić et al. 2012; Rehm et al. 2014), we can state that Croatian is unfortunately among the 21 out of 24 official languages of the European Union that are currently considered *under-resourced*. As a matter of fact, Croatian “tools and resources for [...] deep parsing, machine translation, text semantics, discourse processing, language generation, dialogue management simply do not exist” (Tadić et al. 2012: 77). An observation that is only strengthened by the update study carried out by Rehm et al. (2014), which shows that, in comparison with other European languages, Croatian has *weak to no support* as far as text analytics technologies go and only *fragmentary support* when talking of resources such as corpora, lexical resources and grammars. In this framework, a semantic resource such as CROATPAS could play its part not only in NLP, (e.g. multilingual pattern linking between other existing compatible resources), but also in automatic machine translation, computer-assisted

¹ <http://tpas.fbk.eu> (last visited on July 12th 2019)

² <http://pdev.org.uk> (last visited on July 12th 2019)

³ <https://www.sketchengine.eu/> (last visited on July 12th 2019)

language learning (CALL) and theoretical and applied cross-linguistic studies.

The paper is structured as follows: first a detailed overview of the resource is presented (Section 2), followed by its theoretical underpinnings (Section 3) and a summary of the Croatian-specific challenges we faced while building the resource editor (Section 4). An overview of the existing related works is given in Section 5. Finally, Section 6 hints at the creation of a multilingual resource linking CROATPAS, TPAS (Italian) and PDEV (English) patterns and explores CROATPAS's potential for computer-assisted L2 teaching and learning.

2 Resource overview

CROATPAS, i.e. the Croatian Typed-Predicate Argument Structure resource, is the Croatian equivalent of the Italian TPAS resource (Ježek et al. 2014) and is a corpus-derived collection of Croatian verb argument structures, whose argument slots have been annotated using semantic type specifications (SemTypes).

The first version of the resource is currently being developed at the University of Pavia with the technical assistance of *Lexical Computing Ltd.* in the person of Vit Baisa and will be released in 2020 through an Open Access graphical user interface on the website of the Language Centre of the University of Pavia (CLA)⁴.

CROATPAS contains a sample of 100 medium-frequency Croatian verbs, whose Italian translational counterparts are already available in the TPAS resource: 26 of these verbs are Croatian translational equivalents of Italian “coercive verbs”, i.e. verbs that instantiate metonymic shifts in one of their senses (Ježek & Quochi 2010), while the remaining 74 are Croatian translational equivalents of a sample of Italian *fundamental verbs*, i.e. verbs belonging to that group of approximately 2000 lexemes deemed essential for communicating in Italian and that can be found in any sort of text (De Mauro 2016).

Our 74-verbs sample was selected as follows: we first extracted the frequency counts for all the 452 fundamental verbs on De Mauro's list from a reduced version of the ItWAC (Baroni & Kilgarriff, 2006), which contains over 900 million tokens and is available on the Sketch Engine (Kilgarriff et al. 2014). We then selected

our 74 Italian candidates around the median frequency value after taking out the first and the last 20 verbs on the list. Finally, the Croatian translational equivalents for these verbs were chosen using the 2017 Zanichelli Italian/Croatian bilingual dictionary *Croato compatto*, edited by Aleksandra Špikić.

The theoretical framework the resource relies on is Pustejovsky's Generative Lexicon theory (1995 & 1998; Pustejovsky & Ježek 2008), in light of which verbal polysemy and metonymic shifts can be traced back to compositional operations involving the contextual variation of the SemTypes associated to the valency structure of each verb.

CROATPAS rests on four key-components, namely:

- 1) a representative corpus of Croatian;
- 2) a shallow ontology of SemTypes;
- 3) a methodology for corpus analysis;
- 4) adequate corpus tools.

As for the first component, the corpus used to identify verb patterns is the Croatian Web as Corpus (hrWac 2.2, RELDI PoS-tagged) (Ljubešić & Erjavec, 2011), containing 1.2 billion types and available on the Sketch Engine (Kilgarriff et al. 2014). We chose to work with the Croatian Web as Corpus since the reference corpus for the Italian TPAS resource is a reduced version of the Italian Web as Corpus (Baroni & Kilgarriff, 2006), so as to make the two resources as comparable as possible.

As for the shallow ontology of Semantic Type labels, CROATPAS is based on the same hierarchy shared by TPAS and the PDEV project of 180 SemTypes, which originates from the *Brandeis Shallow Ontology* (BSO) (Pustejovsky et al. 2004) and its initial 65 labels. As pointed out by Ježek (2014: 890), SemTypes “are not abstract categories but semantic classes discovered by generalizing over the statistically relevant list of collocates that fill each position”. For example, the Croatian lexical set for the SemType [BEVERAGE] in the context of the verb pair PITI/POPITI (= TO DRINK, *imperfective/perfective*) contains, among others: {vodu = water, kavu = coffee, koktel = cocktail, vino = wine, čaj = tea, pivo = beer, limonadu = lemonade}, as shown in the following pattern string from the resource.

[Human]_{NON} pije [Beverage]_{ACC} {vodu = water, kavu = coffee}

Figure 1 – One of the pattern strings of PITI

⁴ https://cla.unipv.it/?page_id=53723 (last visited on July 12th 2019)

The corpus analysis methodology used for both TPAS and CROATPAS is a lexicographical methodology called Corpus Pattern Analysis (CPA, Hanks 2004 & 2012; Hanks & Pustejovsky 2005; Hanks et al. 2015), which is based on the Theory of Norms and Exploitations (TNE, Hanks 2004, 2013). TNE divides word uses in two main classes: conventional uses (*norms*) and deviations from the norms (*exploitations*). CPA's potential lies in that it does not try to identify meaning in isolation, but rather associates it with prototypical contexts, thus focusing on the norms. The standard CPA procedure requires:

- 1) sampling concordances for each verb
- 2) identifying its typical patterns – i.e. senses – while going through the corpus lines
- 3) assigning SemTypes to the argument slots in each pattern
- 4) assigning the sampled concordance lines to the identified patterns

This last operation is possible because both the TPAS and CROATPAS editors are linked to their respective language-specific corpora through the Sketch Engine (Kilgariff et al. 2014), which proves once again to be the perfect tool for lexicographic work.

The resource will be evaluated through IAA on pattern identification for a sub-sample of the verb inventory, following the methodology proposed by Cinkova et al. (2012).

3 Generative Lexicon Theory

As pointed out by Hanks (2014: 1), the CPA methodology relies theoretically on the Theory of Norms and Exploitations (TNE), which has its roots in Sinclair's work, but is also influenced by Pustejovsky's Generative Lexicon Theory (1995 & 1998; Pustejovsky & Ježek 2008), thus bridging the gap between corpus linguistics and semantic theories of the lexicon.

In his theory, Pustejovsky tries to account for the semantic richness of natural language focusing on the compositional aspects of lexical semantics. According to this framework, lexical meaning is not an intrinsic feature of lexical items, but is generated by means of their contextual interaction, following the so-called *principles for strong compositionality*. As outlined in Ježek (2016: 78), these principles operate at a sub-lexical level targeting specific aspects of word meaning – such as SemTypes –

and are able to provide different interpretations for a wide range of lexical phenomena.

The *principle of co-composition*, for instance, offers an alternative take on verbal polysemy with respect to traditional accounts. If we consider lexical items expressing verb arguments to be as semantically active and influential as the verb itself (Pustejovsky 2002: 421), we do not need to think of verbs as polysemous, but rather conceive their meaning as contextually defined by the SemTypes of the surrounding arguments. For instance, if we apply this reasoning to the Croatian verb pair PITI/POPITI (= TO DRINK, *imperfective/perfective*), we can notice how its meaning changes depending on what is said to be “drunk”, namely a [BEVERAGE] (1), a [DRUG] (2) or a {GOAL} (3).

- | | | | |
|---|-----------------|------------------------------|--------------------------|
| (1) [[HUMAN] _{NOM}] | PIJE | [[BEVERAGE] _{ACC}] | |
| Djeca | ne piju | kavu. | |
| Children | don't drink | coffee. | |
| (2) [[HUMAN] _{NOM}] | PIJE | [[DRUG] _{ACC}] | |
| Većina ljudi | pije | antibiotike | na svoju ruku. |
| Most people | take | antibiotics | on their own initiative. |
| (3) [[HUMAN_FOOTBALL PLAYER] _{NOM}] | POPIJE | {GOL} | |
| Pavić | je popio | gol. | |
| Pavić | failed to score | a goal. | |

As for metonymic phenomena, in this framework they take the name of *semantic type coercions* (Pustejovsky 2002: 425; Pustejovsky & Ježek 2008, Ježek & Quochi 2010). Unlike co-composition instances, coercions do not cause shifts in verb meaning, but rather operate semantic type adjustments to the verb's selectional requirements within a given pattern. For instance, when a verb such as POPITI combines with a Direct Object with the semantic type [CONTAINER] in a context where it should select [BEVERAGE], it is instantiating a metonymic shift which enables us to interpret the given [CONTAINER] as the [BEVERAGE] itself, like in example (4).

- | | | |
|-------------------------------|----------|-------------------------------|
| (4) [[HUMAN] _{NOM}] | POPIJE | [[CONTAINER] _{ACC}] |
| Stipe | je popio | čašu. |
| Stipe | drank | a glass. |

4 Croatian-specific challenges

Being a Slavic language, Croatian displays a certain number of language-specific features, which had to be taken into account when setting up the new editor for CROATPAS, such as its case system, the consequent absence of prepositions when case markings are providing information on clause roles and verbal aspect. We implemented an editor which is proving to be able to tackle those challenges.

For instance, the following example (5) taken from the verb POSLATI (= TO SEND, *perfective*) shows how the addition of case markings as bottom-right indexes has proven essential to make the resource user-friendly: had they not been there, the absence of the preposition “to” in Croatian would have made Theme and Recipient morphologically undistinguishable from one another.

- (5) [[HUMAN]_{NOM}] POŠALJE [[ARTEFACT]_{ACC}] [[HUMAN]_{DAT}]
 Marija je poslala pismo gradonačelniku.
 Marija sent a letter TO the mayor.

For what concerns sentence structure, like the acronym suggests, the Croatian Typed Predicate Argument Structure resource leans on valency theory, where no distinction is made between subject and obligatory complements, since they are all considered essential verb *arguments* (Ježek 2016: 112). However, the editors of both TPAS and CROATPAS still rely on traditional clause-role labels for the underlying syntactic annotation, thus distinguishing subjects from objects and other obligatory complements.

Also traditional Croatian grammar distinguishes between clause roles, but the classification is heavily influenced by the Croatian case system and the use of prepositions. Croatian makes use of seven morphological cases – nominative, genitive, dative, accusative, vocative, locative and instrumental – which go by the name of *padeži* (Barić et al. 1997: 101)⁵. *Subjects* are usually expressed by the nominative case (6) (*ibidem*, 421), apart from some logical subjects appearing in the dative case (7).

- (6) Ivan-Ø je simpatičan-Ø
 Ivan-NOM is nice-NOM
 ‘Ivan is nice’

- (7) Vrti mi se
 (It) spins I.DAT REFL
 ‘I feel dizzy’

Direct objects (*ibidem*, 431) are expressed either by the accusative (8) or the genitive case (9), in case the context calls for a partitive genitive (*ibidem*, 435).

- (8) Irin-a čita knjig-u
 Irina-NOM reads book-ACC
 ‘Irina reads a book’

- (9) Hoćeš li kruh-a?
 (you) need Q bread-GEN
 ‘Do you want some bread?’

Indirect objects are expressed either by the genitive (10), dative (11) or instrumental case (12) (*ibidem*, 436).

- (10) Bojim se smrt-i
 (I) fear REFL death-GEN
 ‘I am afraid of death’

- (11) Veselim se Božić-u
 (I) rejoice REFL Christmas-DAT
 ‘I look forward to Christmas’

- (12) Revolver-om je lako rukovati
 Revolver-INS (it) is easy to handle
 ‘It is easy to handle a revolver’

Another distinction made in traditional Croatian grammar is the one between *non-prepositional* and *prepositional objects* (*ibidem*, 443): subjects, direct objects and the above-mentioned indirect objects all fall within the first category, whereas those objects in the accusative (13) or locative case (14) requiring a preposition obviously belong to the prepositional ones.

- (13) Preselit ću se u Amerik-u
 To move (I) will REFL to America-ACC
 ‘I am moving to America’

- (14) Živim u Zagreb-u
 (I) live in Zagreb-LOC
 ‘I live in Zagreb’

This being said, in order to facilitate future multilingual linking between resources, an attempt was made to keep the template of clause-role components for CROATPAS as adherent as possible to its Italian counterpart. Here is a list of the final clause-role labels used in CROATPAS:

- 1) **SUBJECT** – nominative and dative subjects
- 2) **OBJECT** – direct objects in the accusative case and partitive genitives
- 3) **INDIRECT COMPLEMENT** – indirect objects in the genitive, dative or instrumental case and prepositional objects
- 4) **ADVERBIAL** – to be used for those obligatory complements expressed by adverbs
- 5) **CLAUSAL** – for both clausal objects and subjects (sub-labels further specify which)
- 6) **PREDICATIVE COMPLEMENT** – of both object and subject (sub-labels further specify which)

Since both TPAS and CROATPAS are first and foremost semantic resources, the same verb pattern can contain different syntactic realizations. For instance, the corpus concordances behind the pattern displayed by example (6) contain sentences where the

⁵ Please note that, for the purpose of this paper, we limit the morphological glosses to case labels. However, the following examples show a number of typological features worth paying attention to, such as the fact that Croatian is a pro-drop language, it does not have articles and has an SVO word order. Here is a list of the abbreviations that we used: NOM (nominative), GEN (genitive), DAT (dative), ACC (accusative), LOC (locative), INS (instrumental), REFL (reflexive particle), Q (question particle).

SemType [INFORMATION] is assigned to both Objects in the accusative case and Clausal Objects, mostly introduced by Croatian complementizers such as DA, ŠTO (both equivalents of THAT) or KAKO (HOW).

- (15) [[HUMAN]_{NOM}] ČUJE [[INFORMATION]_{ACC}] | KAKO[INFORMATION]
Na početku ćete čuti upute. | Nisam čuo kako je bilo.
 At the start you will hear instructions. | I did not hear how it was.

Last but not least, verbal aspect had also to be taken into account during the set up of CROATPAS. Aspect is a grammatical category which applies to verbs only, offering “different ways of viewing the internal temporal constituency of a situation” (Comrie 1976: 3). Those verbs characterised by an imperfective aspect are able to report about actions while they are being carried out, while others – the perfective ones – focus on the completion of such actions. In some languages, aspect can be expressed through the choice of tense (in Italian, *imperfetto* vs. *passato remoto* or *passato prossimo*) or by means of periphrases (in English, the *-ing* form). On the other hand, Slavic languages such as Croatian present a set of prefixes and suffixes that are able to create so-called aspectual pairs or *vidski parnjaci* from one of the two forms (Barić et al. 1997: 226).

to read : ČITATI – PROČITATI (*imperfective/ perfective*)
 to write : PISATI – NAPISATI (*imperfective/ perfective*)
 to announce : OBJAVITI – OBJAVLJIVATI
 (*imperfective/ perfective*)

For each aspectual pair, patterns were extracted keeping the perfective and imperfective variants separate in the resource, as if they were two different verbs. Thus, by comparing the pattern inventories of the two aspects in each pair, we are able to evaluate to what extent aspectual differences influence verb meaning.

5 Related works

As we have already mentioned, CROATPAS is the sister project of the TPAS resource for Italian (Ježek et al. 2014). Both resources follow the CPA methodology (see § 2), which is also applied in the Pattern Dictionary of English Verbs (PDEV, Hanks & Pustejovsky 2005; El Maarouf et al. 2014) and in its Spanish counterpart (PDSV⁶).

⁶ PDSV is being compiled at the Pontifical Catholic University of Valparaíso (Chile) and is available online at: <http://www.verbario.com> (last visited on July 12th 2019). The project is coordinated by Irene Renau.

Existing reference dictionaries for Croatian are the *e-Glava*⁷ online valency dictionary of Croatian verbs (Birtić et al. 2017) and the Croatian Valence Lexicon of Verbs (CROVALLEX⁸, Mikelić Preradović et al. 2009). Unlike CROATPAS, *e-Glava* focuses only on 57 psychological verbs, whose meanings have been selected from pre-existing dictionaries and linked to valency patterns, which have been manually extracted from various Croatian corpora. Each argument in *e-Glava* is described on a morphological, syntactic and semantic level. As for morphology, the resource takes into account cases, prepositions and sentential realisations such as the complementizers ŠTO, DA, KAKO etc. Ten complement classes are specified at a syntactic level, namely Nominative Complement, Genitive Complement, Dative Complement, Accusative Complement, Instrumental Complement, Prepositional Complement, Adverbial Complement, Predicative Complement, Infinitive Complement and Sentential Complement (Birtić et al. 2017: 45). On a semantic level, the resource takes into account semantic role labelling (Agent, Patient, etc.), but has not yet introduced any hierarchically organised tagset of SemTypes as CROATPAS does.

Another important lexicographic reference work for Croatian is CROVALLEX (Mikelić-Preradović et al. 2009), the first project aiming at building a lexicon of valence frames for Croatian verbs. Its syntactic-semantic classes are taken from VerbNet (Kipper-Schuler 2005), which is based on Levin’s verb classes (1993). Once again, morphological information such as case markings and preposition are displayed, as well as semantic roles, but there is no mention of SemTypes. Overall the semantic resource CROATPAS is complementary to existing resources that focus primarily on the morphosyntactic layer.

6 Multilingual pattern linking and computer-assisted language learning

As pointed out by Baisa et al. (2016b), monolingual CPA-based dictionaries offer a unique chance to create multilingual resources by linking corresponding patterns, since they have been created following the same methodology.

⁷ <http://valencije.ihjj.hr/page/sto-je-e-glava/1/> (last visited on July 12th 2019)

⁸ <http://theta.ffzg.hr/crovallex/data/html/generated/alphabet/index.html> (last visited on July 12th 2019)

An early attempt of bilingual pattern linking was carried out by Popescu & Ježek (2013), who aligned CPA patterns of English and Italian using examples from the parallel corpus RTE3. Translation pairs were automatically extracted from the corpus and assigned to the corresponding patterns in the source and target language. The study was aimed at testing whether pattern-based translation is more likely to preserve meaning than Google translations, which was proven to be the case. More recently, Baisa et al. (2016a & 2016b) carried out further studies aimed at linking verb patterns from PDEV and its Spanish counterpart (PDSV) via their shared semantic types following both manual procedures and heuristic-based algorithms. Following Baisa, Vonšovský (2016) worked on the automatic linking of PDEV and VerbaLex (Hlaváčková 2008), a verb valency lexicon for Czech.

Starting in September 2019, an attempt is being made to cross-linguistically align a sample of 50 verb entries from CROATPAS with their Italian and English counterparts in TPAS and PDEV. We are interested in developing a flexible, semi-automatic, Italian-driven procedure able to disambiguate and link verb patterns across languages by matching their overlapping semantic contexts.

Perfect matches are already clearly foreseeable for verb patterns such as the ones in Figure 2, where both Italian, Croatian and English encode the meaning of “drinking a certain amount of alcoholic beverages” using the SemType [HUMAN] associated with the language-specific equivalent of TO DRINK.

T-PAS: [Human] bere
CROATPAS: [Human]_{NOMINATIVE} pije
PDEV: [Human] drink

Figure 2 – Perfect pattern matches

In order to be able to link also verb patterns which are not a perfect match, we are developing an algorithm able to recognize pattern similarity by taking into account also hypernym/hyponym relations between SemTypes. Figure 3 provides a fitting example, which shows how different annotation choices can result into the lumping or separation of semantically connected patterns containing hierarchically related SemTypes, such as [ANIMATE] > [HUMAN] & [ANIMAL] or [BEVERAGE] > [WATER].

T-PAS: [Animate] bere ([Beverage])
[Human] drink [Beverage]
PDEV: [Animal] drink ([Water])

Figure 3 – Hierarchically related SemTypes

On the other hand, CROATPAS has also the potential to become an interesting tool for learners and teachers of Croatian as an L2 in computer-assisted language learning (CALL), especially if combined with a user-friendly SKELL-inspired interface (Kilgarriff et al. 2015).

As its creators put it, SKELL (Sketch Engine for Language Learners) is “a stripped-down, non-scary version of Sketch Engine”, which grants learners access to:

- a summary of a word’s grammatical and collocational behaviour (Word Sketch);
- prototypical example sentences (Good Dictionary Examples) chosen by the GDEX algorithm (Kilgarriff et al. 2008);
- word clouds of similar words, i.e. words that share most collocations with the headword;
- corpus concordance lines

In the case of CROATPAS, displaying Good Dictionary Examples for each of the identified patterns could be a good way to provide real-life context and optional access to more concordance lines could be given to advanced learners. Word clouds displaying the lexical sets populating the SemTypes might also offer an eye-catching opportunity for computer-assisted vocabulary lessons.

At the moment, a resource which is probing these waters is *Woordcombinaties*: a Dutch tool aimed at combining access to collocations, idioms and valency patterns for computer-assisted second language learning and teaching (Colman & Tiberius 2018). This Dutch Collocation, Idiom and Pattern Dictionary focuses on a selection of mid-frequency lexical verbs and aims at offering immediate access to usage patterns from a toolbar, whose search options are: verbs in example sentences, Word Sketches with collocates, pattern-meaning pairs and pragmatic-oriented conversational routines (ibidem. 239). As underlined by the authors, tailor-made examples and Word Sketches can provide a good first impression of an unknown verb, while pattern-meaning pairs are thought for “advanced learners trying to find target

collocates or seeking confirmation of their intuitions regarding a collocation” (ibidem. 240).

7 Conclusion

In this paper, we introduced CROATPAS, a corpus-based digital collection of verb valency structures with the addition of semantic type specifications (SemTypes) to each argument slot. The resource relies on Pustejovsky’s Generative Lexicon theory (1995, 1998; Pustejovsky & Ježek 2008) (Section 3) and is made up of four key-components, namely: 1) a representative corpus of contemporary Croatian (hrWac 2.2. RELDI PoS-tagged); 2) a shallow ontology of SemTypes; 3) a methodology for Corpus Pattern Analysis (CPA, Hanks 2004 & 2013); and 4) the adequate corpus tools (Sketch Engine). We discussed the Croatian-specific challenges we faced while building the editor in Section 4, and provided an overview of the existing related works in Section 5. In Section 6, we anticipated the future multilingual linking of verb patterns from CROATPAS, TPAS and PDEV, which could provide a resource to be exploited in NLP, automatic translation and both theoretical and applied cross-linguistic studies. Moreover, CROATPAS could become an interesting tool for computer-assisted language learning (CALL).

References

- V. Baisa, S. Može, I. Renau (2016a). Linking Verb Pattern Dictionaries of English and Spanish. Presented at the *5th Workshop on Linked Data in Linguistics: Managing, Building and Using Linked Language Resources*. Portorož, Slovenia.
- V. Baisa, S. Može, I. Renau (2016b). Multilingual CPA: Linking Verb Patterns Across Languages. In: *Proceedings of the XVII Euralex International Congress*. Tbilisi, Georgia.
- E. Barić, M. Lončarić, D. Malić, S. Pavešić, M. Peti, V. Zenčević, M. Znika (1997). *Hrvatska gramatika*. Zagreb: Skolska knjiga.
- M. Baroni & A. Kilgariff (2006). Large Linguistically-Processed Web Corpora for Multiple Languages. In: *Proceedings of the XI Conference of the European Chapter of the Association for Computational Linguistics (EACL)*. Trento, Italy.
- M. Birtić, I. Brač, S. Runjaić (2017). The Main Features of the e-Glava Online Valency Dictionary. In: *Proceedings of the 5th eLex conference - Electronic lexicography in the 21st century*. Leiden, Netherlands.
- S. Cinkova, M. Holub, A. Rambousek, L. Smejkalova (2012). A database of semantic clusters of verb usages. In: *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC '12)*. Istanbul, Turkey.
- L. Colman & C. Tiberius (2018). A Good Match: a Dutch Collocation, Idiom and Pattern Dictionary Combined. In: *Proceedings of the XVIII EURALEX International Congress*. Ljubljana, Slovenia.
- B. Comrie (1976). *Aspect: An introduction to the study of verbal aspect and related problems*. Cambridge: Cambridge University Press (6th edition).
- T. De Mauro (2016). *Il Nuovo Vocabolario di Base della lingua italiana*. Available at the website: <https://www.dropbox.com/s/mkcyo53m15ktbnp/nuovovocabolariodibase.pdf?dl=0> (last visited on July 12th 2019).
- I. El Maarouf, J. Bradbury, P. Hanks (2014). PDEV-lemon: a Linked Data implementation of the Pattern Dictionary of English Verbs based on the Lemon model. In: *Proceedings of the 3rd Workshop on Linked Data in Linguistics (LDL): Multilingual Knowledge Resources and Natural Language Processing at the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. Reykjavik, Iceland.
- P. Hanks (2004). Corpus Pattern Analysis. In: *Proceedings of the XI Euralex International Congress*. Lorient, France.
- P. Hanks (2012). How People use words to make Meanings. Semantic Types meet Valencies. In: A. Bulton and J. Thomas (eds.) *Input, Process and Product: Developments in Teaching and Language Corpora*. Brno: Masaryk University Press.
- P. Hanks (2013). *Lexical Analysis: Norms and Exploitations*. Cambridge: The MIT Press.
- P. Hanks, E. Ježek, D. Kawahara, O. Popescu (2015). Corpus Pattern for Semantic Processing. In: *Proceedings of the Tutorials of the 53rd Annual Meeting of the ACL and the 7th IJCNLP*, Beijing, China.
- P. Hanks & J. Pustejovsky (2005). A Pattern Dictionary for Natural Language Processing. In: *Revue française de linguistique appliquée*, 10 (2), pp. 63-82.
- D. Hlaváková (2008). *Databáze slovesných valenčních rámců VerbaLex (Database of Verb Valency Frames VerbaLex)*, PhD Thesis, Masaryk University, Brno, Czech Republic.
- E. Ježek (2016). *The lexicon: An introduction*. Oxford: Oxford University Press.
- E. Ježek & V. Quochi (2010). Capturing Coercions in Texts: a First Annotation Exercise. In: *Proceedings*

- of the VII conference on International Language Resources and Evaluation (LREC). Valletta, Malta.
- E. Ježek, B. Magnini, A. Feltracco, A. Bianchini, O. Popescu (2014). T-PAS: A resource of Typed Predicate Argument Structures for linguistic analysis and semantic processing. In: *Proceedings of the Ninth conference on International Language Resources and Evaluation (LREC)*. Reykjavik, Iceland.
- A. Kilgariff, M. Husák, K. Mcadam, M. Rundell, P. Rychlý (2008). GDEX : automatically finding good dictionary examples in a corpus. In: *Proceedings of the 13th EURALEX International Congress* (pp. 425–432). Barcelona, Spain.
- A. Kilgariff, V. Baisa, J. Bušta, M. Jakubíček, V. Kovár, J. Michelfeit, P. Rychlý, V. Suchomel (2014). The Sketch Engine: ten years on. In: *Lexicography* 1(1), pp. 7-36.
- A. Kilgariff, F. Marcowitz, S. Smith, J. Thomas (2015). Corpora and Language Learning with the Sketch Engine and SKELL. In: *Revue française de linguistique appliquée*, 20(1), pp. 61-80.
- K. Kipper-Schuler (2005). *VerbNet: A broad coverage, comprehensive verb lexicon*, Ph.D. Thesis, University of Pennsylvania, USA.
- B. Levin (1993). *English Verb Classes and Alternations: A Preliminary Investigation*. Chicago: The University of Chicago Press.
- N. Mikelić Preradović, D. Boras, S. Kišiček (2009). CROVALLEX: Croatian Verb Valence Lexicon. In: *Proceedings of the 31st International Conference on Information Technology Interfaces*. Zagreb, Croatia.
- N. Ljubešić & T. Erjavec (2011). hrWaC and slWac: Compiling Web Corpora for Croatian and Slovene. In: *Text, Speech and Dialogue, Lecture Notes in Computer Science*, Springer.
- O. Popescu & E. Ježek (2013), Verbal Phrase Translation, *Tralogy Session 2 - Sense and Machine*. URL: <http://lodel.irevues.inist.fr/tralogy/index.php?id=216&format=print> (last visited on July 12th 2019).
- J. Pustejovsky (1995). *The Generative Lexicon*. Cambridge: The MIT Press.
- J. Pustejovsky (1998). The semantics of lexical underspecification. In: *Folia Linguistica* 32.
- J. Pustejovsky, P. Hanks, A. Rumshisky (2004). Automated Induction of Sense in Context. In: *Proceedings of the 20th International Conference on Computational Linguistics (COLING)*. Geneva, Switzerland.
- J. Pustejovsky & E. Jezek (2008). Semantic Coercion in Language: Beyond Distributional Analysis. In: *Italian Journal of Linguistics*, vol. 20, pp. 181-214.
- G. Rehm & S. Hegele (2018), Language Technology for Multilingual Europe: An Analysis of a Large-Scale Survey regarding Challenges, Demands, Gaps and Needs. In: *Proceedings of the XI Language Resources and Evaluation Conference (LREC 2018)*. Miyazaki, Japan.
- G. Rehm, H. Uszkoreit, I. Dagan, V. Goetcherian, M. U. Dogan, C. Mermer, T. Váradi, S. Kirchmeier-Andersen, G. Stickel, M. Prys Jones, S. Oeter, S. Gramstad (2014). An Update and Extension of the META-NET Study “Europe’s Languages in the Digital Age”. In: *Proceedings of the Workshop on Collaboration and Computing for UnderResourced Languages in the Linked OpenData Era (CCURL 2014)*. Reykjavik, Iceland.
- A. Špikić (2017). *Croato compatto: dizionario croato/italiano e italiano/croato*, Zanichelli: Bologna.
- M. Tadić, D. Brozović-Rončević, A. Kapetanović, (2012). *Hrvatski Jezik u Digitalnom Dobu – The Croatian Language in the Digital Age*. In: META-NET White Paper Series, G. Rehm & H. Uszkoreit (eds.), Springer: Heidelberg, New York, Dordrecht, London.
- J. Vonšovský (2016). *Automatic Linking of the Valency Lexicons PDEV and VerbaLex* (MA Thesis). URL:http://is.muni.cz/th/359500/fi_m/AutomaticLinking.pdf (last visited on July 12th 2019).

Enhancing a Text Summarization System with ELMo

Claudio Mastronardo

DISI - University of Bologna, Italy

claudio.mastronardo@studio.unibo.it

Fabio Tamburini

FICLIT - University of Bologna, Italy

fabio.tamburini@unibo.it

Abstract

Text summarization has gained a considerable amount of research interest due to deep learning based techniques. We leverage recent results in transfer learning for Natural Language Processing (NLP) using pre-trained deep contextualized word embeddings in a sequence-to-sequence architecture based on pointer-generator networks. We evaluate our approach on the two largest summarization datasets: *CNN/Daily Mail* and the recent *Newsroom* dataset. We show how using pre-trained contextualized embeddings on Newsroom improves significantly the state-of-the-art ROUGE-1 measure and obtains comparable scores on the other ROUGE values.

1 Introduction

The amount of human generated data is outstanding: every day we generate about 2 quintillion bytes of unstructured data and this number is expected to grow. With such a huge amount of information, swiftly accessing and comprehending large piece of textual data is becoming more and more difficult. Automatic text summarization constitutes a powerful tool which can provide a useful solution to this problem.

In recent years, automatic text summarization systems have gained a considerable amount of research interest due to deep learning powered NLP impressive results (Mikolov et al., 2013; Bahdanau et al., 2015; Yang et al., 2017; Vaswani et al., 2017; Józefowicz et al., 2016; Devlin et al., 2019). Neural network (NN) based approaches have always been considered data hungry techniques because they often require a large amount

of training data, but, in the latest years, several works have made a huge contribution in this direction (Grusky et al., 2018; Nallapati et al., 2016a; Napoles et al., 2012).

Text summarization systems can be divided into two main categories: *Extractive* and *Abstractive* (Shi et al., 2018). The first generate summaries by purely copying the most representative chunks from the source text (Dorr et al., 2003; Nallapati et al., 2016b), while in the second summarization algorithms make up summaries by using novel phrases and words in order to rephrase and compress the information in the source text (Chopra et al., 2016). Some works shed light on using both approaches through hybrid neural architectures attempting to gather the best characteristics of each world (See et al., 2017; Khatri et al., 2017).

NLP has seen a tremendous amount of attention after several deep learning based important results (Lample et al., 2016; Józefowicz et al., 2016; Hermann et al., 2015). Most of them relied on the concept of distributed representation of words, defining them as real-valued vectors learned from data (Mikolov et al., 2013; Pennington et al., 2014; Bojanowski et al., 2017; Joulin et al., 2017). Recent results were able to generate richer word embeddings by exploiting their linguistic context in order to model word polysemy (Peters et al., 2018; McCann et al., 2017; Peters et al., 2017).

In this paper, we build upon the work of See et al. (2017) on the Pointer-Generator Network for text summarization by integrating it with recent advances in transfer learning for NLP with deep contextualized word embeddings, namely an ELMo model (Peters et al., 2018). We show that, using pre-trained deep contextualized word embeddings, integrating them with pointer-generator networks and learning the ELMo parameters for combining the various model layers together with the text summarization model, we can improve substantially some of the ROUGE evaluation met-

Copyright ©2019 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

rics. Our experiments were based on two datasets commonly used to evaluate this task: *CNN/Daily Mail* (Nallapati et al., 2016a) and *Newsroom* (Grusky et al., 2018).

2 Related work

One of the first neural encoder-decoder approaches to text summarization has been presented by Nallapati et al. (2016a) where they show that an off-the-shelf encoder-decoder framework, used for machine translation, already outperforms the previous systems for text summarization. They also augment input data by concatenating to classical word embeddings part-of-speech tags, named-entity tags and tf-idf statistics. They leverage the *hierarchical attention* mechanism where less important chunks of text are less attended with a chunk-level mechanism attention.

Zhou et al. (2017) propose selective encoding for text summarization by introducing a selective gate network into the encoder with the purpose of distilling salient information from source articles. Then a second layer called “distilled representation” is constructed by multiplying the selective gate to the hidden state of the first layer. Such gate network can control information flow from encoder to the decoder and select salient information, boosting the performances of the sentence summarization task.

Read-Again Encoding (Zeng et al., 2016) follows the human approach of reading several times before writing a summary by using two LSTM encoders reading the source article and a transformed version of the first LSTM output respectively. Another original approach is presented by Xia et al. (2017) where they follow another human-driven approach by first writing a draft and then polishing it looking at the global context. In an encoder-decoder framework there are two decoders, the first attends to encoder states and generates a draft while the second attends to both the encoder and first decoder outputs generating a summary by exploiting information from *two* context vectors. This approach, called *deliberation network*, boosted the performances for both text summarization and machine translation.

Another set of approaches uses reinforcement learning as in Chen and Bansal (2018), where they use two sequence-to-sequence models. The first is defined as an extractive model with the goal of extracting salient sentences from the input source.

The second is an abstractive model which paraphrases and compresses the extracted sentences into a short summary. They make use of convolutional neural networks (ConvNet) to encode tokens and train the two models by using standard policy gradient methods treating them as reinforcement learning agents.

Paulus et al. (2018) presented a new abstractive summarization model achieving state-of-the-art on the New York Times dataset by introducing intra-temporal attention in both encoder and decoder. They use a new objective function by combining maximum-likelihood cross-entropy loss and rewards from policy gradient reinforcement learning in order to reduce the exposure bias and train their architectures by directly optimizing the ROUGE score.

Another research direction goes beyond RNNs to avoid their computational and memory costs by using ConvNet-based encoder-decoder models. Kalchbrenner et al. (2016) adopt one-dimensional convolutions stacking on top of the hidden representation on the encoder/decoder ConvNet. Quasi-Recurrent Neural Networks (Bradbury et al., 2017) use encoders and decoders made of convolutional layers and dynamic average pooling layers, requiring less amount of computational time when compared with LSTMs. Several other approaches attempted to use ConvNets for NLP.

It is also relevant the *transformer* model proposed in (Vaswani et al., 2017) which uses only feed-forward NN and multi-head attention.

3 Datasets

All the experiments in this work have been conducted on two datasets. The first, the *CNN/Daily Mail* dataset (Nallapati et al., 2016a), has been created by scraping news articles from the `cnn.com` website and concatenating news highlights in order to form a multi-sentence summary. It is composed of about 300,000 examples. The second, the recently released *Newsroom* dataset (Grusky et al., 2018) consists of 1.3 million article-summaries pairs. It is the largest and most diverse dataset known in literature. Compared to *CNN/Daily Mail* dataset, *Newsroom* has been created with the explicit goal of summarizing articles over two decades by using 38 major publishers as sources. Authors in (Grusky et al., 2018) also demonstrate that *CNN/Daily Mail* dataset is skewed towards extractive summaries, while the *News-*

room dataset covers a wider range of summarization styles, highly abstractive/extractive summaries and several article-summary compression ratios. For these reasons, even if we will provide the results for both datasets, we will mainly comment them only for the Newsroom dataset.

4 The Proposed Model

Our approach builds upon the work made by See et al. (2017) on pointer-generator networks applied to text summarization. The pointer-generator network is based on the architecture presented in (Nallapati et al., 2016c).

4.1 Pointer-Generator Network

It is an encoder-decoder architecture where tokens of a source text are fed one-by-one to an encoder network (a single layer LSTM) which also generates a sequence of hidden states. The decoder network (a single layer LSTM), at each step t receives the embedding of the emitted word at time $t - 1$ and the current decoder’s hidden state. This architecture makes use of Bahdanau attention (Bahdanau et al., 2015) using:

$$\begin{aligned} e_i^t &= \mathbf{v}^T \tanh(\mathbf{W}_h h_i + \mathbf{W}_s s_t + \mathbf{b}_{\text{attn}}) \\ a^t &= \text{softmax}(e^t) \end{aligned}$$

where s_t represents the decoder’s hidden state at step t , h_i represents the encoder’s hidden state at timestep i and e_i^t represents the weight given to h_i at decoder’s timestep t not yet normalized. Capital letters mark trainable parameters. The tensor a represents a probability distribution over encoder’s hidden states and encodes how much to attend each state in order to alleviate the encoder from the responsibility of encoding all the information into a fixed vector. The tensor a is used to produce a weighted sum of the encoder hidden states called h^* which is concatenated to the decoder’s current hidden state making up the input tensor for the LSTM cell that produces a distribution of probability over the vocabulary.

Pointer-generator networks extend this architecture by leveraging ideas from pointer networks (Vinyals et al., 2015): it is a special kind of architecture being able to point to a specific input token and copy it from the source text to the output sequence. At each time-step t the network produces a *generation probability* value $p_{\text{gen}} \in [0, 1]$ calculated from the context vector h^* , the decoder’s state s_t and the decoder’s input x_t :

$$p_{\text{gen}} = \sigma(\mathbf{w}_{h^*}^T h_t^* + \mathbf{w}_s^T s_t + \mathbf{w}_x^T x_t + \mathbf{b}_{\text{ptr}})$$

again capital letters represent learnable parameters and σ indicates the sigmoid function. p_{gen} is used as a soft switch to choose whether to generate a word from the network’s vocabulary or copy a word from the source text. So, given p_{gen} , the probability of outputting a word w is:

$$P(w) = p_{\text{gen}} P_{\text{vocab}}(w) + (1 - p_{\text{gen}}) \sum_{i:w_i=w} a_i^t$$

where P_{vocab} represents the probability value for the word w at the output layer of the LSTM decoder, $\sum_{i:w_i=w} a_i^t$ is the sum of the attention values given to the hidden states at time t whose input word was the specific word w . In the case of an extremely low p_{gen} , the decoder gives a higher probability value to the input words which produced hidden states who had been attended the most.

At a given time-step t the loss value is computed as the negative log-likelihood of the ground truth word w_t^* for that time-step

$$\text{loss}_t = -\log P(w_t^*)$$

and for a given sequence the loss value is computed by averaging the losses for each word.

In order to cope with the common repetition problem (Mi et al., 2016; Tu et al., 2016; Sankaran et al., 2016), the *coverage loss* (Tu et al., 2016) is used to penalize source-document words attended too much. It is implemented by maintaining a *coverage vector* c^t : $c^t = \sum_{t'=0}^{t-1} a^{t'}$ which tracks the degree of coverage that words have received from the attention mechanism so far. This leads to the augmented version of the attention mechanism including the coverage loss

$$e_i^t = \mathbf{v}^T \tanh(\mathbf{W}_h h_i + \mathbf{W}_s s_t + \mathbf{W}_c c_i^t + \mathbf{b}_{\text{attn}})$$

with \mathbf{W}_c as learnable parameter. Hence, coverage loss is computed by:

$$\text{covloss}_t = \sum_i \min(a_i^t, c_i^t)$$

in order to prevent repeated attention.

4.2 Deep Contextualized Word Embeddings

The original pointer-generator network does not use pre-trained word embeddings, but it learns 128-dimensional word embeddings from scratch during training. Even though learning specialized word embeddings for the summarization task might seem a reasonable approach, we think that using pre-trained word embeddings could improve the overall network performance.

Following Peters et al. (2018) we adopt a transfer learning approach by leveraging the power of

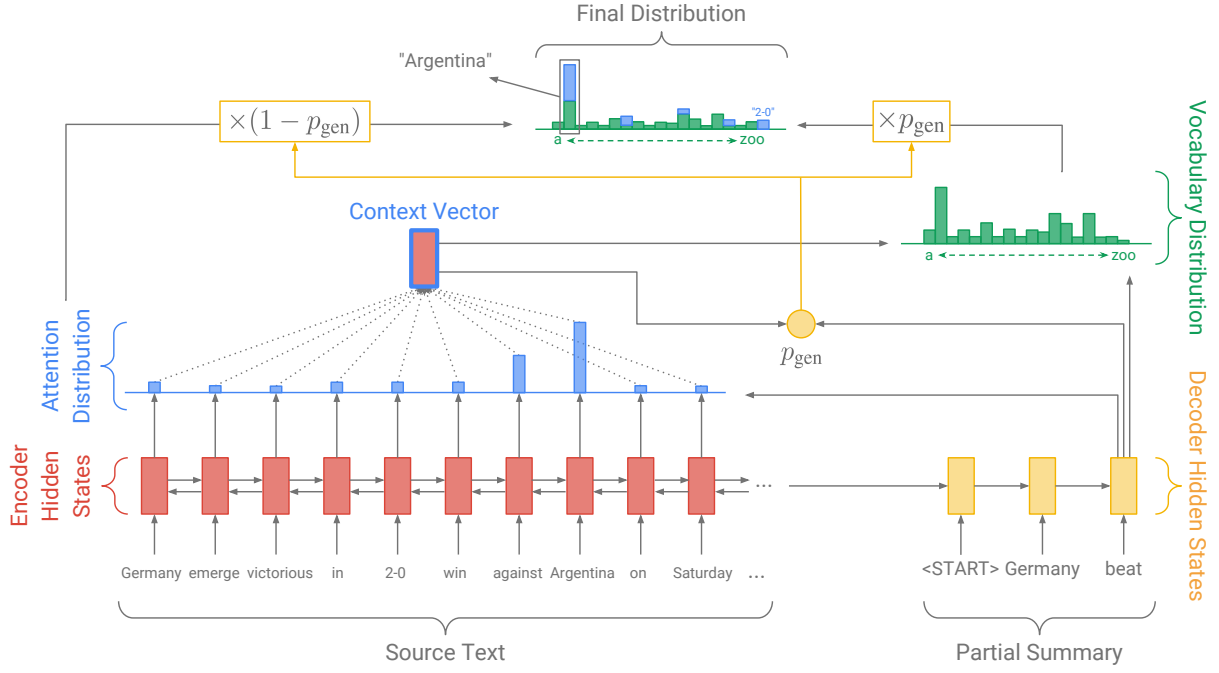


Figure 1: The pointer-generator model. At each time-step the encoder reads a word and outputs an hidden state. The decoder attends to encoder hidden states and generates the attention distribution. After generating p_{gen} , it weights and adds the attention distribution and the vocabulary distribution leading to the final word distribution. Picture courtesy of See et al. (2017).

pre-trained deep contextualized word embeddings. Embedding from Language Model (ELMo) is a particular type of embedding where word representation is a function of the entire input sequence. ELMo trains a bidirectional language modeling architecture inspired by Józefowicz et al. (2016) and Kim et al. (2016), on a large corpus. In order to compute the probability for the token t_k , the language model architecture computes a context-independent token representation via a ConvNet over characters and passes the output to a L -layer bidirectional LSTM. An ELMo representation is the result of a weighted combination of the hidden states of the language modeling architecture. For each token t_k , this architecture computes a set of $2L + 1$ representations: $R_k = \{\mathbf{h}_{k,j}^{LM} | j = 0, \dots, L\}$ where $\mathbf{h}_{k,0}^{LM}$ is the output of the ConvNet token layer and $\mathbf{h}_{k,j}^{LM} = [\vec{\mathbf{h}}_{k,j}^{LM}; \overleftarrow{\mathbf{h}}_{k,j}^{LM}]$ $j > 0$, for each bi-LSTM layer.

More generally, in order to use ELMo for a specific downstream task, word representations are computed by a weighted sum of each intermediate network representation:

$$\text{ELMo}_k^{\text{task}} = \gamma^{\text{task}} \sum_{j=0}^L s_j^{\text{task}} \mathbf{h}_{k,j}^{LM}$$

where s_j^{task} are softmax-normalized learnable

weights and γ^{task} allows to scale the entire produced vector with respect to the downstream task.

Our method feeds ELMo embeddings into a pointer-generator model: as the encoder reads the source text, a pre-trained ELMo model generates contextualized word embeddings. Pointer-generator encoder has two main sources to keep track of what has been read: its own memory and the inner information about past and following words injected into the current word embedding. We learn the s^{task} and γ^{task} weights during training.

We used the “Original (5.5B)” ELMo embeddings¹. The encoder gets 1024 dimensional embeddings which are fed into an LSTM cell of 512 neurons followed by a linear layer. Between the encoder and the decoder there is a neural network called *reduce state* with the aim of reducing the dimensionality of the passed tensors. The decoder is a bidirectional LSTM with size 256 followed by two linear layers of 256 neurons. We use an attention network with Bahdanau’s formula and the coverage mechanism. Decoder’s vocabulary size is set to the first most common 50,000 tokens in the training set. Freezing the model from learning embeddings from scratch reduces the number

¹<https://allennlp.org/elmo>

Paper	R-1	R-2	R-L
(See et al., 2017)	39.53	17.28	36.38
(Paulus et al., 2018)	41.16	15.75	39.08
(Gehrmann et al., 2018)	41.22	18.68	38.64
(Liu, 2019)	43.25	20.24	39.63
This work	38.96	16.25	34.32

Table 1: ROUGE metrics on CNN/Daily Mail test set.

Paper	R-1	R-2	R-L
(Grusky et al., 2018) (Pointer-generator)	26.04	13.24	22.45
(Shi et al., 2018)	39.36	27.86	36.35
This work	40.49	27.15	34.11

Table 2: ROUGE metrics on the Newsroom test set.

of parameters of 2,150,011. We trained our architecture on both CNN/Daily Mail and Newsroom datasets using Adagrad as the optimization algorithm (Duchi et al., 2011) with an initial learning rate of 0.15 and the initial accumulator set to 0.1. During training the batch size has been fixed to 8 and we run the decoder for at least 35 steps. As pre-processing step we just lowercased and tokenized texts using the *nltk* python package. The loss function remained unchanged since we used the negative log-likelihood for the ground truth word with coverage loss.

5 Experimental Results

We trained our model for 455,000 iterations on CNN/Daily Mail and for 520,000 iterations on Newsroom. The best performing models have been tested on both CNN/Daily Mail and Newsroom test sets and the ROUGE metrics are reported in Table 1 and 2 respectively.

The proposed approach achieves state-of-the-art ROUGE-1 value for the Newsroom dataset and competitive values for ROUGE-2 and ROUGE-L. ELMo addition causes an increase of +14.45, +13.91 and +11.66 for the three metrics with respect to basic pointer-generator from Grusky et al. (2018). ELMo s^{task} learned weights are, respectively, 0.4140, 0.4690, 0.1169 and $\gamma^{task} = 0.35$. This shows that the model favours syntactic information (captured at lower LSTM layers) instead of semantic information when generating text embeddings. From a qualitatively point of view we

report some network generated summaries as supplementary material². As we can see the model can generate fairly reasonable summaries, which can differ from the ground truth but still represent valid alternatives. This can explain the high value for ROUGE-1, meaning that summaries’ words have been covered anyway but in a different order (causing a lower ROUGE-L).

6 Discussion and Conclusions

In this work we leveraged recent results in transfer learning for NLP with deep contextualized word embeddings in conjunction with pointer-generator NN for automatic abstractive text summarization. We noticed a considerable increase of model’s performance in terms of the ROUGE score, achieving state-of-the-art on the Newsroom dataset for the ROUGE-1 metric. This is a dataset designed for testing abstractive systems while the other dataset (CNN/Daily Mail) contains summaries formed by sentences extracted from the original texts and it is more suitable for testing extractive systems. Then, it is reasonable that we got improvements only when using the Newsroom dataset.

Intrinsic, corpus-based metrics based on string overlap, string distance, or content overlap, such as BLEU and ROUGE, suffer from the need to have a reference output provided by the gold standard corpus in order to evaluate the system outputs. That seems very problematic (e.g. see Gatt and Krahmer 2018) because the reference summary is only one of the possible summaries that humans can produce. By looking at the supplementary material regarding some examples of our system output, one can immediately recognize that, even if very different from the reference one, the summaries produced by the proposed system are in most cases acceptable.

The definition of proper metrics capturing in the right way the correctness of system outputs remains, in our opinion, a critical open issue. As discussed also in the recent review by Chatzikoumi (2019) about Machine Translation (MT) metrics, “When reference translations are used [...] MT outputs that are very similar to the reference translation are boosted and not similar MT outputs are penalised even if they are good”, the so-called “reference bias”. The same metrics are currently used also in text summarization leading to similar problems.

²<https://bit.ly/2XUJvbd>

Acknowledgments

We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan Xp GPU used for this research.

References

- Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *Proc. of ICLR 2015*, San Diego, CA.
- Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.
- James Bradbury, Stephen Merity, Caiming Xiong, and Richard Socher. 2017. Quasi-recurrent neural networks. In *Proc. of ICLR 2017*, Toulon, France.
- Eirini Chatzikoumi. 2019. How to evaluate machine translation: A review of automated and human metrics. *Natural Language Engineering*, pages 1–25.
- Yen-Chun Chen and Mohit Bansal. 2018. Fast abstractive summarization with reinforce-selected sentence rewriting. In *Proc. of ACL 2018*, pages 675–686, Melbourne, Australia.
- Sumit Chopra, Michael Auli, and Alexander M. Rush. 2016. Abstractive sentence summarization with attentive recurrent neural networks. In *Proc. of NAACL-HLT 2016*, pages 93–98, San Diego, California.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proc. of NAACL 2019*, pages 4171–4186, Minneapolis, Minnesota.
- Bonnie Dorr, David Zajic, and Richard Schwartz. 2003. Hedge trimmer: A parse-and-trim approach to headline generation. In *Proceedings of the HLT-NAACL 03 Text Summarization Workshop*, pages 1–8, Edmonton, Canada.
- John Duchi, Elad Hazan, and Yoram Singer. 2011. Adaptive subgradient methods for online learning and stochastic optimization. *J. Mach. Learn. Res.*, 12:2121–2159.
- Albert Gatt and Emiel Krahmer. 2018. Survey of the state of the art in natural language generation: Core tasks, applications and evaluation. *J. Artif. Int. Res.*, 61(1):65–170.
- Sebastian Gehrmann, Yuntian Deng, and Alexander M. Rush. 2018. Bottom-up abstractive summarization. In *Proc. of EMNLP 2018*, pages 4098–4109, Brussels, Belgium.
- Max Grusky, Mor Naaman, and Yoav Artzi. 2018. Newsroom: A dataset of 1.3 million summaries with diverse extractive strategies. In *Proc. of NAACL2018*, pages 708–719, New Orleans, Louisiana.
- Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *Proc. of NIPS 2015*, pages 1693–1701, Montreal, Canada.
- Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. Bag of tricks for efficient text classification. In *Proc. of EACL 2017*, pages 427–431, Valencia, Spain.
- Rafal Jozefowicz, Oriol Vinyals, Mike Schuster, Noam Shazeer, and Yonghui Wu. 2016. Exploring the limits of language modeling. *CoRR*, abs/1602.02410.
- Nal Kalchbrenner, Lasse Espeholt, Karen Simonyan, Aaron van den Oord, Alex Graves, and Koray Kavukcuoglu. 2016. Neural machine translation in linear time. *CoRR*, abs/1610.10099.
- Chandra Khatri, Gyanit Singh, and Nish Parikh. 2017. Abstractive and extractive text summarization using document context vector and recurrent neural networks. In *Proc. of CoNLL 2016*, pages 280–290, Berlin, Germany.
- Yoon Kim, Yacine Jernite, David Sontag, and Alexander M. Rush. 2016. Character-aware neural language models. In *Proc. of AAAI 2016*, pages 2741–2749, Phoenix, Arizona.
- Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. In *Proc. of NAACL-HLT 2016*, pages 260–270, San Diego, California.
- Yang Liu. 2019. Fine-tune BERT for extractive summarization. *CoRR*, abs/1903.10318.
- Bryan McCann, James Bradbury, Caiming Xiong, and Richard Socher. 2017. Learned in translation: Contextualized word vectors. In *Proc.*

- of *NIPS 2017*, pages 6297–6308, Long Beach, CA.
- Haitao Mi, Baskaran Sankaran, Zhiguo Wang, and Abe Ittycheriah. 2016. Coverage embedding models for neural machine translation. In *Proc. of EMNLP 2016*, pages 955–960, Austin, Texas.
- Tomas Mikolov, G.s Corrado, Kai Chen, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *Proc. of ICLR 2013*, Scottsdale, Arizona.
- Ramesh Nallapati, Bing Xiang, and Bowen Zhou. 2016a. Sequence-to-sequence rnns for text summarization. In *Proc. of Workshop track - ICLR 2016*, San Juan, Puerto Rico.
- Ramesh Nallapati, Feifei Zhai, and Bowen Zhou. 2016b. Summarunner: A recurrent neural network based sequence model for extractive summarization of documents. In *Proc. of AAAI 2016*, Phoenix, Arizona.
- Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Çağlar Gulçehre, and Bing Xiang. 2016c. Abstractive text summarization using sequence-to-sequence RNNs and beyond. In *Proc. of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 280–290, Berlin, Germany.
- Courtney Napoles, Matthew Gormley, and Benjamin Van Durme. 2012. Annotated gigaword. In *Proceedings of the Joint Workshop on Automatic Knowledge Base Construction and Web-scale Knowledge Extraction*, pages 95–100, Montreal, Canada.
- Romain Paulus, Caiming Xiong, and Richard Socher. 2018. A deep reinforced model for abstractive summarization. In *Proc. of ICLR 2018*, Vancouver, Canada.
- Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proc. of EMNLP 2014*, pages 1532–1543, Doha, Qatar.
- Matthew E. Peters, Waleed Ammar, Chandra Bhagavatula, and Russell Power. 2017. Semi-supervised sequence tagging with bidirectional language models. In *Proc. of ACL 2017*, pages 1756–1765, Vancouver, Canada.
- Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proc. of NAACL-HLT 2018*, pages 2227–2237, New Orleans, Louisiana.
- Baskaran Sankaran, Haitao Mi, Yaser Al-Onaizan, and Abe Ittycheriah. 2016. Temporal attention model for neural machine translation. *CoRR*, abs/1608.02927.
- Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proc. of ACL 2017*, pages 1073–1083, Vancouver, Canada.
- Tian Shi, Yaser Keneshloo, Naren Ramakrishnan, and Chandan K. Reddy. 2018. Neural abstractive text summarization with sequence-to-sequence models. *CoRR*, abs/1812.02303.
- Zhaopeng Tu, Zhengdong Lu, Yang Liu, Xiaohua Liu, and Hang Li. 2016. Modeling coverage for neural machine translation. In *Proc. ACL 2016*, pages 76–85, Berlin, Germany.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proc. of NIPS 2017*, Long Beach, CA.
- Oriol Vinyals, Meire Fortunato, and Navdeep Jaitly. 2015. Pointer networks. In *Proc. of NIPS 2015*, pages 2692–2700, Montreal, Canada.
- Yingce Xia, Fei Tian, Lijun Wu, Jianxin Lin, Tao Qin, Nenghai Yu, and Tie-Yan Liu. 2017. Deliberation networks: Sequence generation beyond one-pass decoding. In *Proc. NIPS 2017*, pages 1784–1794, Long Beach, CA.
- Zichao Yang, Zhiting Hu, Yuntian Deng, Chris Dyer, and Alex Smola. 2017. Neural machine translation with recurrent attention modeling. In *Proc. of EACL 2017*, pages 383–387, Valencia, Spain.
- Wenyuan Zeng, Wenjie Luo, Sanja Fidler, and Raquel Urtasun. 2016. Efficient summarization with read-again and copy mechanism. *CoRR*, abs/1611.03382.
- Qingyu Zhou, Nan Yang, Furu Wei, and Ming Zhou. 2017. Selective encoding for abstractive sentence summarization. In *Proc. of ACL 2017*, pages 1095–1104, Vancouver, Canada.

KIParla Corpus: A New Resource for Spoken Italian¹

Caterina Mauri

Università di Bologna

caterina.mauri@unibo.it

Silvia Ballarè

Università di Torino

silvia.ballare@unito.it

Eugenio Gorla

Università di Torino

eugenio.gorla@unito.it

Massimo Cerruti

Università di Torino

massimosimone.cerruti@unito.it

Francesco Suriano

Università di Bologna

francesco.suriano2@studio.unibo.it

Abstract

In this paper we introduce the main features of the KIParla corpus, a new resource for the study of spoken Italian. In addition to its other capabilities, KIParla provides access to a wide range of metadata that characterize both the participants and the settings in which the interactions take place. Furthermore, it is designed to be shared as a free resource tool through the NoSketch Engine interface and to be expanded as a monitor corpus (Sinclair 1991).

1 KIParla corpus: an introduction

The aim of this paper is to describe the design and implementation of a new resource tool for the study of spoken Italian. The KIParla corpus is the result of a joint collaboration between the Universities of Bologna and Turin and is open to further partnerships in the future.

It is characterized by a number of innovative features. In addition to providing access to a wide range of metadata concerning the speakers and the setting in which the interactions take place, it offers transcriptions time-aligned with audio files and is designed to be expanded and upgraded through the addition of independent modules, constructed with a similar attention to the metadata; moreover, it is completely open-access and makes use of open-access technologies, such as the NoSketch Engine platform.

Section 2 provides a detailed description of the corpus design, aimed at featuring the geographic,

social and situational variation that characterizes spoken Italian. In Section 3 we discuss corpus implementation, describing how data have been collected in adherence with ethical requirements, how they have been treated and transcribed, and how they have been made accessible and searchable through NoSketch Engine. Section 4 focuses on the incremental modularity of the corpus, which makes it an open monitor corpus of spoken Italian. The two modules that constitute the current core of KIParla, namely KIP and ParlaTO, are then briefly illustrated, and some prospects for future developments are outlined.

2 Corpus design

This section discusses the parameters taken into account for the creation of the KIParla corpus. In particular, we stress the relevance of extralinguistic factors (regarding both the socio-geographic profile/status of the speakers and the interactional contexts) in order to build a corpus suitable for investigating (socio)linguistic variation in contemporary Italian.

2.1 Aims

The KIParla corpus is designed to overcome some of the shortcomings that characterize previous resources used in the study of spoken Italian. It is intended to bring about major improvements concerning three key aspects of corpus-based research: (i) access to the speakers' metadata, particularly to those concerning age and social group; (ii) the possibility to browse the corpus online as well as to download specific recordings; (iii) text-to-speech alignment.

¹ Copyright © 2019 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

As for (i), the possibility to recover information about the speakers or about the situation in which a conversational exchange has occurred is central in several fields of linguistics, such as sociolinguistics and conversation analysis, and is potentially relevant in many others, such as second language acquisition and language teaching. While some corpora provide general information about the setting of the interaction, at present there is no other corpus of spoken Italian that offers detailed information about single speakers. As for (ii), KIParla will be accessible online through the NoSketch Engine interface, and on the project website it will be possible to download all the recordings (in .wav or .mp3 format) and transcriptions, as previously done for CLIPS (Albano Leoni 2007), VoLIP (Voghera *et al.* 2014), and other corpora. Moreover, with regard to (iii) the research platform will enable users to listen to the results of single queries and download them in .mp3 format, offering text-to-speech alignment.

The philosophy behind KIParla is to pave the way for a collection of spoken corpora, each compiled according to a shared methodology in order to facilitate comparability. For this reason, it was designed as an open resource that is able to receive further implementations from external contributors who want to share their data; therefore, it can also be thought of as a monitor corpus (Sinclair 1991) which grows in size over time thanks to an increasingly wide range of materials.

2.2 The geographic dimension: collecting data in different cities with speakers from all over Italy

The diatopic dimension has always been considered to be of greatest significance when describing the Italian sociolinguistic scenario (see Berruto 2012 *inter al.*); in fact, speech utterances without any regional features are seldom if ever found even among educated speakers and in formal situations. Currently, the only spoken corpora that take into account geographic variation are the LIP corpus and the CLIPS corpus. In the KIParla corpus, thus far we have collected data in Turin and Bologna; the sociolinguistic situation in both urban settings is characterized by the coexistence of Italian and the local dialect, as well as the resulting development of intermediate varieties. Furthermore, even with significant differences, both cities have been and are destinations of internal mobility, and thus we are likely to find several varieties of Italian from other parts of Italy, as well as Italo-Romance dialects. One good exam-

ple of such a scenario is provided in (1); the conversation, recorded in Turin, has two speakers using the progressive periphrasis *stare + a + infinitive* combined with the apocopated form of the lexical verb, which are two typical features of regional varieties of Italian spoken in central Italy.

- (1) GF_TO091: ho capito ma tu sei entrata troppo nella parte **stai a fa'** l'attrice
 "I see but you are getting too much into this, you're putting on an act"

BC_TO089: sì
 "yes"

SF_TO090: no non **sto a fa'** l'attrice io parlo così normalmente come potete notare ragazze

"no, I'm not putting on an act. This is the way I usually speak, as you can see girls"

(KIP corpus, TOA3012)

In order to have a deeper understanding of the situation, information regarding both the city in which the data were collected and the place of origin of each speaker can be retrieved.

2.3 The diastratic dimension: a perspective on Italian society

The speakers involved in the recordings are distinguished primarily by their age and level of education; the latter are traditionally deemed to be the most relevant social factors for the analysis of sociolinguistic variation in Italian (see Berretta 1988). Part of the KIParla corpus (see KIP module in §4.1) is focused on educated speakers, i.e. undergraduates, graduate students, and university professors. In the second data collection sample (see ParlaTO module in §4.2), far more social factors have been taken into account, and both the age range and the level of education of the informants have been broadened. Ideally, the incremental nature of the corpus will make it possible to explore the various dimensions of variation in depth.

2.4 Types of interaction: settings and activities

Building on a central assumption in the conversation analytic framework, i.e. that linguistic practices are often related to specific social activities, we dedicated particular attention to including dif-

ferent types of situations, expecting to find considerable differences between the structures involved in each.

In order to narrow down the field of analysis, for the first bulk of the KIParla corpus we chose to consider various types of interaction occurring in a single sociolinguistic domain (Fishman 1972), namely the academic context.

The different activities were thus classified according to the following external factors: (i) the symmetrical vs asymmetrical relationship between the participants; (ii) the presence vs absence of previously established topics; (iii) the presence vs absence of constraints on turn-taking. We believe, indeed, that using these three very general features is particularly helpful in the task of integrating new data recorded in other situations, without losing comparability with the other parts of the corpus. For example, interviews collected with different types of speakers in the ParlaTO section (§ 4.2) will be comparable to those collected in the academic setting, regardless of any other difference between the two sets.

3 Building the corpus: data collection, transcription, publication, and accessibility

3.1 Data collection: praxis and ethics

All data have been collected by professional researchers; students and interns of the Universities of Bologna and Turin have also been involved in the process, but only after a period of specific training. Increasing the number of data collectors is crucial to avoid unwanted bias caused by the inclusion of informants that belong to the same social network. Furthermore, they acted as second-order contacts (see *friend of a friend* in Tagliamonte 2006: 21-22) and thus played an intermediary role in recording spontaneous speech and interviews.

Whenever data were being collected, speakers were first informed of the main aims of the project and the reasons why we needed to record the interaction. They agreed to the recording and signed a consent form that complies with the European Union's General Data Protection Regulation (G.D.P.R.). The consent form allowed us to collect linguistic material for scientific purposes, to store it in hardware located in Europe and/or via cloud services provided by universities, and to make it available online.

All the collected data are transcribed (see § 3.2) and anonymized before being made available to

the public. The voice of the speakers is the only sensitive data that remains directly accessible.

3.2 Transcription: challenges and solutions

All the recordings have been transcribed by professional researchers and trained students or interns using ELAN software (Sloetjes and Wittenburg 2008). This tool is designed specifically to handle multi-level annotations relating to different speakers in a conversation. It also makes it possible to link each annotation to the media timeline. Thanks to this feature of the software, it was possible to implement text-to-speech alignment within the NoSketch Engine interface (§3.3).

Every tier in the transcription refers to an alphanumeric code that links the spoken production of a single speaker to his/her metadata (e.g. age and level of education); similarly, each transcription file is associated with a code that allows its metadata to be traced (e.g. type of activity, number of participants, time and place of collection).

The most challenging aspect of transcribing spoken data is to strike a balance between a faithful representation of oral production and the "searchability" of the written texts. For this reason, we decided to adopt a simplified version of the Jefferson (2004) conventions used in conversation analysis (see Figure 1). An example of this transcription convention is provided in Figure 2.

,	Rising intonation
.	Falling intonation
:	Prolonged sound (each : corresponds to ca. 20ms)
(.)	Short pause
>hello<	Bracketed speech is delivered more rapidly
<hello>	Bracketed speech is delivered more slowly
[hello]	Overlap between participants
(hello)	Hardly intelligible speech (transcriber's best guess)
xxx	Unintelligible speech
((laughs))	Non-verbal behavior
=	Prosodically attached units

Figure 1: Symbols used in the transcription based on Jefferson (2004)

AG_BO097: e mi ha guardato, io l'ho guardato pero' cioe'
 GG_BO095: ti ha riconosciuta [si e' visto.]
 AG_BO097: [si': pero'.]
 AG_BO097: cioe' non non c'e' stato uno sguardo come dire:::
 AG_BO097: oh mio dio sei tu della lezione

Figure 2: Conversational transcription as shown in the corpus page

The decision to implement conversational transcription was mainly due to the fact that it enables us to obtain a sufficient level of precision, without forcing the researcher to make interpretive choices. This is crucial in the handling of both performance-related phenomena occurring in spoken language (e.g. reformulations and truncated words) and non-standard variants.

However, as will be explained in the next section, we decided to make the data searchable based on the simple orthographic transcription, while the conversational transcript is accessible as an additional option.

3.3 Data publication: From ELAN to NoSketch Engine

The transcriptions obtained through ELAN are in XML format and are automatically time-aligned to the speech audio files; thus, they are ready to be treated and parsed by XML-compatible technologies. Since one of our aims was to make the corpus fully accessible, we decided to make data available through the NoSketch Engine interface (Rychlý 2007).

NoSketch Engine is an open-source tool for corpus management which provides a powerful and user-friendly interface to perform corpus searches, generate word/keyword lists, retrieve collocations based on several statistical measures, and much more. In order to adapt the XML output of ELAN to the format required by NoSketch Engine, we wrote a python script that allows the user to: (i) make the metadata available both as query filters and text information; (ii) search the orthographic and Jefferson transcriptions; (iii) directly link every occurrence with the time-aligned portion of the media file associated with it; (iv) search each module of the corpus separately.

Users can perform a query either by browsing the whole corpus or by selecting one or more metadata concerning the participants or the conversation in which they appear. Figure 3 shows how the metadata can be selected in the corpus. As reported in Figures 4 and 5 respectively, with regard to the KIP module (§ 4.1) conversation metadata include the type of conversation, the city in which it was recorded and the year, the number of participants, and the relationship between them; the participants' metadata include occupation, gender, age, and the region of origin. During data collection, the participants indicated both the city of birth and the city in which they attended high school; however, we decided to retain only the latter information as an indicator of the speakers' region of origin.

Figure 3: Metadata selection

Type of conversation	Spontaneous conversation
	Exams
	Interviews
	Lessons
	Office hours
City	Bologna
	Turin
Number of participants:	1
	2
	3
	4
	5
	6
Year	2017/18
	2019
Relation between the participants	Asymmetrical
	Symmetrical

Figure 4: Conversation metadata

Figures 6 and 7 provide an example of a query in the NoSketch Engine interface; the results appear in KWIC (Keyword-In-Context) format, in which each token is presented within a string of characters containing the words that precede and follow it. By clicking on the conversation name reported in blue in the left portion of the screen, users can access the conversation's metadata, a full transcription of the file, both in Jefferson and text-only format, and a link to the corresponding

audio file (see Figure 6). By clicking on the token, in red, users can open a text box which provides further context (see Figure 7).

Occupation	Professor
	Student
Gender	Male
	Female
Region	Abruzzo
	Basilicata
	Calabria
	...
Age bracket	Under 25
	26-30
	31-35
	36-40
	41-45
	46-50
	51-55
	56-60
	Over 60

Figure 5: Participants' metadata

BOA3013	passione per am // m per questi tortini // quelli tipo col tofu // mhmh // comunque il crudo poi l'ho
BOA3013	a arrigo // e l'c'e' stato questo sguardo tipo // odio siamo seduti a due posti di distanza //
TOD2011	trasferirmi vabbè adesso parlando in grande tipo in america o comunque in posti dove m la grafica e
TOD2011	viene viene molto incontro agli studenti // che tipo di danza fai // e m faccio tip taph cheem
TOD2011	irlanda e precisamente a galway // e ehm chee' a tipo m due ore da dublino // e m e' stata una bellissim
TOD2011	stati ospitati tutti nella stessa struttura tipo un hotel e cose del genere in cui comunque con i
TOD2011	era un formaggio veramente schifoso // ed era tipo m giallognolo una roba del genere e infatti mi
BOD1007	intervista semistrutturata
BOD1007	TO
BOD1007	2
BOD1007	2017/18
BOD1007	simmetrico
BOD1007	http://151.236.39.174/bor
BOD1007	code=TOD2011&begin=1160 tutti gli approcci e a un
BOD1007	annotation.audio_file

Figure 6: Conversation metadata

BOA3013	passione per am // m per questi tortini // quelli tipo col tofu // mhmh // comunque il crudo poi l'ho
BOA3013	a arrigo // e l'c'e' stato questo sguardo tipo // odio siamo seduti a due posti di distanza //
TOD2011	trasferirmi vabbè adesso parlando in grande tipo in america o comunque in posti dove m la grafica e
TOD2011	viene viene molto incontro agli studenti // che tipo di danza fai // e m faccio tip taph che e m
TOD2011	irlanda e precisamente a galway // e ehm chee' a tipo m due ore da dublino // e m e' stata una bellissima
TOD2011	stati ospitati tutti nella stessa struttura tipo un hotel e cose del genere in cui comunque con i
TOD2011	era un formaggio veramente schifoso // ed era tipo m giallognolo una roba del genere e infatti mi
BOD1007	molto forte // che porta a un'estetica // di tipo immediato come se i morti non esistessero come
BOD1007	< previous e ci dava dei panini dei toast eh con e un m un formaggio tipico che usano che si chiama
BOD1007	cheddar // o qualcosa del genere non mi ricordo so solo che era un formaggio veramente schifoso //
BOD1007	ed era tipo m giallognolo una roba del genere e infatti mi ricordo che noi toglievamo sempre sto
BOD1007	formaggio dai toast perché era veramente immangiabile // buono eh però // eh // e m e poi anche la
BOD1007	sera poi m ci riuniamo next >

Figure 7: Context

As of September 2019, the corpus can be accessed online at the website www.kiparla.it. At present, it only consists of the KIP module (see 4.1), but further modules are already being processed and will be uploaded to the same website (see below). The corpus has not yet been lemmatized or POS-tagged, but such steps are planned for the near future.

4 Incremental modularity: an accessible open monitor corpus of spoken Italian

A key feature that makes the KIParla corpus particularly innovative is its incremental modularity,

namely its division into independent modules and the ability to add new modules over time.

Modules contain different corpora of Spoken Italian sharing the same design and a common set of metadata (see §2) which have been transcribed by ELAN and made available through NoSketch Engine by running the same script (see §3). The modules may focus on different dimensions of linguistic variation and may collect data from different geographical areas. However, the shared procedure of data collection and treatment guarantees a high level of mutual comparability.

Easy access to all of the metadata makes the corpus *expandable*, through the addition of further modules focusing on different geographical, socio-cultural, or communicative aspects, and *upgradable*, through the addition of new data to existing modules. Such a dynamic nature of the KIParla corpus makes it a potential monitor corpus, open to additions and upgrades over time. In the following sections, we provide a brief description of the two modules which at present constitute the core of the KIParla corpus.

4.1 KIP module

The KIP subcorpus is the first section that was designed within KIParla and was originally conceived as a self-sufficient unit. It consists of approximately 70 hours of recorded speech collected in Turin and Bologna (35 hours per city approximately) and transcribed between 2016 and 2019.

The subcorpus is domain-specific in that it includes various types of interactions occurring within the academic setting; moreover, from a sociolinguistic perspective, it only includes speakers whose achievements pertain to higher education, namely university students and professors. The social characteristics of the speakers are clearly reflected in speech data, e.g. in the highly educated use of the relative clause in example (2).

- (2) LB_BO100: abbiamo una struttura di dati, abbiamo un algoritmo **attraverso il quale** ci muoviamo tra queste strutture di dati

“we have a data structure, we have an algorithm **through which** we move among these data structures.”

(KIP corpus, BOD1007)

The structure of this subcorpus is intended to maximize diaphasic variability, according to the parameters described in 2.4 (symmetrical *vs* asymmetrical relations; presence *vs* absence of a

moderator; presence *vs* absence of a fixed topic). This resulted in the selection of the contexts listed in Figure 8, which represent ideal combinations between such parameters.

Activity	Bologna	Turin
spontaneous conversation	10:00:37	06:22:24
exams	03:09:34	03:10:48
lessons	12:19:39	13:25:33
interviews	06:18:37	07:47:38
office hours	02:59:11	03:49:08
TOTAL	34:47:38	34:35:30

Figure 8: Hours recorded for each interaction type in Turin and Bologna

The complete KIP module is currently available on the www.kiparla.it website.

4.2 ParlaTO module

ParlaTO is a corpus of spontaneous speech collected in Turin between 2018 and 2019. The corpus is being compiled in an effort to portray a contemporary multilingual urban setting. In fact, Turin has been, and still is, the scene of contact between different languages, partly because of the endogenous coexistence of Italian and Piedmontese, and partly as the result of both internal and external migration patterns.

Basically, the corpus contains speech data coming from three categories of individuals: (i) speakers of Piedmontese origin, (ii) speakers from other parts of Italy, and (iii) speakers of foreign origin, i.e. first and second-generation immigrants. Accordingly, the collection of data accounts for different languages and language varieties, namely Italian – either as L1 or L2 – and, to a lesser extent, immigrant minority languages and Piedmontese, as well as other Italo-Romance dialects. Therefore, the corpus makes it possible to investigate a wide range of phenomena. Below are just a couple of examples of Italian as L1: a case of substratum interference in (3), i.e. the absence of a preverbal negative marker (which characterizes most Northern Italo-Romance dialects), and a typical feature of uneducated speech in (4), i.e. the use of *ci* as 3pl indirect object clitic pronoun.

- (3) PST035: in quei tempi q- c’era proprio niente da mangiare

“in those days there was really nothing to eat”

(ParlaTO corpus, PTB009)

- (4) PMM017: c’erano gli altri ragazzi **ci** ho fatto dei nomi

“the other boys were there, I gave **them** some names”

(ParlaTO corpus, PTB002)

Data has been collected through semi-structured interviews about city life and personal experiences (urban initiatives, policies for neighborhoods, leisure time activities, etc.). The corpus provides a rich set of metadata, geared to fostering the investigation of linguistic variation across socio-economic classes and social groups. It includes such categories as age, level of education, gender, employment status, place of birth (of both the individual and their parents), mother tongue, and knowledge of other languages, as well as duration of stay and duration of study in Italy for first and second-generation immigrants. The occurrence of Italo-Romance dialects and/or foreign languages in speech utterances is being tagged as well.

ParlaTO is thus meant to fill some crucial gaps in the *panorama* of Italian speech corpora. In particular, the spontaneous speech of such social groups as young speakers with limited educational qualifications and first and second-generation immigrants can, for the first time, be the subject of targeted corpus-based searches online.

The corpus currently amounts to approximately 60 hours of speech, one third of which is from speakers of foreign origin. However, ParlaTO is still under construction and will not be available online until early 2020.

5 Conclusions and future prospects

The ParlaTO corpus has been added to the KIP corpus, thereby creating two modules within the larger KIParla corpus. We aim to make this resource grow over time through subsequent additions and upgrades. The leading idea is that the greater the variety of interactions, speakers, and geographical areas recorded in the KIParla data, the more the corpus will become representative of the language(s) and language varieties spoken in

Italy. Moreover, as the corpus is upgraded over time, it will tell us more and more about the sociolinguistic situation in the Italian peninsula.

We envision the future development of the corpus to proceed in two main directions. On the one hand, we intend to collaborate with existing projects, in order to verify whether data already collected for different purposes may be adapted into new modules of the KIParla corpus. The only requirement in such cases is the ability to trace and access a core set of metadata for the speakers (gender, age, geographical information, level of education, and occupation) and for the interaction (interview, free conversation, etc.). Further metadata would of course be welcome. Moreover, new data collection efforts have already started or are scheduled to start in different regions (e.g. in Lombardy). A data collection project parallel to ParlaTO is also planned for Bologna.

The second direction along which KIParla will grow has to do with data annotation. For the moment, KIParla data are available as prosodic and orthographic transcriptions, time-aligned with the speech audio file and linked to the metadata of speakers and interactions. Further functions are offered by NoSketch Engine, such as word sketches, thesaurus, and keyword computation.

We plan two further stages of annotation, namely lemmatization and POS-tagging, which will significantly enhance data retrieval. Due to space constraints, we are unable to discuss the problems that lemmatization and POS-tagging raise when applied to spoken data (cf. Panunzi, Picchi, Moneglia 2004), and leave such a crucial discussion to future work.

References

- Albano Leoni, Federico (2007), "Un frammento di storia recente della ricerca (linguistica) italiana. Il corpus CLIPS". In: *Bollettino d'Italianistica*, IV, (2), 122-130.
- Berretta, Monica (1988), "Italienisch: Varietätenlinguistik des Italienischen/Linguistica delle varietà". In: *Lexicon der Romanistischen Linguistik*, vol. IV 762-774.
- Berruto, Gaetano (2012), *Sociolinguistica dell'italiano contemporaneo. Seconda edizione*, Roma, Carocci.
- De Mauro, Tullio, Federico Mancini, Massimo Vedovelli and Miriam Voghera (1993), *Lessico di frequenza dell'italiano parlato*, Milano, Etaslibri.
- Fishman, Joshua (1972), "Domains and the relationship between micro- and macrosociolinguistics. In: Gumperz, John and Dell Hymes (eds.), *Directions in sociolinguistics. The ethnography of communication*, New York, Holt, Rinehart and Winston, 435-453.
- Jefferson, Gail (2004), "Glossary of transcript symbols with an introduction". In: Lerner, Gene H. (ed.), *Conversation Analysis: studies from the first generation*, Amsterdam, John Benjamins, 13-31.
- Tagliamonte, Sali A. (2006), *Analysing sociolinguistic variation*, Cambridge, Cambridge University Press.
- Panunzi, Alessandro, Eugenio Picchi and Massimo Moneglia (2004), "Using PiTagger for Lemmatization and PoS Tagging of a Spontaneous Speech Corpus: C-Oral-Rom Italian". In: *Proceeding of Fourth Language Resources and Evaluation Conference (LREC 2004)*.
- Rychlý, Pavel (2007), "Manatee/Bonito – A Modular Corpus Manager". In: *1st Workshop on Recent Advances in Slavonic Natural Language Processing*, Brno, Masaryk University, 65-70.
- Sinclair, John (1991), *Corpus, Concordance, Collocation*, Oxford, Oxford University Press.
- Voghera, Miriam, Claudio Iacobini, Renata Savy, Francesco Cutugno, Aurelio De Rosa and Iolanda Alfano (2014), "VoLIP: A searchable Italian spoken corpus". In: Vaseľovská, Ludmila and Markéta Marjanebová (eds.), *Complex visibles out there. Proceedings of the Olomouc Linguistics Colloquium: Language use and linguistic structure*, Olomouc, Palacký University, 628-640.

Evaluating Speech Synthesis on Mathematical Sentences

Alessandro Mazzei

Università degli Studi di Torino
alessandro.mazzei@unito.it

Michele Monticone

Università degli Studi di Torino
michele.monticone@edu.unito.it

Cristian Bernareggi

Università degli Studi di Torino
cristian.bernareggi@google.com

Abstract

English. In this paper we present the main features of a rule-based architecture to transform a \LaTeX encoded mathematical expression into its equivalent *mathematical sentence* form, i.e. a natural language sentence expressing the semantics of the mathematical expression. Moreover, we describe the main results of a first human based evaluation of the system for Italian language focusing on speech synthesis engines.

Italiano. *In questo lavoro presentiamo le caratteristiche principali di un'architettura software a regole per trasformare un'espressione matematica, codificata in \LaTeX , nella sua equivalente frase matematica, cioè una frase del linguaggio naturale che esprima la stessa semantica dell'espressione originale. Inoltre, descriviamo i primi risultati di una valutazione del sistema fatta da esseri umani per la lingua italiana riguardante principalmente i motori di sintesi del parlato.*

1 Introduction

Computational linguistics can help people in many ways, especially in the field of assistive technologies. In the case of mathematical domain, blind people can access to a mathematical expression by listening its \LaTeX source. However, this process has several drawbacks. First of all, it assumes the knowledge of the \LaTeX . Second, listening \LaTeX is slow and error-prone, since \LaTeX is a typographical language, that is a language designed

for specifying the details of typographical visualization rather than for efficiently communicate the semantics of a mathematical expression. For instance, the simple \LaTeX expression $f(x)$ is a typographical description and so it represents both the function application of f to x , and the multiplication of the variable f for the variable x surrounded by parenthesis.

There are many lines of research to enable people with sight impairments to access mathematical contents. It is possible to embed mathematical expressions in web pages not only as images but through MathML or MathJax (Cervone, 2012) and in PDF documents produced from \LaTeX (Ahmetovic et al., 2018). Other research directions concern conversion into Braille (Soiffer, 2016) and speech reading (Raman, 1996; Waltraud Schweikhardt, 2006; Sorge et al., 2014).

In this paper we follow another direction: we consider the possibility to produce a *mathematical sentence*, i.e. a natural language sentence expressing the semantics of a mathematical expression. Indeed, the idea to use mathematical sentences for improving the accessibility of mathematical expressions has been previously presented and experimented for Spanish in (Ferres and Fuentes Sepúlveda, 2011; Fuentes Sepúlveda and Ferres, 2012). However, in contrast to previous work on mathematical sentences, in this work we use a natural language generation (NLG) architecture rather than a template-based one for generating sentences. By using NLG architecture we obtain (i) more portability, and (ii) a major and simple customization of the output.

We have two research goals in this paper. The first goal is to describe a system for transforming a mathematical expression natively encoded in \LaTeX in its equivalent mathematical sentence (cf. Figure 1). The processing flow follows a well-known approach, called *interlingua* in the field of machine translation (Hutchins and Somer, 1992).

Copyright © 2019 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

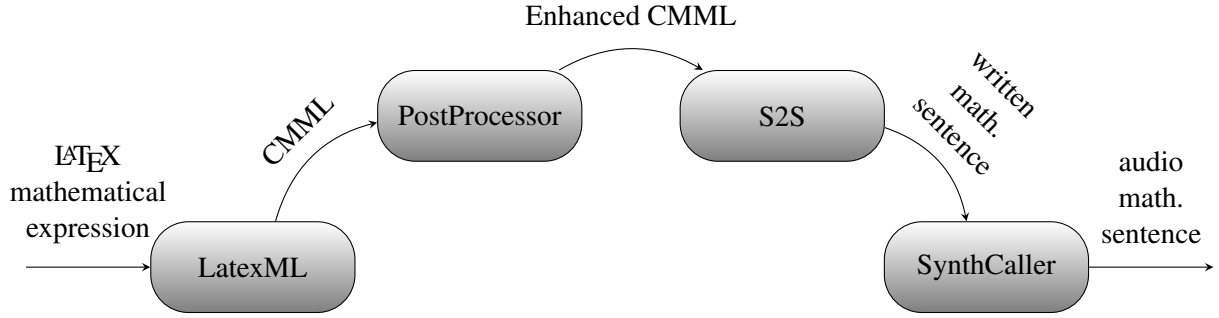


Figure 1: The software architecture for the generation of mathematical sentences. The process starts from (1) the $\text{L\textsubscript{A}T\textsubscript{E}X}$ representation of the expression, (2) its translation in CMML, (3) enhancement of CMML, (4) generation of the written form of the mathematical sentence, (5) production of the audio form of the mathematical sentence.

Indeed, the process of generating a mathematical expression from its $\text{L\textsubscript{A}T\textsubscript{E}X}$ source is a two-step algorithm. In the first step the $\text{L\textsubscript{A}T\textsubscript{E}X}$ is analyzed and its semantics is represented in *Content MathML* (CMML henceforth), a W3C standard for expressing the syntax and the semantics of mathematical expressions¹. In the second step, the CMML representation is used as input of the S2S (Semantics to Speech) module, that is a NLG module generating the mathematical sentence. Note that the S2S module inserts in the sentence parenthesis and pauses too. The sentence will finally be transformed in audio format encoding by an external synthesis engine.

The second goal of this paper is to give a first evaluation of the performance of two distinct synthesis engines in the domain of mathematical sentence. With a pilot experimentation conducted with four blind people, we will compare the perception of the mathematical sentences of a neural-network based speech engine and of a formant-based speech engine.

In Section 2 we will describe the main features of the developed system, in Section 3 we will describe the experimentation and finally in Section 4 we end the paper with some conclusions and introducing future work.

2 Building Mathematical Sentences

The first step of our algorithm is the generation of CMML associated to a $\text{L\textsubscript{A}T\textsubscript{E}X}$ formula. We based this step on an external tool named *LatexML* (Miller, 2007). However, the CMML obtained from this tool needed to be enhanced by a post-processing procedure for (1) uniform them

to CMML standard and (2) to remove ambiguity as for the case $y = f(x)$. In Figure 2 we report the CMML representation for the mathematical expression $x > b \implies |f(x)| < M$.

Mathematical notation has been conceived with the aim of representing mathematical concepts using a specific written symbolic language. As working hypothesis, we decided to assume a “specialized” syntactic analysis for a number of mathematical objects. For instance, `x plus three` indicates the action of adding one quantity to another, so it can be represented as a declarative structure. As a consequence, `plus` can be analysed as verb and this assumption can be extended to all the mathematical sentences. In this paper we considered only the mathematical structures belonging to the subfield of the mathematical analysis. In particular, we considered all the expressions in an Italian analysis book (Pandolfi, 2013). By using this corpus of expressions and by assuming that all numbers and variables can be treated as nouns and that all arithmetic operators can be treated as verbs, we found eight additional categories for representing all complex mathematical expressions and we defined a specific syntactic construction for each category.

In Table 1, we reported some examples of syntactic constructions for mathematical expressions. We decided to analyse and represent the mathematical sentences of relational operators as copula sentences (*a è maggiore di b*, *a is greater than b*), algebraic operators as declarative sentences (*a prodotto cartesiano b*, *a cartesian product b*), logical operators as conjunctions (*a o b*, *a or b*), elementary operators (e.g. *radice*, *radical*), sequence (e.g. *limite*, *limit*), calculus (e.g. *integrale*, *inte-*

¹<https://www.w3.org/TR/MathML3/chapter4.html>

```

<apply>
  <implies/>
  <apply>
    <gt/>
    <ci>x</ci>
    <ci>b</ci>
  </apply>
  <lt/>
  <apply>
    <abs/>
    <apply>
      <ci>f</ci>
      <ci>x</ci>
    </apply>
  </apply>
  <ci>M</ci>
</apply>

```

Figure 2: The CMML representation of the mathematical expression $x > b \implies |f(x)| < M$.

Mathematical Expression	Construction
$>, \geq, \gg, \dots$	Copula
$+, -, *, \dots$	Declarative
$\wedge, \vee, \neg, \dots$	Coordination
\sin, \cos, \tan, \dots	Noun Phrase
$\sum_{[x=a]}^{[b]} [f(x)], \dots$	Noun Phrase
$\int_{[a]}^{[b]} f(x) \, dx$	Noun Phrase
$\{[vars] \mid conditions\}$	Reduced Relative
$([x], [y])$	Reduced Relative

Table 1: Mathematical expressions and their linguistic constructions.

gral) as noun phrases (La radice quadrata di x , *the square root of x*), pairs and conditional sets as reduced relatives (L'insieme delle x tali che x è minore di 3, *the set of x such that x is less than 3*). Our syntactic representations for mathematical operators in the analysis domain could have alternative representations or could be specialized in a more refined classification (c.f. (Chang, 1983)), but we decided to use only eight category for sake of simplicity.

Traditional NLG architectures split the generation process into three distinct phases, that are *document planning*, *sentence planning* and *realization* (Reiter and Dale, 2000; Gatt and Krahmer, 2018). In particular document planning decides *what* to say and sentence planning and realization decides *how* to say it. In the system architecture depicted in Figure 1, the content of the communication is

specified by the input mathematical expression, so the content selection phase is not necessary at all. In Section 2.1 we will give some details on the rule-based sentence planner designed for managing mathematical sentences and in Section 2.2 we will describe the use of the SimpleNLG-it realizer for the case of mathematical domain.

2.1 Building a Sentence Planner for Mathematical Sentences

The input of the sentence planner is a mathematical expression in the form of enhanced CMML. In order to associate a *sentence plan*, that is a sort of under-specified tree-based syntactic structure, we devised a recursive algorithm that traverses top-down the CMML structure.

By considering the eight categories used to classify all mathematical expressions, for each category we designed a prototypical sentence plan that will be used in the recursive process. Each prototype builds a specific linguistic construction (e.g. *copula*, *reduced relative* etc.), that is designed for giving syntactic roles to the arguments of the specific mathematical construction. For instance, on the left of the Figure 3, we reported the prototypical sentence plan for the *conditional set* mathematical structure and on the right of we reported an example of its instantiation. In the final produced structures we have that, (1) the leaves of the sentence plan are lemmas rather than words, (2) the syntactic relations among the nodes are expressed using both dependency relations (e.g. subj, complement) as well as constituency nodes (e.g. Prepositional Phrase, PP). Note that this is the input format for sentence plan required by the SimpleNLG realizer (see Section 2.2).

In order to build a complete sentence plan for a mathematical sentence by using the eight categories for mathematical expressions, there are two important issues.

The first issue concerns the perception of precedence of the arithmetic operator. Listening mathematics has some peculiarities with respect to reading it. For instance, division is granted a higher precedence than addition, and during the reading process the expression $a + b/c$ is parsed as $a + \frac{b}{c}$ without ambiguities. A different result arises if one listens the equivalent mathematical sentence *a plus b divided by c* without reading the expression: we experimented that the most frequent perceived parse is $\frac{a+b}{c}$. After a limited num-

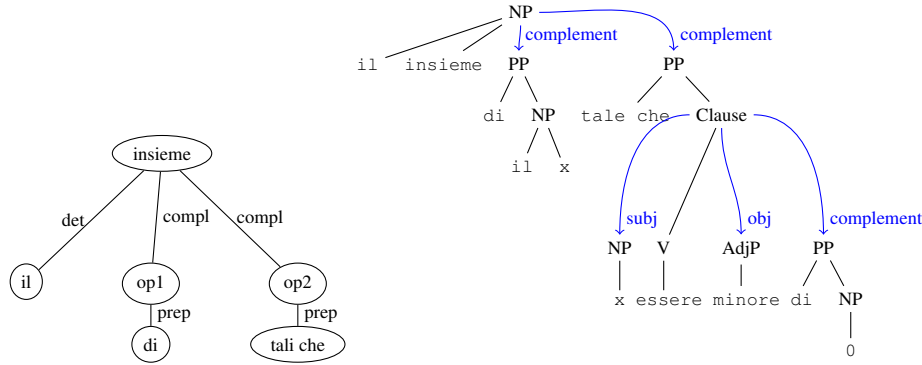


Figure 3: The prototypical sentence plan for the *conditional set* mathematical structure (left), and its fulfillment producing the sentence *L'insieme degli x tali che x è minore di 0* (right, *the set of all x such that x is lesser than 0*).

ber of experiments in listening arithmetic expressions with distinct (blind and not blind) people, we decided to state as working hypothesis that *the precedence of the arithmetic operators are perceived in the reverse order when one listens a mathematical expressions without reading it*².

A second issue is how to represent the correct structures of the operators. In other words, how we can build a mathematical sentence unambiguously equivalent to $\frac{a+b}{c}$? A trivial but effective solution is to use parenthesis, that is to produce the mathematical sentence *open parenthesis a plus b close parenthesis divided by c*. However, the drawback of this solution is the length of the sentence that, for very complex expressions, can augment substantially.

In order to account for both the issues, we modified the sentence planner in two ways. First, we decided to model parenthesis as lexical items, that is we considered *open-parenthesis* and *closed-parenthesis* as two new lexical items of the SimpleNLG lexicon which can be used as pre-modifier and post-modifier of a mathematical sentence respectively. Second, similar to (Fuentes Sepúlveda and Ferres, 2012), we allowed to use a *speech pause* as a synonymous of open/closed-parenthesis items. Moreover, in order to experiment both with parentheses and pauses in the understanding of a mathematical sentence, we decided to implement three distinct parenthesis strategies, called *parenthesis*, *pause*, and *smart*. In the *parenthesis strategy*, all the necessary parentheses are inserted in the sentence plan.

²We have not been able to find any scientific reference on this point.

Note that a parenthesis has to be considered necessary with respect to the inverted precedence order hypothesis stated above. In the *pause strategy*, all the necessary pauses are inserted in the sentence plan. In the *smart strategy*, all the necessary parentheses are inserted in the higher nodes of the sentence plan, and the necessary pauses are inserted close to the leaves of the sentence plan. This is a hybrid strategy that combines parentheses and pauses in order to have a less verbose mathematical sentence.

2.2 NLG for spoken mathematics

In order to produce a spoken mathematical sentences in Italian with the SimpleNLG-it realizer (Mazzei et al., 2016), we needed to account for the construction of a domain specific lexicon for the field of the mathematical analysis. SimpleNLG-it is the Italian porting of the SimpleNLG realizer, that was originally designed only for English (Gatt and Reiter, 2009). As default Italian lexicon, SimpleNLG-it uses a basic vocabulary of around 7000 words, that is a *simple* lexicon studied to be perfectly understood by most Italian people (Mazzei, 2016; Conte et al., 2017; Ghezzi et al., 2018). However, for this specific project we needed to augment the basic lexicon with both (i) a mathematical specialized lexicon, that contains both new lexical entries (as *arcotangente*, *arctangent*), and (ii) new values for lexical entries which are yet in the basic lexicon (as the value *noun* for the part of speech of the lemma *integrale*, *integral*). This specialized lexicon contains 113 entries which are mostly categorized as nouns (e.g. *logaritmo*, *logarithm*), verbs (e.g. *intersecare*, *intersect*), adjective (e.g. *iperbolico*, *hyperbolic*). In the lexicon, there

are only two new instances of adverbs (that are *relativamente* and *propriamente*, *relative*, *properly*), and only one instance of “prepositional locution” (that is *tale che*, *such that*). Finally, we added specific lexical items to realize both parenthesis (that are *parentesi aperta* and *parentesi chiusa*, *open/closed parenthesis*) and speech pause. This latter item will be finally realized by using the SSML (Speech Synthesis Markup Language) tag `<break/>`, that can be processed by many speech synthesis engines³.

The actual version of the mathematical sentence generator has been interfaced with two speech synthesis engines, that are the web service provided by the IBM-Watson framework⁴ (*W-engine* henceforth), and the Espeak API⁵ (*E-engine* henceforth). *W-engine* is a commercial, closed software based on deep learning, while *E-engine* is a free, open-source software based on formant synthesis algorithms. Note that for not visual impaired people *W-engine* sounds more fluent but, in contrast, for visual impaired people *E-engine* sounds more familiar since it is used by a widespread free screen reader.

3 Evaluation

In order to have a first evaluation of the generation system, we built a web-based test explicitly designed for visually impaired people. We designed a questionnaire composed by a 6 multiple choices questions concerning personal data, a core of 25 open questions each one concerning the listening of a mathematical sentence and its comprehensibility, 1 Likert-scale question globally comparing \LaTeX and system comprehensibility, 1 open question for free comments.

The 25 core questions have a all the same schema: there is a audio file encoding a mathematical sentence and there is a open form for transcribing it. In the compilation instructions, we asked the users to fill this section by using “ \LaTeX or with other non ambiguous formal representation”. The mathematical expressions obtained have been manually translated to CMML for evaluation. We implemented the questionnaire by using the Google Form framework, that was

³<https://www.w3.org/TR/speech-synthesis11/>

⁴<https://www.ibm.com/watson/services/text-to-speech/>

⁵<http://espeak.sourceforge.net>

ID	Formula
E1	$A \times B = \{(x, y) \mid x \in A, y \in B\}$
E4	$x > b \implies f(x) < M$
E6	$\lim_{x \rightarrow x_0} \left\{ \frac{f(x) - f(x_0)}{x - x_0} - f'(x_0) \right\} = 0$
E8	$\int \frac{1}{\sqrt{m^2 - x^2}} dx = \arcsin \frac{x}{m} + c$
E10	$\lim \left(1 + \frac{1}{n} \right)^n = e$

Table 2: The five mathematical expressions used for experimentation.

preliminarily judged accessible by a blind person.

In this paper we discuss the results of 10 core questions of the questionnaire that have been created by using the 5 mathematical expressions belonging to the Table 2. We use the *W-engine* to build 5 mathematical sentences and the *E-engine* to build other 5 mathematical sentences. Note that we change the names of the variables in the two set of sentences.

In order to score the comprehension of the user we decided to use the *SPICE* (Anderson et al., 2016) metric. *SPICE* is obtained by computing the F-score of the overlap between two trees: the overlap is measured by decomposing trees in typed elementary substructures, that are operands, operators and their relations. For instance, the expression $x - 1$ is decomposed as $\{1, x, \text{minus}, (\text{op: minus, first: } x), (\text{op: minus, second: } 1)\}$ (cf. (Anderson et al., 2016) for more details). For the experimentation, we recruited 4 visually impaired people with personal invitation without any rewards. All users are Italian mother tongue, have a good knowledge of mathematical analysis and have a bachelor degree (only one related to mathematics).

In Table 3 we reported the averaged values of *SPICE* for *W-engine* and *E-engine*. A first view of data seems suggest a preference for the *E-engine*, but there is not a significant effect on the performance of the system: by applying the t-test we obtained for 0.08 (two-tailed p-value), indicating no statistical significance. So, new experiments with more trials and users are necessary to statistically confirms the preference of for the *E-engine*.

In Table 4, we report the The distribution of the answers in Likert scale for the question of the web form concerning comprehensibility, that is “Quanto sei d’accordo con la frase: - La frase pronunciata è facile da capire -” (*How much do*

Engine	U1	U2	U3	U4
W-engine	0.96 (0.06)	0.95 (0.12)	0.97 (0.06)	0.97 (0.06)
E-engine	0.99 (0.03)	0.99 (0.03)	0.97 (0.04)	0.97 (0.04)

Table 3: The averaged SPICE measures and standard deviations for the speech synthesis W-engine and E-engine.

Engine	U1	U2	U3	U4
W-engine	4.60 (0.55)	5.20 (1.10)	4.00 (0.71)	4.00 (1.22)
E-engine	4.60 (0.89)	4.00 (1.73)	5.60 (1.14)	4.40 (1.52)

Table 4: The distribution of the answers in Likert scale (1 – 7) for the question concerning comprehensibility.

you agree with the sentence: - *The pronounced sentence is easy to understand* -”). The value 1 corresponds to “per nulla” (*nothing*), the value 7 corresponds to “completamente” (*completely*). It seems from data that there is not notable difference between the perceived comprehensibility of the W-engine with respect to the E-engine and the t-test we obtained for the Likert score is 0.67 (two-tailed p-value).

4 Conclusion

In this paper we have presented a study on the generation of mathematical sentences, i.e. natural language sentences encoding mathematical expressions⁶. In particular, we have described the main features of the system and the a first experimentation centred on the evaluation of two distinct speech engine. The results of the experimentation suggests a good performance of the formant-based synthesis engine with respect to the neural-network base synthesis engine. However, more data is necessary to achieve statistical significance.

In future work we intend to repeat the evaluation of the system for Italian with a larger number of users and to repeat the experiment by using English lanaguage too.

References

- [Ahmetovic et al.2018] Dragan Ahmetovic, Tiziana Armano, Cristian Bernareggi, Michele Berra, Anna Capietto, Sandro Coriasco, Nadir Murru, Alice Ruighi, and Eugenia Taranto. 2018. Axessibility: A latex package for mathematical formulae accessibility in pdf documents. In *Proceedings of the 20th International ACM SIGACCESS Conference on Computers and Accessibility*, ASSETS '18, pages 352–354, New York, NY, USA. ACM.
- [Anderson et al.2016] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 2016. SPICE: semantic propositional image caption evaluation. *CoRR*, abs/1607.08822.
- [Cervone2012] Davide Cervone. 2012. Mathjax: A platform for mathematics on the web. *Notices of the American Mathematical Society*, 59, 02.
- [Chang1983] Lawrence A. Chang. 1983. Handbook for spoken mathematics (larry’s speakeasy). Lawrence Livermore Laboratory, The Regents of the University of California., 1.
- [Conte et al.2017] Giorgia Conte, Cristina Bosco, and Alessandro Mazzei. 2017. Dealing with italian adjectives in noun phrase: a study oriented to natural language generation. In *Proceedings of the Fourth Italian Conference on Computational Linguistics (CLiC-it 2017)*, Rome, Italy, December 11-13, 2017., December.
- [Ferres and Fuentes Sepúlveda2011] Leo Ferres and José Fuentes Sepúlveda. 2011. Improving accessibility to mathematical formulas: the wikipedia math accessor. In *Proceedings of the International Cross-Disciplinary Conference on Web Accessibility, W4A 2011, Hyderabad, Andhra Pradesh, India, March 28-29, 2011*, page 25.
- [Fuentes Sepúlveda and Ferres2012] José Fuentes Sepúlveda and Leo Ferres. 2012. Improving accessibility to mathematical formulas: The wikipedia math accessor. *New Rev. Hypermedia Multimedia*, 18(3):183–204, September.
- [Gatt and Krahmer2018] Albert Gatt and Emiel Krahmer. 2018. Survey of the state of the art in natural language generation: Core tasks, applications and evaluation. *J. Artif. Intell. Res.*, 61:65–170.
- [Gatt and Reiter2009] Albert Gatt and Ehud Reiter. 2009. SimpleNLG: A Realisation Engine for Practical Applications. In *Proceedings of the 12th European Workshop on Natural Language Generation*, ENLG '09, pages 90–93, Stroudsburg,
- ⁶The described system can be freely downloaded at <https://bitbucket.org/tesimagistralemoniticone/formula-to-speech/>

- PA, USA. Association for Computational Linguistics.
- [Ghezzi et al.2018] Ilaria Ghezzi, Cristina Bosco, and Alessandro Mazzei. 2018. Auxiliary selection in italian intransitive verbs: A computational investigation based on annotated corpora. In Proceedings of the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018), pages 1–6, Berlin. CEUR.
- [Hutchins and Somer1992] W. John Hutchins and Harold L. Somer. 1992. An Introduction to Machine Translation. London: Academic Press.
- [Mazzei et al.2016] Alessandro Mazzei, Cristina Battaglino, and Cristina Bosco. 2016. SimpleNLG-IT: adapting SimpleNLG to Italian. In Proceedings of the 9th International Natural Language Generation conference, pages 184–192, Edinburgh, UK, September 5-8. Association for Computational Linguistics.
- [Mazzei2016] Alessandro Mazzei. 2016. Building a computational lexicon by using SQL. In Pierpaolo Basile, Anna Corazza, Francesco Cutugno, Simonetta Montemagni, Malvina Nissim, Viviana Patti, Giovanni Semeraro, and Rachele Sprugnoli, editors, Proceedings of Third Italian Conference on Computational Linguistics (CLiC-it 2016) & Fifth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2016), Napoli, Italy, December 5-7, 2016., volume 1749, pages 1–5. CEUR-WS.org, December.
- [Miller2007] Bruce Miller. 2007. LaTeXXML: A LaTeX to XML converter.
- [Pandolfi2013] Luciano Pandolfi. 2013. ANALISI MATEMATICA 1. Dipartimento di Scienze Matematiche “Giuseppe Luigi Lagrange”, Politecnico di Torino.
- [Raman1996] T. V. Raman. 1996. Emacs-peak—direct speech access. In Proceedings of the Second Annual ACM Conference on Assistive Technologies, Assets '96, pages 32–36, New York, NY, USA. ACM.
- [Reiter and Dale2000] Ehud Reiter and Robert Dale. 2000. Building Natural Language Generation Systems. Studies in Natural Language Processing. Cambridge University Press.
- [Soiffer2016] Neil Soiffer. 2016. A study of speech versus braille and large print of mathematical expressions. In Lecture Notes in Computer Science, volume 9758, Berlin. Springer.
- [Sorge et al.2014] Volker Sorge, Charles Chen, T. V. Raman, and David Tseng. 2014. Towards making mathematics a first class citizen in general screen readers. In Proceedings of the 11th Web for All Conference, W4A '14, pages 40:1–40:10, New York, NY, USA. ACM.
- [Waltraud Schweikhardt2006] Nadine Jessel Benoit Encelle Margaret Gut Waltraud Schweikhardt, Cristian Bernareggi. 2006. Lambda: A european system to access mathematics with braille and audio synthesis. In Lecture Notes in Computer Science, volume 4061, Berlin. Springer.

Automated Short Answer Grading: A Simple Solution for a Difficult Task

Stefano Menini[†], Sara Tonelli[†], Giovanni De Gasperis[‡], Pierpaolo Vittorini[‡]

[†]Fondazione Bruno Kessler (Trento), [‡]University of L'Aquila

{menini, satonelli}@fbk.eu

{giovanni.degasperis, pierpaolo.vittorini}@univaq.it

Abstract

English. The task of short answer grading is aimed at assessing the outcome of an exam by automatically analysing students' answers in natural language and deciding whether they should pass or fail the exam. In this paper, we tackle this task training an SVM classifier on real data taken from a University statistics exam, showing that simple concatenated sentence embeddings used as features yield results around 0.90 F1, and that adding more complex distance-based features lead only to a slight improvement. We also release the dataset, that to our knowledge is the first freely available dataset of this kind in Italian.

1 Introduction

Human grading of open ended questions is a tedious and error-prone task, a problem that has become particularly pressing when such an assessment involves a large number of students, like in an Academic setting. One possible solution to this problem is to automate the grading process, so that it can facilitate teachers in the correction and enable students to receive immediate feedback. Research on this task has been active since the '60s (Page, 1966), and several computational methods have been proposed to automatically grade different types of texts, from longer essays to short text answers. The advantages of this kind of automatic assessment do not concern only the limited time and effort required to grade tests compared with a manual assessment, but include also the reduction of mistakes and bias introduced by humans, as well as a better formalization of assessment criteria.

In this paper, we focus on tests comprising short answers to natural language questions, proposing

a novel approach to binary *automatic short answer grading* (ASAG). This has proven particularly challenging because an understanding of natural language is required, without having much textual context, while grading multiple-choice questions can be straightforwardly assessed, given that there is only one possible correct response to each question. Furthermore, the tests considered in this paper are taken from real exams on statistical analyses, with low variability, a limited vocabulary and therefore little lexical difference between correct and wrong answers.

The contribution of this paper is two-fold: we create and release a dataset for short-answer grading containing real examples, which can be freely downloaded at <https://zenodo.org/record/3257363#.XRsrn5P7TLY>. Besides, we propose a simple approach that, making use only of concatenated sentence embeddings and an SVM classifier, achieves up to 0.90 F1 after parameter tuning.

2 Related Work

In the literature, several works have been presented on automated grading methods, to assess the quality of answers in written examinations. Several types of answers have been addressed, from essays (Kanejiya et al., 2003; Shermis et al., 2010), to code (Souza et al., 2016). Here we focus on works related to short answers, which are the target of our tests. With short answers we refer to open questions, given in natural language, usually with the length of one paragraph, recalling external knowledge (Burrows et al., 2015). When assessing the grading of short answers we face two main issues, *i)* the grading itself and *ii)* the presence of appropriate datasets.

ASAG can be tackled with several approaches, including pattern matching (Mitchell et al., 2002), looking for specific concepts or keywords in the answers (Callea et al., 2001; Leacock and Chodorow, 2003; Jordan and Mitchell, 2009), using bag of

Copyright ©2019 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

words and matching terms (Cutrone et al., 2011) or relying on LSA (Klein et al., 2011). Some other solutions rely more heavily on NLP techniques, for example by extracting metrics and features that can be used for text classification such as the overlap of n-grams or POS between student's and teacher's answers (Bailey and Meurers, 2008; Meurers et al., 2011). Some attempts have been made also to use similarity between word embeddings as a feature (Sultan et al., 2016; Sakaguchi et al., 2015; Kumar et al., 2017).

Another aspect that can affect the performance of different ASAG approaches is the target of automated evaluation. We can for instance assess the quality of the text (Yannakoudakis et al., 2011), its comprehension and summarization (Madnani et al., 2013), or, as in our case, the knowledge of a specific notion. Each task would therefore need a specific dataset as a benchmark. Other dimensions affecting the approach to ASAG and its performance are also the school level for which an assessment is required (e.g. primary school vs. university) as well as its domain, e.g. computer science (Gütl, 2007), biology (Siddiqi and Harrison, 2008) or math (Leacock and Chodorow, 2003). As for Italian, we are not aware of existing automated grading approaches, nor of available datasets specifically released to foster research in this direction. These are indeed the main contributions of the current paper.

3 Task and Data Description

The short grading task that we analyse in this paper is meant to automatize part of the exam that students of Health Informatics in the degree course of Medicine and Surgery of the University of L'Aquila (Italy) are required to pass. It includes two activities: a statistical analysis in R and the explanation of the results in terms of clinical findings. While the evaluation of the first part has already been automatized through automated grading of R code snippets (Angelone and Vittorini, 2019), the second task had been addressed by the same authors using a string similarity approach, which however did not yield satisfying results. Indeed, they used Levenshtein distance to compute the distance between the students' answer and a gold standard (i.e. correct) answer, but the approach failed to capture the semantic equivalence between the two sentences, while focusing only on the lexical one.

For example, an exam provided students with

data about surgical operations, subjects, scar visibility and hospital stay, and asked to compute several statistical measures in R, such as the absolute and relative frequencies of the surgical operations. Then, students were required to comment in plain text on some of the analyses, for example state whether some data are extracted from a normal distribution. For this second part of the exam, the teacher prepared a "gold answer", i.e. the correct answer. Two real examples from the dataset are reported below.

Correct answer pair:

(Student) *Poiché il p -value è maggiore di 0.05 in entrambi i casi, la distribuzione è normale, procediamo con un test parametrico per variabili appaiate.*

(Gold) *Siccome tutti i test di normalità presentano un $p > 0.05$, posso utilizzare un test parametrico.*

Wrong answer pair:

(Student) *Siccome $p < 0.05$, la differenza fra le due variabili è statisticamente significativa.*

(Gold) *Siccome il t-test restituisce un p -value $>$ di 0.05, non posso generalizzare alla popolazione il risultato osservato nel mio campione, e quindi non c'è differenza media di peso statisticamente significativa fra i figli maschi e femmine.*

The goal of our task is, given each pair, to train a classifier and label correct and wrong students' answers. An important aspect of our task is that the correctness of an answer is not defined with respect to the question, which is not used for classification. For the moment we also focus on binary classification, to determine whether an answer is correct or not, without providing a numeric score on how much it is correct or wrong. With the data organized into student-professor answers pairs, the classification is done considering *i*) the semantic content of the answers (represented through word embeddings *ii*) features related to the pair structure of the data such as the overlap or the distance between the two texts. The adopted features are explained in detail in Section 4.1.

3.1 Dataset

The dataset available at <https://zenodo.org/record/3257363#.XR5i8ZP7TLY> has been partially collected using data from real statistics exams

spanning different years, and partially extended by the authors of this paper. The dataset contains the list of sentences written by students, with a unique sentence ID, the type of statistical analysis it refers to (if either given for the hypothesis or normality test), its degree in a range from 0 to 1, and its fail/pass result, flanked with a manually defined gold standard (i.e. the correct answer). The degree is a numerical score manually assigned to each answer, which takes into account whether an answer is partially correct, mostly correct or completely wrong. Based on this degree, the pass/fail decision was taken, i.e. if $\text{degree} < 0.6$ then fail, otherwise pass.

In order to increase the number of training instances and achieve a better balance between the two classes, we manually negated a set of correct answers and reversed the corresponding fail/pass result, adding a set of negated gold standard sentences for a total of 332 new pairs. We also manually paraphrased 297 of the original gold standard sentences, so that we created some additional pairs. Overall the dataset consists of 1,069 student/gold standard answer pairs, 663 of which are labeled as “pass” and 406 as “fail”.

4 Classification framework

Although several works have explored the possibility to automatically grade short text answers, these attempts have mainly focused on English. Furthermore, the best performing ones strongly rely on knowledge bases and syntactic analyses (Mohler et al., 2011), which are hard to obtain for Italian. We therefore test for the first time the potential of sentence embeddings to capture pass or fail judgments in a supervised setting, where the only required data are *a)* a training/test set and *b)* sentence embeddings (Bojanowski et al., 2017) trained using fastText .

4.1 Method

Since we cast the task in a supervised classification framework, we first need to represent the pairs of student/gold standard sentences as features. Two different types of features are tested: **distance-based features**, which capture the similarity of the two sentences using measures based on lexical and semantic similarity, and **sentence embeddings** features, whose goal is to represent the semantics of the two sentences in a distributional space.

All sentences are first preprocessed by removing the stopwords such as articles and prepositions, and by replacing mathematical notations with their transcription in plain language, e.g. “>” with “*maggiore di*” (*greater than*). We also perform part of speech tagging, lemmatisation and affix recognition using the TINT NLP Suite for Italian (Aprosio and Moretti, 2018). Then on each pair of sentences the following distance-based features are computed:

- Token overlap: a feature representing the number of overlapping tokens between the two sentences normalised by their length. This feature captures the lexical similarity between the two strings.
- Lemma overlap: a feature representing the number of overlapping lemmas between the two sentences normalised by their length. Like the previous one, this feature captures the lexical similarity between the two strings.
- Presence of negations: this feature represents whether a content word is negated in one sentence and not in the other. For each sentence, negations are recognised based on the NEG PoS tag or the affix ‘a-’ or ‘in-’ (e.g. *indipendente*), and then the first content word occurring after the negation is considered. We extract two features, one for each sentence, and the values are normalised by their length.

Other distance-based features are computed at sentence level, and to this purpose we employ fastText (Bojanowski et al., 2017), an extension of word embeddings (Mikolov et al., 2013; Pennington et al., 2014) developed at Facebook that is able to deal with rare words by including subword information, and representing sentences basically by combining vectors representing both words and subwords. To generate these embeddings we start from the pre-computed Italian language model trained on Common Crawl and Wikipedia. The latter, in particular, is suitable for our domain, since it includes also scientific content and statistics pages, therefore the language of the exam should be well represented in our model. The embeddings are created using continuous bag-of-words with position-weights, a dimension of 300, character n-grams of length 5, a window of size 5 and 10 negatives.

<https://fasttext.cc/docs/en/crawl-vectors.html>

<https://fasttext.cc/>

Then, the embedding of the sentences written by the students and the gold standard ones are created by combining the word and the subword embeddings with the fastText library. Each sentence is therefore represented through a 300 dimensional embedding. Based on this, we extract four additional distance-based features:

- Embeddings cosine: the cosine between the two sentence embeddings is computed. The intuition behind this feature is that the embeddings of two sentences with a similar meaning would be close in a multidimensional space
- Embeddings cosine (lemmatized): the same feature as the previous one, with the only difference that the sentences are first lemmatised before creating the embeddings
- Word Mover’s Distance (WMD): WMD is a similarity measures based on the minimum amount of distance that the embedded words of one document need to move to reach the embedded words of another document (Kusner et al., 2015) in a multidimensional space. Compared with other existing similarity measures, it works well also when two sentences have a similar meaning despite having few words in common. We apply this algorithm to measure the distance between the solutions proposed by the students and the ones in the gold standard.
- Word Mover’s Distance (lemmatized): the same feature as the previous one, with the only difference that the sentences are first lemmatised before creating the embeddings

The sentence embeddings used to compute the distance features are also tested as features in isolation: a 600 dimensional vector is indeed created by concatenating each sentence embeddings composing a student answer – gold standard pair. This representation is then directly fed to the classifier. We adopt this solution inspired by recent approaches to natural language inference using the concatenation of premise and hypothesis (Bowman et al., 2015; Kiros and Chan, 2018).

As for the supervised classifier, we use support vector machines (Scholkopf and Smola, 2001), which generally yield satisfying results in classification tasks with a limited number of training instances (as opposed to deep learning approaches).

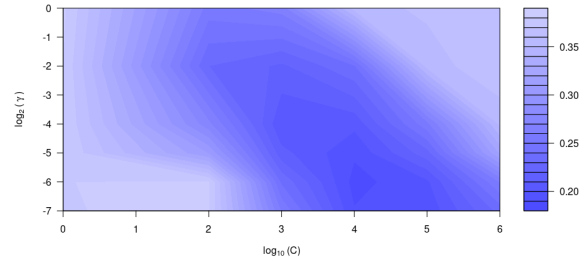


Figure 1: Plot for parameter tuning

We then proceeded to find the best C and γ parameters by means of grid-search tuning (Hsu et al., 2016), through a 10-fold cross-validation to prevent to overfit the model. Finally, with the parameters that returned the best performance, we finalised the classifier and calculated its accuracy and F1 score. The analyses were performed using R 3.6.0 with CARET v6.0-84 and E1071 v1.7-2 packages (R Core Team, 2018).

4.2 Results

Figure 1 shows the plot summarising the tuning process. In summary, within the explored area, the best parameters were found to be $C = 10^4$ and $\gamma = 2^{-6}$. The resulting tuned model produced the following results:

- Accuracy = 0.891 (balanced accuracy = 0.876);
- F1 score = 0.914;

With a similar approach, we also tuned the classifier when fed with only the concatenated sentence embeddings as features (i.e., without distance-based features). With best parameters $C = 10^3$ and $\gamma = 2^{-3}$, the results were:

- Accuracy = 0.885 (balanced accuracy = 0.870);
- F1 score = 0.909;

To evaluate the quality of the model learned with these two configurations, and make sure that it does not overfit, we perform an additional test: we collect a small set of students’ answers from a different statistics exam than the one used to create the training set. This is done on novel data by collecting students’ answers from a small number of new questions, and manually creating new gold answers to be used in the pairs. Overall, we obtain

77 new answer pairs, consisting of 14 wrong and 63 correct answers. We then run the best performing model with all features and using only sentence embeddings (same C and γ as before). The results are the following:

- Accuracy using all features = 0.7838 (balanced accuracy = 0.5965);
- F1 score 0.8710;

while the results achieved using only sentence embeddings are:

- Accuracy = 0.7973 (balanced accuracy = 0.6349);
- F1 score = 0.8780;

5 Discussion

The results presented in the previous section show only a small increase in performance when using the distance-based features in addition to the sentence embeddings after tuning both configurations. This outcome highlights the effectiveness of using sentence embeddings to represent the semantic content of the answers in tasks where student's and gold solutions are very similar to each other. In fact, the sentence pairs in our dataset show a high level of word overlap, and the only discriminant between a correct and a wrong answer is sometimes only the presence of "<" instead of ">", or a negation.

The second experiment, where the same configuration is run on a test set taken from a statistics exam on different topics, shows an overall decrease in performance as expected, but the classification accuracy is still well above the most frequent baseline. In this setting, using only the sentence embeddings yields a slightly better performance than including the other features, showing that they are more robust with respect to a change of topic.

In general terms, despite the accurate parameter tuning, the classification approach seems to be applicable to short answer grading tests different from the data on which the training was done, provided that the student's and gold answer types are the same as in our dataset (i.e. limited length, limited lexical variability).

6 Conclusions

In this paper, we have presented a novel dataset for short answer grading taken from a real statistics exam, which we make freely available. To our

knowledge, this is the first dataset of this kind. We also introduce a simple approach based on sentence embeddings to automatically identify which answers are correct or not, which is easy to replicate and not computationally intensive.

In the future, the work could be extended in several directions. First of all, it would be interesting to use deep-learning approaches instead of SVM, but for that more training data are needed. These could be collected in the upcoming exam sessions at University of L'Aquila. Another refinement of this work would be to grade the tests by assigning a numerical score instead of a pass/fail judgment. Since such scores are already included in the released dataset (the degrees), this would be quite straightforward to achieve. Finally, we plan to test the classifier by integrating it in an online evaluation tool, through which students can submit their tests and the trainer can run an automatic pass/fail assignment.

References

- Anna Maria Angelone and Pierpaolo Vittorini. 2019. The Automated Grading of R Code Snippets: Preliminary Results in a Course of Health Informatics. In *Proc. of the 9th International Conference in Methodologies and Intelligent Systems for Technology Enhanced Learning*. Springer.
- Alessio Palmero Aprosio and Giovanni Moretti. 2018. Tint 2.0: an all-inclusive suite for NLP in Italian. In *Proceedings of the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018)*, Torino, Italy, December 10-12, 2018.
- Stacey Bailey and Detmar Meurers. 2008. Diagnosing meaning errors in short answers to reading comprehension questions. In *Proceedings of the Third Workshop on Innovative Use of NLP for Building Educational Applications*, pages 107–115. Association for Computational Linguistics.
- Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.
- Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal, September. Association for Computational Linguistics.
- Steven Burrows, Iryna Gurevych, and Benno Stein. 2015. The eras and trends of automatic short answer grading. *International Journal of Artificial Intelligence in Education*, 25(1):60–117.

- David H Callear, Jenny Jerrams-Smith, and Victor Soh. 2001. Caa of short non-mcq answers.
- Laurie Cutrone, Maiga Chang, et al. 2011. Auto-assessor: Computerized assessment system for marking student's short-answers automatically. In *2011 IEEE International Conference on Technology for Education*, pages 81–88. IEEE.
- Christian Gütl. 2007. e-examiner: towards a fully-automatic knowledge assessment tool applicable in adaptive e-learning systems. In *Proceedings of the 2nd international conference on interactive mobile and computer aided learning*, pages 1–10. Citeseer.
- Chih-Wei Hsu, Chih-Chung Chang, and Chih-Jen Lin. 2016. A Practical Guide to Support Vector Classification. Technical report, National Taiwan University.
- Sally Jordan and Tom Mitchell. 2009. e-assessment for learning? the potential of short-answer free-text questions with tailored feedback. *British Journal of Educational Technology*, 40(2):371–385.
- Dharmendra Kanejiya, Arun Kumar, and Surendra Prasad. 2003. Automatic evaluation of students' answers using syntactically enhanced lsa. In *Proceedings of the HLT-NAACL 03 workshop on Building educational applications using natural language processing-Volume 2*, pages 53–60. Association for Computational Linguistics.
- Jamie Kiros and William Chan. 2018. Inferlite: Simple universal sentence representations from natural language inference data. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 4868–4874.
- Richard Klein, Angelo Kyrilov, and Mayya Tokman. 2011. Automated assessment of short free-text responses in computer science using latent semantic analysis. In *Proceedings of the 16th annual joint conference on Innovation and technology in computer science education*, pages 158–162. ACM.
- Sachin Kumar, Soumen Chakrabarti, and Shourya Roy. 2017. Earth mover's distance pooling over siamese lstms for automatic short answer grading. In *IJCAI*, pages 2046–2052.
- Matt Kusner, Yu Sun, Nicholas Kolkin, and Kilian Weinberger. 2015. From word embeddings to document distances. In *International Conference on Machine Learning*, pages 957–966.
- Claudia Leacock and Martin Chodorow. 2003. C-rater: Automated scoring of short-answer questions. *Computers and the Humanities*, 37(4):389–405.
- Nitin Madnani, Jill Burstein, John Sabatini, and Tenaha O'Reilly. 2013. Automated scoring of a summary-writing task designed to measure reading comprehension. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 163–168.
- Detmar Meurers, Ramon Ziai, Niels Ott, and Stacey M Bailey. 2011. Integrating parallel analysis modules to evaluate the meaning of answers to reading comprehension questions. *International Journal of Continuing Engineering Education and Life-Long Learning*, 21(4):355–369.
- Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space.
- Tom Mitchell, Terry Russell, Peter Broomhead, and Nicola Aldridge. 2002. Towards robust computerised marking of free-text responses.
- Michael Mohler, Razvan Bunescu, and Rada Mihalcea. 2011. Learning to grade short answer questions using semantic similarity measures and dependency graph alignments. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1, HLT '11*, pages 752–762, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Ellis B Page. 1966. The imminence of grading essays by computer. *The Phi Delta Kappan*, 47(5):238–243.
- Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of EMNLP*.
- R Core Team. 2018. R: A Language and Environment for Statistical Computing.
- Keisuke Sakaguchi, Michael Heilman, and Nitin Madnani. 2015. Effective feature integration for automated short answer scoring. In *Proceedings of the 2015 conference of the North American Chapter of the association for computational linguistics: Human language technologies*, pages 1049–1054.
- Bernhard Scholkopf and Alexander J Smola. 2001. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press.
- Mark D Shermis, Jill Burstein, Derrick Higgins, and Klaus Zechner. 2010. Automated essay scoring: Writing assessment and instruction. *International encyclopedia of education*, 4(1):20–26.
- Raheel Siddiqi and Christopher Harrison. 2008. A systematic approach to the automated marking of short-answer questions. In *2008 IEEE International Multitopic Conference*, pages 329–332. IEEE.
- Draylson M Souza, Katia R Felizardo, and Ellen F Barbosa. 2016. A systematic literature review of assessment tools for programming assignments. In *2016 IEEE 29th International Conference on Software Engineering Education and Training (CSEET)*, pages 147–156. IEEE.
- Md Arafat Sultan, Cristobal Salazar, and Tamara Sumner. 2016. Fast and easy short answer grading with

high accuracy. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1070–1075.

Helen Yannakoudakis, Ted Briscoe, and Ben Medlock. 2011. A new dataset and method for automatically grading esol texts. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 180–189. Association for Computational Linguistics.

Games for Learning Old and Special Alphabets – The Case Study of Gamifying *Mrežnik*

Josip Mihaljević

Institute of Croatian Language and Linguistics

jmihalj@ihjj.hr

Abstract

This paper presents many different custom made web games which are created for learning the Glagolitic script, the sign language, and the Braille alphabet. These games were created within *The Croatian Web Dictionary Project – Mrežnik* where the author works on gamifying dictionary content. The games for learning the Glagolitic script, sign language, and Braille alphabet will be connected to the entries *glagoljica* (the Glagolitic script), *brajica* (Braille alphabet), and the subentry *znakovni jezik* (sign language) of the entry *jezik* (language) in *Mrežnik*. In the paper, each of these games will be presented by stating the game type, mechanics, and gamification elements such as scoring, leaderboards, levels, and badges, etc. The position of these games in the structure of *Mrežnik* will be shown and the reception of the published games through Facebook likes and shares will be presented. For Glagolitic games, a statistical analysis will also be given to show how many players have completed the game, submitted their results, and replayed the game. At the end of the paper technology used for creating, testing, and publishing these games will also be analyzed.¹

1 Introduction

Games have evolved as a new media and are being more and more used in everyday life. What makes a game more engaging than other media is its interactivity with the player. In a game, content is constantly changing based on players reactions in the physical world. With dynamic content and unlimited ability to do different things in the virtual world, games can be used as a powerful tool for educational purposes (Gros, 2007). Some contents in which educational games occur are mili-

tary schools, driving schools, and hospitals which are using virtual simulation to simulate real-life situations. There are also a lot of websites and applications for learning foreign languages such as Duolingo and Memrise. Online dictionaries such as Merriam-Webster and The Free Dictionary have some games for learning definitions, grammar, spelling, etc. The popularity of games for educational purposes in all fields can be attributed to new trends such as e-learning, gamification and game-based learning (Strmečki et al., 2015). The purpose of e-learning methods and techniques is to improve the quality of the class, communication between teachers, instructors, students, and other participants in the learning process, and to allow easier exchange and access to learning material (SRCE, 2016). There is no unique definition of gamification. One of the most quoted papers on gamification (Deterding et al., 2011) *From Game Design Elements to Gamefulness: Defining Gamification* defines gamification as a process which uses the existing game elements in situations which are not considered as a game. Gamification elements, which include scoring, ranking, levels, rewards, ect., are abstracted from many different games. Research has been conducted on the use of gamification mostly in the field of computer science (Ortiz et al., 2016). A study conducted by professor of management Traci Sitzmann (2011) at *Colorado Denver Business School* demonstrates that staff which completed their training with the help of video games learned more facts and accomplished more skills and long-term knowledge than staff that was trained in a less interactive environment. However, there were many critical points about using gamification elements such as leaderboards because some students don't do well when they are compared against others they know and scoring can sometimes be misused,

¹ Copyright © 2019 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

misinterpreted or not implemented correctly. Sometimes assignments are not scored correctly or the games or systems automatically give scores to meaningless actions such as clicking the answers without reading the text (Armando et al., 2018). Gamification can also be used in combination with crowdsourcing where the community can play a certain game in which they solve or offer a solution for certain tasks. This type of gamification used through virtual games is called GWAP (Game with a Purpose) where the player is rewarded with entertainment rather than money (Venhuizen et al., 2013). GWAPs challenge players to score high on specifically designed tasks, thereby contributing their knowledge. GWAPs were successfully pioneered in NLP by initiatives such as ‘Phrase Detectives’ for anaphora resolution (Chamberlain et al., 2008) and ‘JeuxDeMots’ for term relation (Artignan et al., 2009). Venhuizen et al. (2013) have created a gaming platform Wardrobe (wordrobe.org) in which players answer multiple choice questions in which they guess if a certain word in a sentence is a noun or a verb or in another game, they must identify correct senses of a word. Players are through their virtual profile awarded with points and virtual achievements to keep them motivated. Player's answers are used for annotating the text. The amount of points the player gets depends on the agreement with fellow players. The working assumption is that the right sense of a word can be determined by the answers given by the players. The answer which has more selection or is selected by a more experienced player in a game is usually considered to be the correct one. However, that doesn't mean that this system of annotation is good because people tend to have wrong assumptions and make mistakes so in the end the overall data is checked by expert annotators. However, if most of the words are annotated correctly by non-expert annotators it makes the job of checking annotation much quicker for expert annotator (Venhuizen et al., 2013). This is one of the examples of using gamification in NLP.

2 The Croatian Web Dictionary – *Mrežnik*

In the Institute of Croatian Language and Linguistics in Zagreb, the Croatian web dictionary called *Mrežnik* is compiled. *Mrežnik* will be the first web-born dictionary of Croatian. It is corpus-based (based on web corpora: *hrWaC*, *Riznica*

Croatian language corpus), written in TLex² and compiled using Sketch grammar and Word Sketches especially compiled within the *Mrežnik* project. Corpus and word sketches are searchable through Sketch Engine program for corpora managing. *Mrežnik* pays special attention to collocations and examples of word usage extracted from the corpus. It also has external links to different databases and web sites compiled at the Institute. So it contains much more content than digitized versions of paper dictionaries (such dictionaries exist for Croatian on web pages *Hrvatski jezični portal* and *riječnik.hr*). *Mrežnik* consists of three separate modules (the module for adult native speakers of Croatian, the module for elementary school children, and the module for foreigners learning Croatian). The three modules are connected by the fact that the data is coordinated and synchronized. However, each module functions as a separate dictionary compiled for a different target group. The module for adult native speakers of Croatian consists of 10,000 entries. The module for elementary school children consists of 3000 entries, and the module for foreigners consists of 1000 entries. Each dictionary module has a different dictionary grammar which is based on the specific needs of the dictionary user (Hudeček and Mihaljević, 2017). An additional content that is being developed for *Mrežnik* are games which are being placed as external links in certain entries of all three modules. Games compiled for children are e.g. games with fruit, animals, vegetables, professions, etc., games for non-native speakers of Croatian, i.e. foreigners learning Croatian, are e.g. games which help the foreigner produce correct verbal forms or use correct verbal aspect in Croatian, and games for adult native speakers are e.g. games for finding appropriate Croatian words for English loan words often used in Croatian as well as the presented games for learning old and special alphabets. Some of these games have already been published on the portal *Croatian in School* (hrvatski.hr/). This paper will focus on games compiled for learning the old script called Glagolitic and games for learning the sign language and Braille alphabet. These three scripts have been included in the *Croatian Orthographic Manual*, which will also be connected with *Mrežnik*. The structure of the entry *glagoljica* (Glagolitic script) in the module for adult native speakers of Croatian (this module includes children older than 14) is shown below:

² TLex (aka TshwaneLex) is a professional, feature-rich, fully internationalised, off-the-shelf software application

suite for compiling dictionaries or terminology lists. URL: <https://tshwanedje.com/tshwanelex/> (23.9.2019.)

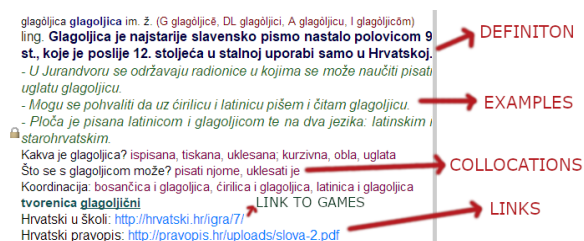


Figure 1: The structure of entry Glagolitic script displayed through TLex program

3 Games for learning the Glagolitic script

The Glagolitic script (Croatian *glagoljica*) is the oldest known Slavic alphabet. It was created in the 9th century by Saint Cyril. The alphabet was created for Slavs in Moravia but was also used in Pannonia, Macedonia, Bulgaria, Bohemia, and Croatia (Damjanović, 2003). After the 12th century, it only survived in Croatia where it was actively used until the middle of the 19th century (Gadžijeva et al., 2014). Today the Glagolitic script is a symbol of national identity and is often used in Croatian art, design, subculture (e.g. Glagolitic tattoos are very popular), and marketing. Although the Glagolitic script is recognized in Croatia and mentioned in schools during history and Croatian language classes, most Croats cannot read or write in the Glagolitic script. Games developed for learning the Glagolitic script focus on recognizing each letter with its Latin equivalent. The two games that will be presented in this paper were created for the Institute of Croatian Language and Linguistics and are published online on 21st February of 2019 on the web site *Croatian in School* (hrvatski.hr/) and advertised on the Institute Facebook page, the day before the official proclamation of the *Day of Croatian Glagolitic Script and Glagoliticism* by the Croatian Parliament. The first game *Glagoljica pamtilica* (engl. *Glagolitic memory*) is a memory game in which the players have to match cards with Glagolitic and Latin letters (hrvatski.hr/games/pamtilica-glagoljica/). At the beginning of the game, the player can choose if he wants to turn on or off the colors for letters. If the player chooses to play with colors they will help him find pairs because the Glagolitic and Latin pairs have the same color. This was done to help the beginners to learn the Glagolitic script. Players who know the Glagolitic script can play without the assistance of colors for matching the pairs. The game also allows players to choose the level of difficulty of the game based on the number of

pairs they want to have (4, 8, 12). Player's results for each game are scored based on the number of tries and the time needed to finish the game. The player can submit his score to online leaderboards by using his written username with a certain emoticon. Emoticons will be displayed next to the username on leaderboards. If the player wins any of the first three places he gets a medal (bronze, silver or gold) and joyful music plays in the background. Leaderboards are different for different levels of difficulty.

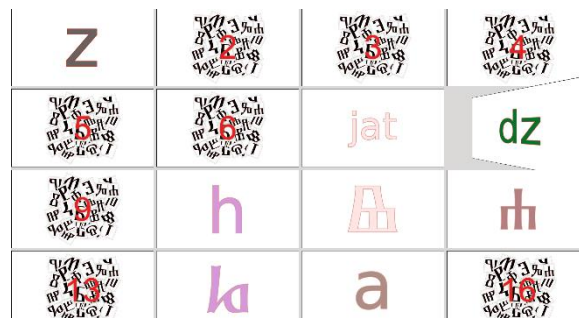


Figure 2: Example of a memory game in which the player connects Glagolitic and Latin letters

The next game *Znam glagoljicu* (engl. I know the Glagolitic script) is a quiz in which the player has 10 seconds to recognize the Glagolitic letter and choose one of the four given answers (hrvatski.hr/games/kviz-glagoljica/). The question is put for all the letters of the Glagolitic script but the order of the letters differs so the correct answer is never the same. The player can choose if he wants to play a game with the angular Glagolitic script, the script that was mostly used in Croatian history or the round Glagolitic script that was more used in Bulgaria than in Croatia. The player always gets feedback after each question. If the answer is correct the feedback will be given on how many points the player has gained on this question. If the player gives the wrong answer, he loses points and get feedback in the form of the correct answer. The quiz also allows players to submit their scores to online leaderboards similar to the previous game. Player's results for each game are scored based on the number of tries and the time needed to finish the game. If a player answers question quicker (e.g. 5 seconds for 20-second question) he gets additional points from remaining time left for answering a question (e.g. gets 15 points for answering a question in 5 seconds). With this type of a scoring system where there are more points, the results from players differ more.

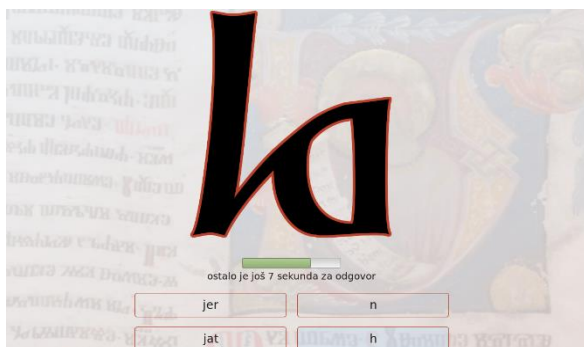


Figure 3: Example of a quiz with a time limit for guessing the Glagolitic letter

The third Glagolitic content present on the site *Croatian in School* is not a game but a web application used to facilitate the learning of writing the Glagolitic letters (hrvatski.hr/games/crtanje-glagoljica/). The user can choose a certain Latin letter for which he will receive a virtual canvas for drawing which displays same Glagolitic letters with reduced opacity. On the canvas, there are some arrows that show directions in which he must drag the mouse or finger on a touchscreen to write the letter correctly. The mentioned game types for learning the Glagolitic alphabet could also be used for learning other alphabets like the Greek alphabet or Chinese or Japanese symbols.

3.1 Analyzing game reception

All three mentioned interactive Glagolitic contents published on the site *Croatian in School* were well received on the Facebook page of the Institute of Croatian Language and Linguistics. They currently have 559 likes (187 on the original post, 372 on the shared post) and 106 shares of the post³. The post about these games is among the most popular posts on the Facebook page reaching more than 16,546 people, only outnumbered by posts on the mobile language advice application that has reached more than 39,623 people, the post about official proclamation of the *Day of the Croatian Glagolitic Script and Glagoliticism* by the Croatian Parliament which has reached 17,701 people and the post about a Croatian language quiz for preparing high school seniors for the state exam in the Croatian language which has reached 26,810 people. The analysis of the database containing the results of the players that have finished the games and submitted their results shows that currently there is a total of 758 submitted results for the memory game and 378 submitted results

for the quizzes. This means that these games have been played many times and by many users. Some of the recorded results were submitted by the same player since the player used the same username for each new round of the game thus showing that he liked the game and enjoyed playing it more than once. In Glagolitic quizzes, there are 195 unique usernames and 107 of those users have submitted their quiz results more than once. Out of those 107 users, 71 of them even submitted results more than twice. The maximum number of times a certain user submitted his score in quizzes is currently 22. In memory games, there are 279 unique usernames. 148 have submitted their results more than once, and 79 of them have submitted results more than twice. The maximum number of times a certain user submitted his score in the memory game is currently 22 times. These games have also been presented on the television show *School hours* on Croatian Radiotelevision Two (skolski.hrt.hr/emisije/1070/jezicne-igrice).

4 Games for learning Braille alphabet and sign language

In addition to different ancient alphabets, the other special letters and symbols that are an important part of human culture and knowledge are the ones made for people with certain disabilities. Braille is a writing system used by people who are visually impaired. It is traditionally written with embossed paper. Braille symbols are formed within units of space known as braille cells. A full braille cell consists of six raised dots arranged in two parallel vertical columns of three dots. 63 combinations are possible using one or more of these six dots. Cells can be used to represent a letter of the alphabet, number, punctuation, part of a word or even the whole word. The braille system was created by Louis Braille, a Frenchman who lost his sight in 1824 as a result of a childhood accident. It is still the most popular writing system for visually impaired people today although its usage has decreased because of the development of screen readers (Wiazowski, 2014). However, it is still largely present in the paper format. There is no substitute for the ability to read, and therefore no digital alternative can replace the braille alphabet completely. Visually impaired people learn braille letters by touch so creating a computer game for them is pointless since it is a visual media. However, teachers, parents, and others who are not visually impaired tend to read braille by sight rather

³ Games for learning Glagolitic script. URL: <https://www.facebook.com/ihjj.hr/photo/a.687321037952455/2715935941757611/?type=3&theater> (21. 6. 2019.)

than by touch. Since some people who are not blind will want to learn braille it is good for them to know how to read the system and explain it to a person who is learning it. That is the reason why braille alphabet was included in the *Croatian Orthography Manual* (Jozić et. al, 2013: 125).

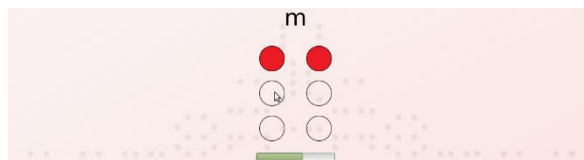


Figure 4: Example of the game in which the player has to press the correct cells to get a certain letter in the Braille alphabet

The game for learning braille is not yet publically available. It is currently stored on GitLab repository, but it can be accessed through this link: bit.ly/2XYHHO1. In the game, each player gets a certain symbol and six empty braille cells. The player has 15 seconds to click on certain braille cells to get the symbol. He can use a hint to know how many cells need to be click on, he can also unselect a cell if he thinks it is not a part of the symbol. The player always gets feedback for his answers. If he answers incorrectly or runs out of time the right answer will be displayed. The symbols are displayed in random order and they include Croatian alphabet, punctuation, and symbols for mathematical operations. The plan is to officially publish this game on 15th October, i.e. on the White Cane Safety Day. Because the game has not been published at the time of writing of this paper, there is no way to evaluate its success yet. The game for learning the sign language alphabet is similar to the quiz game for learning the Glagolitic script. Sign language is a language in which communication is done by using hands and sometimes the movement of other body parts. It is unclear how many sign languages currently exist in the world. Each country generally has its own, native sign language, and some have more than one (Lewis et al., 2013). The game for learning the sign language alphabet only covers the Croatian language. The game is available through this link: ihjj.hr/znakovni. It was officially published on 23rd September 2019, the International Day of Sign Languages. The initial reaction to the game was positive with 337 likes, comments and shares and 321 results submitted to leaderboards. The players can choose if they want to learn one or two hand alphabet. For each question, the player has 10 seconds to answer and he always gets a feedback for his answers.

5 Technology used for creating the games

All of the mentioned games were created for web browsers. They have a responsive design so they can be played even on mobile devices. Game logic and animations were programmed by using jQuery language. Questions, answers, and additional question data were stored in JSON format. Development of these games started on GitLab repositories which allow distributed but private storing of digital files which can be accessed and modified through various computers. GitLab also has a version control system the author could experiment easily while developing games without the fear of spoiling the final product. GitLab also allows users to generate a preview for the websites so they can send private links to testers or publishers. Since there was no database present on the server, the scores and players usernames for games are stored on Google Sheets. The website in the background reads, sorts, and displays data from the sheets so players don't notice that the data is stored elsewhere. One thing these games have to implement in the future is social play in which you can challenge individual opponents through social networks such as Facebook.

6 Conclusion

Game design is constantly evolving so we can expect more educational games in the area of language learning and lexicography. In the paper, some examples of games for learning special alphabets are presented and it is shown how they are incorporated within a dictionary project and received by users. These types of games could be applied for other special alphabets of other languages. The compilation of games for *Mrežnik* is still in progress and many different games for learning grammar, definitions, spelling, etc. are being developed. By gamifying the Croatian dictionary and grammar we can interest children and teenagers but also help foreigners learn Croatian language and culture and popularise language contents. Research on the influence of gamified content on non-native speakers learning Croatian is still in progress within the *Mrežnik* project.

Acknowledgments

This paper is written within the research project Croatian Web Dictionary – *MREŽNIK* (IP-2016-06-2141), financed by the Croatian Science Foundation.

Reference

- Armando Toda, Pedro Henrique Dias Valle and Seiji Isotani. 2018. The Dark Side of Gamification: An Overview of Negative Effects of Gamification in Education. In *Higher Education for All: From Challenges to Novel Technology-Enhanced Solutions*. Springer, page 154. https://doi.org/10.1007/978-3-319-97934-2_9.
- Begoña Gros. 2017. Digital Games in Education: The Design of Games-Based Learning Environments. *Journal of Research on Technology in Education*, 40(1):31-36.
- Daniel Strmečki, Andrija Bernik and Danijel Radošević. Gamification in E-Learning: Introducing Gamified Design Elements into E-Learning System. *Journal of Computer Sciences*, 11(2):1108-1109. <https://doi.org/10.3844/jcssp.2015.1108.1117>.
- Dijana Plantak Vukovac, Maja Škara i Goran Hajdin. 2018. Teachers' Usage and Attitudes Towards Gamification in Primary and Secondary Schools. In *Zbornik Veleučilišta u Rijeci (Volume 6: Long Papers)*. Veleučilišta u Rijeci, page 183. <https://doi.org/10.31784/zvr.6.1.14>.
- Facebook - Institute of Croatian Language and Linguistics. 2019. *Games for learning Glagolitic script*. URL: <https://www.facebook.com/ihjj.hr/photos/a.687321037952455/2715935941757611/?type=3&theater> (21. 6. 2019.)
- Guillaume Artignan, Mountaz Hascoët and Mathieu Lafourcade. 2009. Multiscale visual analysis of lexical networks. In *13th International Conference on Information Visualisation*. Institute of Electrical and Electronics Engineers (IEEE), pages 685–690.
- Jaroslaw Wiazowski. 2014. Can Braille Be Revived? A Possible Impact of High-End Braille and Mainstream Technology on the Revival of Tactile Literacy Medium. *Assistive Technology*, 26(4): 227. <https://doi.org/10.1080/10400435.2014.928389>.
- Jon Chamberlain, Massimo Poesio and Udo Kruschwitz. 2008. Addressing the Resource Bottleneck to Create Large-Scale Annotated Texts. In *Semantics in Text Processing. STEP 2008 Conference Proceedings*. College Publications, pages. 375–380.
- Lana Hudeček, Milica Mihaljević. 2017. The Croatian Web Dictionary Project – *Mrežnik*. In *Electronic lexicography in the 21st century: Proceedings of eLex 2017 conference*. Lexical Computing CZ s.r.o., pages 175–176.
- M. Paul Lewis, Gary F. Simons and Charles D. Fennig. 2013. *Ethnologue: Languages of the World*. SIL International, Dallas
- Margarita Ortiz, Katherine Chiliza and Martin Valcke. 2016. Gamification in Higher Education and STEM: A Systematic Review of Literature. In *Conference: 8th Annual International Conference on Education and New Learning Technologies. Edulearn16*, pages 6548–6558. <https://doi.org/10.21125/edulearn.2016.0422>.
- Noortje Venhuizen et al. 2013. Gamification for Word Sense Labeling. In *The 10th International Conference on Computational Semantics (IWCS)*. Association for Computational Linguistics, pages 397–402.
- School hours. 2019. *Language games*. URL: <https://skolski.hrt.hr/emisije/1070/jezicne-igrice> (21. 6. 2019.)
- Sebastian Deterding, Dan Dixon, Rilla Khaled and Lennart Nacke. 2011. From Game Design Elements to Gamefulness: Defining Gamification. In *Proceedings of the 15th International Academic Mind-Trek Conference: Envisioning Future Media Environments*. ACM, pages 9–15. <https://doi.org/10.1145/2181037.2181040>.
- Sofija Gadžija, Ana Kovačević, Milan Mihaljević, Sandra Požar, Johannes Reinhart, Marinka Šimić, Jasna Vince. 2014. *Hrvatski crkvenoslavenski jezik*. Hrvatska sveučilišna naklada, Staroslavenski institut, Zagreb.
- SRCE: University of computing center. 2016. *Osnove e-učenja*. URL: <https://lms3.srce.hr/module/mod/page/view.php?id=1149> (18. 4. 2019.).
- Stjepan Damjanović. 2003. *Staroslavenski jezik*. Hrvatska sveučilišna naklada, Zagreb
- Traci Sitzmann. 2011. A Meta-analytic Examination of The Instructional Effectiveness of Computer-Based Simulation Games. *Personnel Psychology*, 64(2): 489–528. <https://doi.org/10.1111/j.1744-6570.2011.01190.x>.
- TshwaneDJe Software. 2009. *TLex Lexicography, Terminology and Corpus Software*. URL: <http://tshwanedje.com/> (23. 9. 2019.)
- Željko Jozić et al. 2013. *Hrvatski pravopis*. Institut za hrvatski jezik i jezikoslovlje, Zagreb

Text Frame Detector: Slot Filling Based On Domain Knowledge Bases

Martina Miliani, Lucia C. Passaro and Alessandro Lenci

CoLing Lab, Dipartimento di Filologia, Letteratura e Linguistica (FiLeLi), Università di Pisa

`martina.miliani@fileli.unipi.it`

`lucia.passaro@fileli.unipi.it`

`alessandro.lenci@unipi.it`

Abstract

English. In this paper we present a system called *Text Frame Detector* (TFD) which aims at populating a frame-based ontology in a graph-based structure. Our system organizes textual information into frames, according to a predefined set of semantically informed patterns linking pre-coded information such as named entities, simple and complex terms. Given the semi-automatic expansion of such information with word embeddings, the system can be easily adapted to new domains.

1 Introduction

Textual data are still the most widespread content around the Web (Smirnova and Cudré-Mauroux, 2018). Information Extraction (IE) is a key task to structure textual information and make it machine understandable. IE can be modelled as the process of filling *semantic frames* specified within a domain ontology and consisting of a collection of slots typed with their possible values (Minsky, 1974; Jurafsky and Martin, 2018). Therefore, each frame can be seen as a set of relations whose participants are the values of the slots. Following Jean-Louis et al. (2011), we refer to such relations as complex relations, namely any n -ary relation among typed entities.

Relation extraction techniques have been widely applied to populate semantic frames (Surdeanu, 2013; Zhenjun et al., 2017). However, both supervised and unsupervised methods have shown their limits. On the one hand, supervised approaches (Zelenko et al., 2003; Mooney and Bunescu, 2005; Nguyen and Grishman, 2015; Zhang et al., 2017) model frame filling as a classification task, hence they require labelled data,

with the consequent high cost of long annotation time. On the other hand, unsupervised approaches do not need any training data, but mapping extraction results onto predefined relations or ontologies is often quite challenging with this kind of methods (Fader et al., 2011).

Moreover, semi-supervised methods exploit bootstrap learning, so that any new relation requires a small set of labelled data to be extracted (Agichtein and Gravano, 2000; Chen et al., 2006; Weld et al., 2008).

Finally, another kind of approach has been proposed, which relies on knowledge bases (KBs) to produce training data. Introduced by Mintz et al. (2009), *distant supervision* detects relations on semantically annotated texts where entities which co-occur in the same sentence match with entity-pairs contained in the KB. Then a classifier is trained using features extracted from the annotated relations (Smirnova and Cudré-Mauroux, 2018). Although this approach has been proven to be effective, the supervised step could suffer from scarce amount of data, especially if the relations occur with low frequency in small corpora.

In this paper, we present a system to populate a frame-based ontology, whose values are stored in a graph-based structure. Our method exploits some aspects of distant supervision, leveraging on domain specific KB to infer the relations, and populates the frames with specific information (i.e., the participants) as well as the portions of text (i.e., the snippets) which contain them. Thus, the output of the system for a single frame is a set of snippets, one for each of its slots. Each snippet is also associated with a weight encoding how likely it is expected to contain the information about a certain relation. Such a weight is calculated with a scoring function based on similarity measures and textual distance information. The system has been tested on the administrative domain, with the goal of gathering information related to

Copyright ©2019 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

taxes and agenda events. Indeed, since the KB can be semi-automatically enriched with Named Entities (NEs) and vocabularies of simple and complex terms, our approach can be easily adapted to different domains. Furthermore, system recall can be increased by expanding the frame and attribute vocabulary by exploiting word embeddings (Mikolov et al., 2013).

Our approach differs from existing systems like PIKES (Concoglioni et al., 2016), Framester (Gangemi et al., 2016), FRED (Gangemi et al., 2017), and Framebase (Rouces et al., 2015) primarily for the notion of semantic frame we have adopted. The works above are mainly based on Fillmore’s (1976) definition of frame as encoded in FrameNet: frames and associated roles describe situations evoked by lexical expressions (i.e. *Lexical Units*). In our system a frame represents a domain entity (e.g. “tax”) by means of attributes and relations associated to that domain. Unlike FrameNet frames, these attributes and relations are activated by a set of distributed lexico-syntactic cues.

This paper is structured as follows: in section 2 we describe the general methodology of the system, we define terminology and notation and we describe the main features of the proposed approach. The system implementation is illustrated in section 3, which shows the extraction algorithm as well as the indexing methods in the knowledge graph. Evaluation and results are reported in section 4.

2 Methodology

Following Riedel et al. (2010), we assume that “if two entities $\langle e_1, e_2 \rangle$ participate in a relation $\langle r \rangle$, then there is at least one sentence $\langle s \rangle$ in the text expressing such relation”. We adopt this hypothesis for both simple and complex relations (cf. infra), by considering the sentence $\langle s \rangle$ itself and the $[\langle s - k \rangle, \dots, \langle s + k \rangle]$ adjacent ones, where k is a system parameter.

In order to identify sentences where one or more relations are expressed, we developed a system called *Text Frame Detector* (TFD).

Given a KB where domain terms are associated to a given set of frames, TFD populates them, by making explicit the semantic relation between terms and named entities (NEs). In particular, TFD exploits linguistic analysis and IE algorithms: texts are processed up to part of speech

tagging, then NEs (Passaro et al., 2017) and multiword terms are identified (Passaro and Lenci, 2016). *Co-occurrence Analysis* (Asim et al., 2018) is then performed to identify the participants of each relation by considering terms and NEs co-occurring in the same sentence or in adjacent ones. The relations are filtered and ranked by applying a scoring process (cfr. Section 3.2) to the snippets containing them. The number of slots for each frame is not fixed, therefore we decided to store frames data in the graph-based database (GBD) Neo4j¹. Compared to relational databases, GBDs do not require a pre-defined set of relations, allowing for a more flexible object-oriented data storage. Moreover, GBDs can be updated in real-time and show a better performance in terms of query execution time.

In order to increase the system recall of relevant information, we also used the semantic neighbors of the terms defining the frames. For example, if a text contains the word “versamento” (‘deposit’) but the KB only contains the word “pagamento” (‘payment’), the term “versamento” may be extracted because it is a semantic neighbor of the latter (see Table 1).

Neighbor	Cosine Similarity
rimborso (‘refund’)	0.89
versamento (‘deposit’)	0.86
versare (‘to deposit’)	0.78

Table 1: Semantic neighbors of “pagamento” (‘payment’) and their cosine similarity score.

We trained *fastText* word embeddings (Bojanowski et al., 2017) on a combination of La Repubblica corpus (Baroni et al., 2004) and PAWAC (Passaro and Lenci, 2016) for administrative domain specific knowledge.

Currently, KB terms are expanded with their 10 nearest semantic neighbors in terms of cosine similarity, which can be filtered through a parametric threshold.

2.1 Definitions and terminology

Frame: Terms and entities contained in the KB are organized in frames. Frames allow to structure the implicit knowledge contained in texts around concepts that define the relevant semantic categories in a domain. For instance, the frame EVENT corresponds to en-

¹<http://neo4j.com/>

titles like concerts, shows, etc. Each frame is defined by its *frame triggers* and *attributes*.

Frame trigger: It corresponds to an instance of the semantic class described by the frame (e.g., in the administrative domain, the frame TAX is expressed by its instances: “TARI” (‘Garbage tax’), “IMU” (‘Municipal tax’)). Frame triggers suggest the presence of frame attributes in the text.

Attribute: A frame is composed by a set of slots, which must be filled by specific instances or data (Minsky, 1974). Each slot value is a participant in a relation with the frame trigger. This relation is referred to as an “attribute”, and describes an aspect of the concept represented by the frame. For instance, the EVENT frame, requires the following attributes: **when**, to be filled with time and date, **where**, which corresponds to a location and **cost**, such as the ticket price. Depending on the way they are expressed in texts, we distinguish between *simple attributes* and *complex attributes*.

Simple attribute: Their values correspond to simple and complex terms, NEs or Temporal Expressions (TEs) identified during the IE step. The EVENT frame attributes are considered simple because they usually appear right near the frame trigger (cfr. Figure 1).

Complex attribute: The values of these attributes do not correspond to a single entity, but are expressed by whole text segments. Concerning the TAX frame, the **deadline** attribute cannot be filled by simply extracting the due dates from the text, because the reported information would be incomplete if taken out of context (cfr. Figure 2). Therefore, it is necessary to return the entire text snippet, which includes the *attribute triggers* that allow to identify the complex attribute.

Attribute trigger: They represent the linguistic cues of an attribute instance. They are manually selected by domain experts and stored in the KB with a standard form t and a small number of orthographic and morphosyntactic variants v . Attribute triggers can be: (i) single and multiword terms, like “bollettino postale” (‘postal order’), “saldo” (‘balance’),

NEs, such as “Firenze” (‘Florence’) or TEs, like “18 giugno” (‘18th June’); (ii) complex patterns, such as “non inferiore a” (‘not lower than’).

3 Implementation

In order to fill the frame slots, textual data are analyzed by TFD in various steps. After linguistic annotation, NER, and term extraction, TFD looks for frame triggers and for its attribute triggers, in the same sentence or in the sentences around it. More specifically, given a snippet, a frame instance F is expressed by a frame trigger F_t , and a set of attributes A , containing both simple (A_s) and complex (A_c) attributes, so that $F = \{F_t, A\}$ where $a_i \in A_s \cup A_c$.

3.1 Frame and attribute retrieval

Since both simple and complex attributes of a frame are expressed by means of the set T of their attribute triggers, we can say that F is instantiated in a text by the joint occurrence of a frame trigger F_t and a set of attribute triggers T related to one or more of its attributes, namely $F = \{F_t, T\}$ where $T = \{t_1, \dots, t_n\}$.

In order to retrieve a frame F in a portion of text, first of all we look for its frame triggers. Once a F_t has been detected, we search for its potential attributes. Given such F , its potential instances in the text consist of the co-occurrence of F_t and a subset of T . To guarantee a certain degree of flexibility, we decided to provide each of the elements in T with a binary feature that can be set to 1 if the attribute trigger t_i is mandatory to extract the F , and to 0 if the attribute trigger is optional. A further implementation could consider to convert these features in continuous weights. In this way the TFD would be able to consider some triggers as more relevant than others to populate the frame.

Moreover, the attribute triggers of F belonging to T are selected within terms and entities used to express its attribute instances. Such triggers are then exploited by the attribute retrieval system of the TFD. Concerning the retrieval of simple attributes, see the extraction of the EVENT frame from the sentence in Figure 1.

The trigger for the EVENT frame (“spettacolo di Roger Waters”) in Figure 1 is a clue for the presence of its attributes which populate the frame instance showed in Table 2.

Moreover, the TFD stores the raw text in Figure 1 as the relevant snippet for both the attributes

Lo [spettacolo di Roger Waters]*nome_evento*
si terrà il [26 giugno]*data* allo [stadio di
Firenze]*luogo*.

Figure 1: Example of a snippet (‘Roger Waters’ show will take place on 26th June at the Florence Stadium’) containing simple attributes.

EVENT	spettacolo di Roger Waters
when	26 giugno
where	Stadio di Firenze
cost	-

Table 2: An instance of the EVENT frame.

when and **where**.

Il [versamento]*pagamento* dell’[IMU]*tassa*
deve essere effettuato con [bonifico
bancario]*mod_pagamento* o [bollettino
postale]*mod_pagamento* in due [rate]*somma*:
l’[acconto]*somma* entro il [18 giugno]*data* e il
[saldo]*somma* entro il [17 dicembre]*data*.

Figure 2: Example of a snippet (‘The Municipality tax disbursement must be made through wire transfer or postal order in two installments: down payment by June 18th and balance by December 17th’) containing complex attributes.

Examples of complex attributes can be found in the TAX frame, namely **deadline**, indicating the due date of the tax payment, and **methods of payment**, indicating how it is possible to pay it. For example, the triggers detected for the attribute **deadline** in Figure 2 are “somma” (‘sum’), “pagamento” (‘payment’) and two TEs, namely “18 giugno” (‘June 18th’) and “17 dicembre” (‘December 17th’). The snippet contains also the attribute **methods of payment**, which is expressed by the triggers “pagamento” (‘payment’) and “mod_pagamento” (‘methods_payment’), expressed by “bonifico bancario” (‘wire transfer’) and “bollettino postale” (‘postal order’). Table 3 shows the TAX frame instantiated with the extracted attributes. Also in this case, the full snippet (the raw text in Figure 2) is stored for both the attributes **deadline** and **methods of payment**.

3.2 Snippet selection and ranking

The binary features associated to each attribute trigger in a frame instance lead also the snippet

TAX	IMU
deadline	18 giugno, 17 dicembre
methods of payment	bonifico bancario, bollettino postale

Table 3: An instance of the TAX frame.

selection and ranking system. Given a potential instance of a frame, its attribute triggers are associated with a binary feature indicating their compulsory presence in order to associate the attribute with a certain snippet. On the basis of how many features are set to 1, the TFD will be more or less strict in the selection phase. For example, given the following sentences, where the frame triggers appear in bold and attribute triggers are underlined (the standard form for “pagata” is “pagamento” and “17 giugno” is marked as “data”), Table 4 shows which snippets are extracted according to the binary values associated to each attribute trigger.

- A “L’**IMU** va pagata entro il 17 giugno” (‘The Municipality tax must be paid before June 17th’)
- B “La scadenza dell’**IMU** è fissata al 17 giugno” (‘The deadline for the Municipality tax payment is on June 17th’)

Line ID	pagamento (‘payment’)	scadenza (‘deadline’)	data (‘date’)	snippet extracted
1	0	0	0	A,B
2	0	0	1	A,B
3	0	1	0	B
4	0	1	1	B
5	1	0	0	A
6	1	0	1	A
7	1	1	0	-
8	1	1	1	-

Table 4: Mandatoriness of attribute triggers.

Each line of the table represents a potential combination of attribute triggers, with the respective mandatoriness. According to these features, the absence of mandatory attribute triggers (line 1) allows the retrieval of both the snippets A and B. Otherwise, if the system is expected to find all the attribute triggers (line 8), none of the two snippets is extracted because “pagamento” and “scadenza” never appear in the same sentence. This system is useful in order to balance the extraction flexibility based on the domain. For example, in administrative documents, where the language is bounded to stereotyped phrases (Brunato, 2015) a more strict approach is preferable, whereas in general domain ones it might be better to work with a higher number of optional triggers.

Moreover, a second objective of the TFD is to rank the extracted snippets according to their relevance with respect to a given attribute. Such relevance is calculated through a co-occurrence analysis, which employs measures based on semantic and distance features. One of these measures is the *Sentence score*, defined as:

$$SS = |t| \times |v| \quad (1)$$

where t is the number of attribute triggers (standard forms) and v is the total of their variants.

This formula takes into account the ratio between the number of attribute triggers and their variants. In particular, the TFD favours the snippets containing the highest number of distinct attribute triggers, namely their standard forms. In the case of simple attributes, t represents the number of entity types and v the number of NEs.

Furthermore, although different frame triggers may be found all over a given document, they may refer to the same domain entity, hence to the same frame instance. For example, we observed that Italian municipality web pages dedicate entire articles to a single tax, which can be mentioned in different ways, such as their full names and their acronyms (e.g., the Italian Tax “Imposta Municipale Propria” (‘Municipality tax’) can be mentioned also with the acronym, “IMU”). In order to avoid that attributes belonging to the same frame are associated to different ones and affect the scoring process, our system can be set to apply a “fuzzy normalization” strategy that is able to associate all the triggers of a document to a frame referring to the same entity. For example, the snippets extracted from a municipality web page and associated to the **deadline** attribute of the TAX frame can be ranked together, regardless the frame triggers they contain, such as “Imposta Municipale Propria” (‘Municipality tax’) or its acronym, “IMU”.

At a document level, the snippet selected is simply the one with the highest *Sentence Score*, but we provide an additional level of analysis, which is applied when the snippet has to be chosen within a group of documents, instead of a single one. In that case, TFD selects the snippet with the highest *Document score* (DS), which encodes how likely the document contains a relevant information about a certain attribute. The Document score is calculated as follows:

$$DS = \frac{\sum_{i=1}^n TS}{l} \quad (2)$$

where l is the sentence length in terms of tokens, and TS is the *Trigger score* of a given variant v . TS is defined as:

$$TS = \frac{1}{d} \times \cos \quad (3)$$

where d is the distance between the attribute trigger (or NEs) and the frame trigger, and \cos is the cosine similarity between the trigger variant contained in the KB and the neighbor found in the text (the cosine is equal to 1 for the KB terms).

3.3 Storage

Extracted frame instances are stored in a Neo4j GDB. The Knowledge Graph (KG) contains several root nodes, one for each of the frames detected in the document or in the collection of documents (Figure 3).

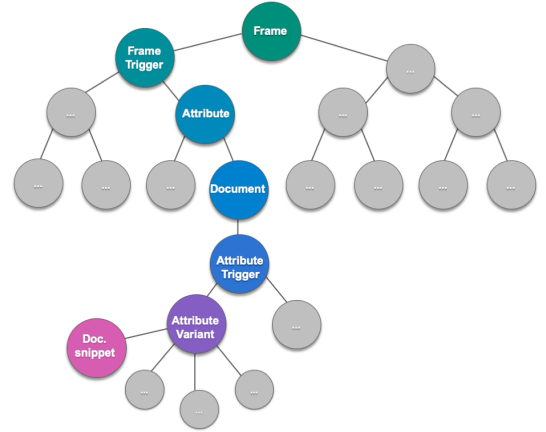


Figure 3: Information levels in the Knowledge Graph.

For instance, there are two root-nodes corresponding to the EVENT and TAX frames. If we consider the frame TAX (the node “Frame” in Figure 3), the nodes “Frame Trigger” can be populated with instances like “Imposta Municipale Propria” (‘Municipality tax’) or its acronym, “IMU”. Each frame trigger node is linked to the corresponding frame attributes (“Attribute” node in Figure 3) which can be populated with information like “scadenza” (‘deadline’) and “modalità di pagamento” (‘methods of payment’). Document-nodes (“Document” node in Figure 3), labelled by document names, are placed between attribute-nodes and attribute-trigger-nodes in order to facilitate the retrieval phase. Each document node

is associated with the snippet having the highest *Sentence score* for the connected attribute-node (e.g., ‘deadline’), along with its *Document score*. In the retrieval phase, unless the information is extracted from a single document, the snippet with the higher *Document score* is selected and returned (see Section 3.2). The other levels of the graph contain information extracted from each document. Every attribute-trigger-node (“Attribute Trigger” node in Figure 3) is labelled by the standard form of the attribute trigger extracted from the connected document-node (e.g., ‘sum’). Then, each attribute-trigger-node is connected to one or more nodes representing the trigger variants (“Attribute Variant” node in Figure 3). Continuing with this example, attribute variants can consist in ‘installments’, ‘balance’ and ‘down payment’. Finally, the last node of the graph consists of the snippet-node (“Doc. snippet” node in Figure 3), storing the snippet containing the information extracted. For example, the node can be populated with a snippet like the one reported in Figure 2: “Il versamento dell’IMU deve essere effettuato con bonifico bancario o bollettino postale in due rate: l’acconto entro il 18 giugno e il saldo entro il 17 dicembre” (‘The Municipality tax disbursement must be made through wire transfer or postal order in two installments: down payment by June 18th and balance by December 17th’).

4 Evaluation and Results

The extraction of attributes related to TAX and EVENT frames were evaluated on Italian language texts by an administrative domain expert. We decided to evaluate these frames because the first one is very specific of the administrative domain, whereas the second one can be seen as a general purpose one. The gold standard includes both administrative documents as well as social media texts and news published on the municipalities websites. Both frames were evaluated on 50 texts, including information about taxes (municipality online guidelines), events (administrative acts, press releases, Facebook statuses and tweets) and other topics (municipality web pages). For municipality guidelines web pages, the “fuzzy normalization” strategy has been applied (see Section 3.2). The results of the TFD are shown in Table 5.

Since simple attribute values consist mostly of NEs, these results are strictly dependent on the

Frame	Precision	Recall	F1
TAX	0.771	0.519	0.621
EVENT	0.808	0.955	0.875
Total	0.799	0.793	0.796

Table 5: TFD evaluation results.

generalization capability of the models used to extract those entities. In other cases, a wrong snippet is selected as relevant for an attribute, although triggers and NEs are correctly annotated and extracted. Moreover, additional errors depend on the absence of attribute triggers variants in the Knowledge Graph.

More specifically, errors are mainly related to a wrong NE annotation (35%). In the 22.8% of cases, a wrong sentence is selected as relevant for a certain attribute, although triggers and NEs are correctly annotated and extracted. False negative errors are caused by relevant information spread in several sentences (8.8%), whereas each extracted snippet consists of a single sentence, by unknown triggers describing an attribute (7.5%), by partial information contained in the extracted sentence (5%), by wrong lemmatization (1.75%) or by the overlapping of named entities and events (1.75%) (e.g., ‘Roger Waters’ show’ is not annotated as an event, however ‘Roger Waters’ is extracted as a named entity). In other cases (3.5%), attribute triggers are too distant from their frame trigger to be extracted. Although this span is customizable, an excessive distance between frame and attribute triggers could produce noise in the retrieval phase. Finally, the application of the “fuzzy normalization” strategy (see Section 3.2) led to errors in the ranking phase (14.3%). One of the municipality web pages in which the strategy has been applied contained information on more than one tax, but only one frame instance has been returned. This kind of errors can be limited by automatically checking the frame triggers cited on the text, and deciding whether applying or not the normalization according to external lexical resources, such as gazetteers or dictionaries.

5 Conclusions

In this paper we presented a domain independent system for slot filling that exploits a graph to populate a frame-based ontology. The Text Frame Detector extracts a relevant snippet for each frame attribute from textual information with good results in terms of *F1 score* (0.796). Nonetheless, the

evaluation showed that there is room for improvement in some of the TFD modules. For example, the annotation of the semantic neighborhood of single and multiword terms, which are particularly relevant in technical domains, should lead to further improve recall performances for complex attributes.

Moreover, although we did not adopt Fillmore's semantic frames in the present work, we would like to explore the possibility of integrating our domain frames with FrameNet ones, which might contribute to enhance the system flexibility.

Finally, in the near future, we plan to fine-tune parameters and to implement additional features such as to associate multiple snippets to the same attribute. Furthermore, we intend to convert the binary features used in the snippet selection system into continuous weights. These weights, along with the collected data about frame population, would be also employed to train a supervised model for slot filling, in order to test TFD across new domains.

Acknowledgments

This research has been funded by the Project "SEM il Chattadino" (SEM), funded by Regione Toscana (POR CreO Fesr 2014-2020). The project brings together the CoLing Lab and the companies ETI3 s.r.l. (coordinator), BNova s.r.l. and Rigel Engineering s.r.l.

References

- Eugene Agichtein and Luis Gravano. 2000. Snowball: Extracting relations from large plain-text collections. In *Proceedings ACM 2000, the fifth conference of the Association for Computing Machinery on Digital libraries*, pages 85–94, New York, NY, USA.
- Muhammad Nabeel Asim, Muhammad Wasim, Muhammad Usman Ghani Khan, Waqar Mahmood, and Hafiza Mahnoor Abbasi. 2018. A survey of ontology learning techniques and applications. *Database: the journal of biological databases and curation* 2018.
- Marco Baroni, Silvia Bernardini, Federica Comastri, Lorenzo Piccioni, Alessandra Volpi, Guy Aston, and Marco Mazzoleni. 2004. Introducing the la repubblica corpus: A large, annotated, TEI(XML)-compliant corpus of newspaper Italian. In *Proceedings LREC'04, the fourth International Conference on Language Resources and Evaluation*, Lisbon, Portugal. European Language Resources Association (ELRA).
- Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5(Dec):135–146.
- Dominique Brunato. 2015. *A Study on Linguistic Complexity from a Computational Linguistics Perspective. A Corpus-based Investigation of Italian Bureaucratic Texts*. Ph.D. thesis, Università di Siena.
- Jinxu Chen, Donghong Ji, Chew Lim Tan, and Zhengyu Niu. 2006. Relation extraction using label propagation based semi-supervised learning. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 129–136, Sydney, Australia. Association for Computational Linguistics.
- Francesco Concoglioni, Marco Rospocher, and Alessio Palmero Aprosio. 2016. Frame-based ontology population with pikes. *IEEE Transactions on Knowledge and Data Engineering*, 8(12):3261–3275.
- Anthony Fader, Oren Etzioni, and Stephen Soderland. 2011. Identifying relations for open information extraction. In *Proceedings of EMNLP 2011, the Conference on Empirical Methods in Natural Language Processing*, pages 1535–1545, Edinburgh, Scotland, UK.
- Charles J. Fillmore. 1976. Frame semantics and the nature of language. *Annals of the New York Academy of Sciences: Conference on the origin and development of language and speech*, 280(1).
- Aldo Gangemi, Mehwish Alam, Luigi Asprino, Valentina Presutti, and Diego Reforgiato Recupero. 2016. ramerster: a wide coverage linguistic linked data hub. In *Proceedings European Knowledge Acquisition Workshop*, Cham. Springer.
- Aldo Gangemi, Valentina Presutti, Diego Reforgiato Recupero, Andrea Giovanni Nuzzolese, Francesco Draicchio, and Misael Mongiovi. 2017. Semantic web machine reading with fred. *Semantic Web*, 8(6):873–893.
- Ludovic Jean-Louis, Romaric Besançon, and Olivier Ferret. 2011. Text segmentation and graph-based method for template filling in information extraction. In *Proceedings of IJCNLP 2011, the fifth International Joint Conference on Natural Language Processing*, pages 723–731, Chiang Mai, Thailand.
- Dan Jurafsky and James H. Martin. 2018. Speech and language processing. Third edition draft on webpage: <https://web.stanford.edu/~jurafsky/slp3/>. Accessed: 3 July 2019.
- Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Proceedings of NIPS 2013, 26th Conference*

- on *Advances in Neural Information Processing Systems*, pages 171–178, Lake Tahoe, Nevada, USA.
- Marvin Minsky. 1974. *A framework for representing knowledge*. Massachusetts Institute of Technology, Cambridge, MA.
- Mike Mintz, Steven Bills, Rion Snow, and Daniel Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 1003–1011, Suntec, Singapore. Association for Computational Linguistics.
- Raymond J. Mooney and Razvan C. Bunescu. 2005. Subsequence kernels for relation extraction. In *Proceedings of NIPS 2005, 18th Conference on Advances in Neural Information Processing Systems*, pages 171–178, Vancouver, British Columbia, Canada.
- Thien Huu Nguyen and Ralph Grishman. 2015. Relation extraction: Perspective from convolutional neural networks. In *Proceedings of VS@NAACL-HLT 2015, the 1st Workshop on Vector Space Modeling for Natural Language Processing*, pages 39–48, Denver, Colorado.
- Lucia C. Passaro and A. Lenci. 2016. Extracting terms with Extra. *Computerised and Corpus-based Approaches to Phraseology: Monolingual and Multilingual Perspectives*, pages 188–196.
- Lucia C. Passaro, Alessandro Lenci, and Anna Gabbolini. 2017. Informed pa: A ner for the italian public administration domain. In *Proceedings of Clic-It 2017. The fourth Italian Conference on Computational Linguistics*, pages 246–252, Rome, Italy.
- Sebastian Riedel, Limin Yao, and Andrew McCallum. 2010. Modeling relations and their mentions without labeled text. In *Proceedings of ECML PKDD 2010, the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases*, pages 148–163, Barcelona, Catalonia, Spain. Springer.
- Jacobo Rouces, Gerard De Melo, and Katja Hose. 2015. Framebase: Enabling integration of heterogeneous knowledge. In *Proceedings European Semantic Web Conference*, Cham. Springer.
- Alisa Smirnova and Philippe Cudré-Mauroux. 2018. Relation extraction using distant supervision: A survey. *ACM Computing Survey*, 51(5):1–35.
- Mihai Surdeanu. 2013. Overview of the tac2013 knowledge base population evaluation: English slot filling and temporal slot filling. In *Proceedings of TAC 2013, the Sixth Text Analysis Conference*, Gaithersburg, Maryland USA.
- Daniel S. Weld, Raphael Hoffmann, and Fei Wu. 2008. Using wikipedia to bootstrap open information extraction. *SIGMOD record*, 37(4):62–68.
- Dmitry Zelenko, Chinatsu Aone, and Anthony Richardella. 2003. Kernel methods for relation extraction. *Journal of machine learning research*, 3(Feb):1083–1106.
- Meishan Zhang, Yue Zhang, and Guohong Fu. 2017. End-to-end neural relation extraction with global optimization. In *Proceedings EMNLP 2017, conference on Empirical Methods in Natural Language Processing*, pages 1730–1740, Copenhagen, Denmark.
- Ming Zhenjun, Yan Yan Guoxin Wang, Janet K. Allen Joseph Dal Santo, and Farrokh Mistree. 2017. An ontology for reusable and executable decision templates. *Journal of Computing and Information Science in Engineering*, 17(3):031008.

Defining Action Types: on the roles of Thematic Structures and Argument Alternations

Massimo Moneglia
Università di Firenze

Alessandro Panunzi
Università di Firenze

Rossella Varvara
Università di Firenze

massimo.moneglia@unifi.it alessandro.panunzi@unifi.it rossella.varvara@unifi.it

Abstract

English. The paper focuses on the relation between Action Types (ontological objects that identifies the referential potential of a verb) and the Thematic Structure and Alternations of verbs. The empirical analysis shows that these linguistic features are not properties of the verb itself, but vary in relation to its referential variation. Given this evidence, we argue that Thematic Structure and Argument Alternation can help in the identification of the different types of action to which a verb refers, providing evidences to define the granularity of action concepts in the development of an ontology of actions.

1 Introduction

Action verbs are among the most frequent words in ordinary communication, and their correct processing constitutes an underpinning element for a wide series of human-machine interaction tasks. The formalization of action verb meanings has often been linked to propositional representations within decompositional approaches (Dowty, 1979; Rappaport Hovav and Levin, 2012), in which the semantic core of the verb remains a non-analyzed lexical root. Other traditional components used in the representation and annotation of the meaning of action verbs are: the temporal and aspectual properties of verbs (Vendler, 1957; Pustejovsky, 1991); the thematic roles of participants (Fillmore, 1967; Gruber, 1965); the force dynamics and causal relations implied (Talmy, 1988; Croft, 2012; Gärdenfors, 2014). Nevertheless, even if these semantic components are usually assumed to reflect the general structure

of action conceptualization, the linguistic and the cognitive levels of categorization are not equivalent and should not be confused (Croft, 2012; Moneglia, 2014). As a matter of fact, the lexical category instantiated by an action verb can refer to more than one cognitive entity.

For instance, the verb *to push* can refer to actions in which the force causes the movement of the object in space (e.g. in a sentence like *John pushes the basket under the table*), as well as to actions in which the object does not move (e.g. *John pushes the fabric into a ball*). This differential property is more than enough to cognitively distinguish these events in different action concepts. As a consequence, the need for a cognitive level of action categorization which is independent from the lexical one becomes clear.

In this paper, we investigate the role of one type of linguistic information, specifically Thematic Structure and Argument Alternations, in the definition of action types, i.e. types of action concepts that gather together single datapoint in the IMAGACT ontology of actions. We point out that Thematic Structure is not a property of the verb itself, since different structures may be present for the same verb. Our aim is to show how these features are linked to action types and how this correlation can be useful for the definition and the induction of Action Types¹.

In section 2, we show the innovative methodology assumed by the IMAGACT Ontology of Action for representing the meaning of action verbs, focusing on their referential properties rather than on their intensional definition. In sections 3 and 4, we will see through a case study that the induction of the referential variation of verbs can take advantage of linguistic features. Thematic Struc-

Copyright ©2019 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

¹Similarly, previous work in Word Sense Disambiguation (Dang and Palmer, 2005; Roberts and Kordoni, 2012) have shown that thematic information can improve verb sense disambiguation.

tures and their Alternations can have an impact in the definition and characterization of the objects in an ontology of actions.

2 The IMAGACT ontology

In the IMAGACT multilingual Ontology of Actions² (Moneglia et al., 2012b; Panunzi et al., 2014) action concepts are not defined through a propositional and truth conditional perspective, but they are rather identified and visually represented through scenes. Each scene is conceived as a prototypical instance (Rosch, 1983) of an action concept and constitutes the basic entity of reference of the action ontology. Scenes have been derived from a complex annotation procedure (Moneglia et al., 2012a) of the occurrences of action verbs³ in two large spoken resources of English and Italian. After this bootstrapping phase, the ontology has been extended to many other languages exploiting competence judgments by native speakers (Brown et al., 2014; Pan et al., 2018; Moneglia et al., 2018b). The whole IMAGACT database is currently comprised of 1,010 scenes linked to more than 8,700 verbs in 13 languages⁴. As a result, action concepts have been represented by language independent scenes, each one linked to the series of verbs that can be used to refer to it. The scenes are described by linguistic captions (i.e. short sentences) that have as predicates each of those verbs. Simultaneously, each verb is connected to a set of scenes in the ontology, representing in this way its referential variation.

The scenes linked to a verb have been then grouped in broader categories called Action Types⁵ (hereafter also *ATs* or *Types*). ATs are defined as higher level concepts which fall in the semantic variation of a verb, useful to represent its referential potential in a more compact way, reducing an excessive granularity in the representation of meaning⁶. ATs have been created exploiting similarity judgments among scenes and considering Local Equivalent verbs, i.e. all the verbs

that could co-extend to the same scene (Moneglia et al., 2018a). An additional validation, in which raters were asked to assign scenes to ATs, was conducted with an overall agreement of 0.8 (Gagliardi, 2014). Lastly, during the ontology's development, Thematic Structure, Alternations and Aktionsart were manually annotated for the linguistic captions of each scene. These latter annotations will be the starting point of the present study, in which we analyze the correlation between ontological entities (ATs) and linguistic features, specifically Thematic Structure and Alternations.

3 Thematic Structure and Action Types

Thematic structure and syntactic frame information of verbs are usually provided by most lexical resources, such as VerbNet (Kipper-Schuler, 2005), FrameNet (Fillmore et al., 2004) and PropBank (Palmer et al., 2005). In these resources, the different entries of a verb are associated to their possible thematic structures. They include manually annotated data and have been useful for the development of statistical approaches for Semantic Role Labeling (Gildea and Jurafsky, 2002) and for various NLP applications (e.g. information extraction (Surdeanu et al., 2003), summarization (Melli et al., 2006), and machine translation (Boas, 2002)).

In this section, we show that Thematic Structure (TS) is not a property of the verb and we will verify: 1) to what extent it can be considered a property of the action types in the variation of a verb; 2) to what extent it can provide a differential feature for the identification of ATs. We consider as TS the minimal thematic structure⁷ which is necessary to interpret a verb as an instance of a specific AT.

There are cases in which the TS is the same all through the verb variation. Frequently, one specific thematic structure is associated to activity verbs that show almost no variation in their meaning. This is the case of the verb *to drink*, who has only one AT. The verb *to close*, on the contrary, shows a significant variation in the IMAGACT ontology (7 ATs, four of them represented in table 1), but all types present the same TS (AG-V-TH). In these cases, thematic structure cannot play any role in the definition of different types, which are

²Freely accessible at <http://www.imagact.it/>

³Only in their basic, physical meaning, so excluding all metaphorical and phraseological uses.

⁴Besides English and Italian, the list of fully mapped language comprehends: Arab, Chinese, Danish, German, Hindi, Japanese, Polish, Portuguese, Serbian, Spanish, Greek.

⁵See, for instance, Table 1 which represents the main variation of the action verb *to close*.

⁶As a matter of fact, some verbs in IMAGACT can be linked to several dozen scenes, and the most general ones, like *to take* and *to put*, are linked to about 100 scenes.

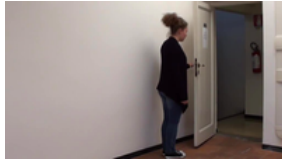
⁷The set of roles used in IMAGACT is based mainly on the set used in VerbNet.



Mary closes her hand



Mary closes the umbrella



Mary closes the door



Mary closes the lock

Table 1: Variation of *to close*

identified on the basis of ontological features of the theme (e.g. a body part vs an artifact) or by the kind of result produced (spatial consequences vs functional consequences), and even by the set of equivalent verbs which provide a differential property of each ATs (*to shut* vs. *to lock* vs. *to close up* vs. *to clench*).

Verbs like *to close* shows that TS is not a necessary differential of ATs, but, as the next examples will point out, it can help to select among the interpretations of a general verb. This is the case with verbs like *to press* and *to push* which record different TSs possibilities across their variation. Let's consider the verb *to press*. In the IMA-GACT ontology it shows 10 different ATs. We can observe groups of Types that share the same TS. Types *a* (table 2) and *b* (table 3) present Agent-Verb-Theme-Destination structure. In both cases, the destination is necessary to represent the event type, which cannot be identified otherwise. In type *a*, the Agent compacts the Theme into a block, changing its shape but not its volume. In type *b* the Agent squeezes the Theme, reducing its volume.

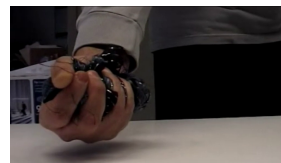


John presses the scraps into a block
AG-V-TH-DEST
to push, to compact

Table 2: *To press*, type *a*

Types *c*, *d* and *e* (tab.4, 5, 6) differ from types *a* and *b* since Destination is not necessary and AG-V-TH is sufficient to identify the action.

Despite the common Thematic Structure, they clearly identify different actions for cognitive reasons. In type *c* the Theme is a humans body part,



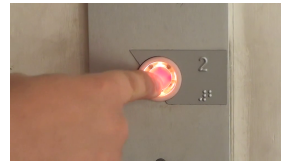
Mary presses the fabric into a ball
AG-V-TH-DEST
to push, to squeeze,
to compress

Table 3: *To press*, type *b*



The doctor presses the shoulder
AG-V-TH
to push, to poke

Table 4: *To press*, type *c*



John presses the button
AG-V-TH
to push

Table 5: *To press*, type *d*



John presses the pedal
AG-V-TH
to push

Table 6: *To press*, type *e*

and the concept implies a compression as an inter-subjective activity, whereas in type *d* the Theme is an object and the compression implies a functional correlation. In type *e* the Theme undergoes a continuous scalar pressure, not limited to a single impulse.

Although these TS commonalities among types show that TS is not necessarily predictive of a single type, TS helps in the distinction of action types. For example, TS restricts the range of possible interpretation of a general verb like *to press* in the case of type *a* and type *b* (table 2 and 3) versus type *c*, *d* and *e* (tables 4, 5 and 6). The distinction between these groups of types (which is independently motivated) is mirrored by the different TSs.

TS may constitute an important feature for the definition of granularity of action types in the verb variation. Type *c* (*the doctor presses the shoulder*, tab.4) and type *f* (*the thief presses the gun into Marys back*, tab.7) are distinguished in reason of their TS: they are similar actions from a cognitive point of view and they can be paraphrased both with *to push*, but the TS of the verb in the

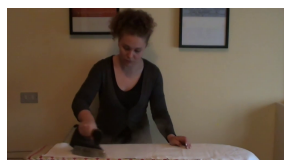
two events is different.



*The thief presses the gun
into Mary's back*
AG-V-TH-DEST
to push

Table 7: *To press*, type *f*

Two cases in the variation of *to press* are characterized by a specific TS: type *g* (AG-V-TH-INSTR) and type *h* (AG-V-TH-RESULT). Type *g* (tab. 8) necessarily requires the instrument in its minimal structure, contrary to all other types; type *h* (tab. 9) requires the expression of the result of the action. TS is predictive of the Action Type in those cases.



*The tailor presses the
cloth with the iron*
AG-V-TH-INSTR
to push

Table 8: *To press*, type *g*



John presses the can flat
AG-V-TH-RESULT
to push

Table 9: *To press*, type *h*

Considering the variation of a verb like *to press*⁸, we can conclude that TS is not peculiar of the verb but is related to its types. Given the cases in which one TS is shared by multiple types, it is clear that types distinction is not a function of the thematic variation (which is actually related to the intersection of multiple features). However, TS has a role in type prediction, since it helps identifying the features of a type.

4 The role of Argument Alternation

Argument Alternations (AAs) are one of those properties of the verb that have received great attention in a large body of literature after Levin (1993). As we will see, also AAs are not properties of the whole verb, but rather characterizes the verb in its types. Considering the verb *to press*, it

⁸Further similar examples have been extracted from the IMAGACT ontology; however, for space limitations, we refer only to the cases already discussed.

has been previously stated that it shows the *conative alternation*, i.e. “a transitivity alternation in which the objects of the verb in the transitive variant turns up in the intransitive conative variant as the object of the preposition in a prepositional phrase headed by the preposition *at* (sometimes *on* with certain verb of ingesting and the push/pull verbs)” (Levin, 1993, p.42). However, only four types of *press* allow for the *conative alternation*, as illustrated in the examples below:

- Type *c*: The doctor presses the shoulder → The doctor presses on the shoulder
- Type *d*: John presses the button → John presses on the button
- Type *e*: John presses the pedal → John presses on the pedal

Other types result in agrammatical sentences when the *conative alternation* is applied:

- Type *a*: *John presses at/on the scraps into a block
- Type *g*: *The tailor presses at/on the cloth with the iron

Considering now *to push*, a verb that shares many interpretations with *to press*, only some types of *to push* (types *a*, *b*, *c*, *d* but not *e*, *f* and *g*) allow this alternation:

- Type *a*: John pushes the button → John pushes on the button
- Type *b*: John pushes the shoulder → John pushes on the shoulder
- Type *c*: John pushes the lever → John pushes on the lever
- Type *d*: John pushes the pedal → John pushes on the pedal
- Type *e*: Mary pushes the chair → *Mary pushes on the chair
- Type *f*: Mary pushes the toothpaste → *Mary pushes on the toothpaste
- Type *g*: Mary pushes the fabric → *Mary pushes on the fabric

In addition to the conative alternations, other two alternations can be seen in the variation of the verbs considered: the *resultative construction* and the *theme-instrument alternation*. The *resultative construction* presents a phrase that describes the state achieved by the referent of a noun phrase as a result of the action. As noted already by Levin (1993, p. 100), it cannot be predicated in case of oblique:

- (1) a. The silversmith pounded the metal flat.
- b. *The silversmith pounded on the metal flat.

This alternation is found only in type *h* for *to press*:

- John presses the can → John presses the can flat

Lastly, we find an alternation between the Theme and the Instrument, not listed in Levin (1993). In this case, the Instrument from sentence 2b (which can be seen as the result of a conative alternation) becomes the Theme in sentence 2c.

- (2) a. The doctor pushes the shoulder with his hand
- b. The doctor pushes on the shoulder with his hand
- c. The doctor pushes his hand on the shoulder

This alternation can be considered as a particular case of *locative alternation*. In terms of Levin (1993), the noun *shoulder* would represent the location argument, whereas *hand* would be considered the *locatum*. Also in this case, the *theme-instrument alternation* does not apply to all types of the variation of *to press*, but rather characterizes specific types.

- Type *g*: the tailor presses the clothes with the iron → The tailor presses the iron on the clothes
- Type *c*: the doctor presses the shoulder → the doctor presses the shoulder with the hand → the doctor presses the hand on the shoulder
- Type *d*: John presses the button → John presses the button with the hand → John presses the hand on the button

- Type *f*: the thief presses the gun into Marys back → the thief presses Marys back with the gun⁹

For the verb *to push*, only types *i* and *d* allow it:

- Type *i*: The thief pushes the gun into Marys back → The thief pushes Marys back with the gun
- Type *d*: John pushes the pedal → John pushes the foot on the pedal

As a whole, if considered together, TS and AA can reduce the underdetermination of types. In other words, when two types share the same TS, they can be predicted from a difference in their argument alternations. This is the case, for example, for types *a* (table 2) and *f* (table 7) of *to press*, which share the TS AG-V-TH-DEST, but differ with respect to the theme-instrument alternation: only type *f* allows it, not type *a*.

In the variation of *to push*, types *e* and *a* share the same TS (AG-V-TH) but type *e* does not allow the conative alternation (**Mary pushes on the chair*), contrary to types *a* (*John pushes on the button*).

5 Conclusion

In this paper we have investigated the relation between Thematic Structure and Action Types. The empirical analysis reveals that Thematic Structure and Argument Alternations are not properties of the whole verb, but rather of the verb in its Types. We have provided evidence about the saliency of both Thematic Structure and Argument Alternations in the identification of Action Types. Although TS and AA do not determine the variation of a verb across different ATs, these linguistic features can, indeed, reveal characterizing features of a Type, helping us in the disambiguation of concepts and in the recognition of the necessary level of granularity in building our ontologies.

References

Hans Christian Boas. 2002. Bilingual framenet dictionaries for machine translation. In *Proceedings of LREC*.

⁹Other types do not allow the theme-instrument alternation: *John presses the hand on the scraps (type *a*); *Mary presses the hand on the clothes (type *i*). For completeness, we report some examples of *to push* for which this alternation does not hold: *Mary pushes on the chair with her hand (type *e*); *Mary pushes the hand on the box (type *h*).

- Susan Windisch Brown, Gloria Gagliardi, and Massimo Moneglia. 2014. Imagact4all. mapping spanish varieties onto a corpus-based ontology of action. *CHIMERA: Journal of Romance Corpora and Linguistic Studies*, (1):91–135.
- William Croft. 2012. *Verbs: Aspect and causal structure*. OUP Oxford.
- Hoa Trang Dang and Martha Palmer. 2005. The role of semantic roles in disambiguating verb senses. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 42–49. Association for Computational Linguistics.
- David Dowty. 1979. *Word Meaning and Montague Grammar*. Reidel Publishing Co, Dordrecht.
- Charles J Fillmore, Josef Ruppenhofer, and Collin F Baker. 2004. Framenet and representing the link between semantic and syntactic relations. *Frontiers in linguistics*, 1:19–59.
- Charles J. Fillmore. 1967. The case for case. In E. Bach and R. Harms, editors, *Universals in Linguistic Theory*, pages 1–89. Holt, Rinehart and Winston, New York.
- Gloria Gagliardi. 2014. *Validazione dell'ontologia dell'azione IMAGACT per lo studio e la diagnosi del Mild Cognitive Impairment*. Ph.D. thesis, University of Florence.
- Peter Gärdenfors. 2014. *The geometry of meaning: Semantics based on conceptual spaces*. MIT Press, Cambridge (MA).
- Daniel Gildea and Daniel Jurafsky. 2002. Automatic labeling of semantic roles. *Computational linguistics*, 28(3):245–288.
- Jeffrey Gruber. 1965. *Studies in Lexical Relations*. Ph.D. thesis, M.I.T.
- Karin Kipper-Schuler. 2005. *VerbNet: A broad-coverage, comprehensive verb lexicon*. Ph.D. thesis, University of Pennsylvania.
- Beth Levin. 1993. *English verb classes and alternations: A preliminary investigation*. University of Chicago press.
- Gabor Melli, Yang Wang, Yudong Liu, Mehdi M Kashani, Zhongmin Shi, Baohua Gu, Anoop Sarkar, and Fred Popowich. 2006. Description of squash, the sfu question answering summary handler for the duc-2005 summarization task. *Proceedings of the HLT/EMNLP Document Understanding Workshop (DUC)*.
- Massimo Moneglia, Gloria Gagliardi, Alessandro Panunzi, Francesca Frontini, Irene Russo, and Monica Monachini. 2012a. Imagact: Deriving an action ontology from spoken corpora. In *Proceedings of the Eight Joint ACL - ISO Workshop on Interoperable Semantic Annotation (ISA-8)*. Pisa, October 3-5, 2012, pages 42–47.
- Massimo Moneglia, Monica Monachini, Omar Calabrese, Alessandro Panunzi, Francesca Frontini, Gloria Gagliardi, and Irene Russo. 2012b. The imagact cross-linguistic ontology of action. a new infrastructure for natural language disambiguation. In Nicoletta Calzolari, editor, *Proceedings of the Eight International Conference on Language Resources and Evaluation*, pages 948–955. European Language Resources Association (ELRA).
- Massimo Moneglia, Alessandro Panunzi, and Lorenzo Gregori. 2018a. Action identification and local equivalence of action verbs: the annotation framework of the imagact ontology. In James Pustejovsky and Ielka van der Sluis, editors, *Proceedings of the LREC 2018 Workshop AREA Annotation, Recognition and Evaluation of Actions*, pages 23–30. European Language Resources Association (ELRA).
- Massimo Moneglia, Alessandro Panunzi, and Lorenzo Gregori. 2018b. Taking events in hindi. a case study from the annotation of indian languages in imagact. In *Proceedings of the LREC 2018 Workshop WILDRE4 4th Workshop on Indian Language Data: Resources and Evaluation*, pages 46–51. LREC.
- Massimo Moneglia. 2014. Natural language ontology of action: A gap with huge consequences for natural language understanding and machine translation. In Z. Vetulani and J. Mariani, editors, *Human Language Technology. Challenges for Computer Science and Linguistics.*, pages 370–395. Springer, Berlin/Heidelberg.
- Martha Palmer, Daniel Gildea, and Paul Kingsbury. 2005. The proposition bank: An annotated corpus of semantic roles. *Computational linguistics*, 31(1):71–106.
- Yi Pan, Massimo Moneglia, Alessandro Panunzi, and Lorenzo Gregori. 2018. Imagact4all. una ontologia per immagini dell'azione per l'apprendimento del lessico verbale di base delle lingue seconde. In Anna De Meo and Margaret Rasulo, editors, *Usare le lingue seconde*, pages 120–150. Officinaventuno.
- Alessandro Panunzi, Irene De Felice, Lorenzo Gregori, Stefano Jacoviello, Monica Monachini, Massimo Moneglia, and Valeria Quochi. 2014. Translating action verbs using a dictionary of images: the imagact ontology. In *Proceedings of the XVI EU-RALEX International Congress: The User in Focus. Bolzano: EURAC research*, pages 1163–1170.
- James Pustejovsky. 1991. The syntax of event structure. *Cognition*, 41:47–81.
- Malka Rappaport Hovav and Beth Levin. 2012. Building verb meanings. In Miriam Butt and Wilhelm Geuder, editors, *The projection of arguments: Lexical and compositional factors*, pages 97–134. CSLI Publications, Stanford, CA.
- Will Roberts and Valia Kordoni. 2012. Using verb subcategorization for word sense disambiguation. In *LREC*, pages 829–832.

- Eleanor Rosch. 1983. Prototype classification and logical classification: The two systems. *New trends in conceptual representation: Challenges to Piaget's theory*, pages 73–86.
- Mihai Surdeanu, Sanda Harabagiu, John Williams, and Paul Aarseth. 2003. Using predicate-argument structures for information extraction. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*.
- Leonard Talmy. 1988. Force dynamics in language and cognition. *Cognitive science*, 12(1):49–100.
- Zeno Vendler. 1957. Verbs and times. *The philosophical review*, 56:97–121.

HATECHECKER: a Tool to Automatically Detect *Hater* Users in Online Social Networks

Cataldo Musto

University of Bari

Dip. di Informatica

cataldo.musto@uniba.it

Angelo Pio Sansonetti

University of Bari

Dip. di Informatica (*Bachelor Student*)

a.sansonetti6@studenti.uniba.it

Marco Polignano

University of Bari

Dip. di Informatica

marco.polignano@uniba.it

Giovanni Semeraro

University of Bari

Dip. di Informatica

giovanni.semeraro@uniba.it

Marco Stranisci

Associazione ACMOS

Torino

marco.stranisci@acmos.net

Abstract

In this paper we present HATECHECKER, a tool for the automatic detection of *hater* users in online social networks which has been developed within the activities of "Contro L'Odio" research project.

In a nutshell, our tool implements a methodology based on three steps: (i) all the Tweets posted by a target user are gathered and processed. (ii) sentiment analysis techniques are exploited to automatically label intolerant Tweets as *hate speeches*. (iii) a lexicon is used to classify hate speeches against a set of specific categories that can describe the target user (e.g., racist, homophobic, anti-semitic, etc.).

Finally, the output of the tool, that is to say, a set of labels describing (if any) the intolerant traits of the target user, are shown through an interactive user interface and exposed through a REST web service for the integration in third-party applications.

In the experimental evaluation we crawled and annotated a set of 200 Twitter profiles and we investigated to what extent our tool is able to correctly identify *hater* users. The results confirmed the validity of our methodology and paved the way for several future research directions.

1 Background and Motivations

According to a recent study¹, 58% of the Italian population regularly uses online social networks as Twitter, Facebook, Instagram and LinkedIn.

Such a huge diffusion of these platforms is providing the users with many new opportunities and services, just think that almost everyone now uses social media to get information, discuss, express opinions and stay in touch with friends. Unfortunately, due to the lack of control and the absence of a clear management of the concept of *identity* of the users, social networks have become the *perfect place* to spread hate against minorities and people having different cultures, values and opinions.

As pointed out by several works (Mathew et al., 2018), the diffusion of *hate speeches* in online social media is continuously growing and the countermeasures adopted by the single platforms are neither effective nor timely, even if a big effort is done to make the process of removing hate speeches faster and more precise². Accordingly, the research line related to the development of tools and methods for the *automatic detection of hate speeches* gained more and more attention. Techniques for detecting hate speeches are obviously based on NLP techniques, and range from simple lexicon-based approaches (Gitari et al., 2015) to more sophisticated techniques that exploit word embeddings (Djuric et al., 2015) and deep learning methods (Badjatiya et al., 2017).

Similar research attempts were also proposed for the Italian language. One of the most popular initiative is the Italian HateMap project (Musto

¹<https://wearesocial.com/it/blog/2018/01/global-digital-report-2018>

²<https://www.cnn.com/2019/02/04/facebook-google-and-twitter-are-getting-faster-at-removing-hate-speech-online-eu-finds-.html>

et al., 2016), a research project that exploits semantic analysis and opinion mining to identify the most-at-risk areas of the Italian country, that is to say, the areas where the users more frequently publish hate speeches. The interest of the research community for the topic was confirmed by the recent work by Bosco et al. (Bosco et al., 2017), who studied hate speech against immigrants, and by Anzovino et al. (Anzovino et al., 2018) who detected misogyny on Twitter. Moreover, as shown by the organization of a specific task in the EVALITA evaluation campaign, an important effort is now devoted to the automatic detection of misogyny (Fersini et al., 2018) and hate speeches in general (Bosco et al., 2018; Basile et al., 2019).

In order to continue the investigation in this research line ACMOS³, a no-profit association based in Torino, recently launched "Contro l'Odio"⁴, a joint research project with the University of Bari, University of Torino and several local associations. The project aims to develop tools and methodologies to monitor (and *hopefully* tackle) online hate speeches and intolerant behaviors.

One of the outcomes of the research is HATE-CHECKER, a tool that aims to automatically identify *hater* users on Twitter by exploiting sentiment analysis and natural language processing techniques. The distinguishing aspect of the tool with respect to the work we have previously introduced is the *focus* of the tool itself. Indeed, differently from most of the literature, that focused on the analysis of single Tweets, HATECHECKER aims to analyze the users *as a whole*, and to identify *hater users* rather than *hate speeches*. Clearly, both the tasks are in close correlation, since techniques to detect hate speeches can be used to detect *hater* users as well.

However, through this work we want to move the focus on the latter since, up to our knowledge, this a poorly investigated research direction. Just think that no datasets of *hater users* is currently publicly available.

To sum up, the contributions of the work can be summarized as follows:

- We present a workflow that allows to detect *hater users* in online social networks;

- We evaluate several configurations (on varying of lexicons and sentiment analysis algorithms) of the pipeline and we identified the most effective one to tackle our specific task;
- We share the first publicly available dataset for automatic detection of *hater users* on Twitter.

In the following, we will first describe the methodology we designed to implement our system, then we will discuss the effectiveness of the approach by analyzing the results we obtained on a (publicly available) dataset of 200 Twitter users.

2 Methodology

The workflow carried out by the HATECHECKER tool is reported in Figure 1.

Generally speaking, the pipeline consists of four different modules, that is to say, a SOCIAL DATA EXTRACTOR, a SENTIMENT ANALYZER, a PROFILE CLASSIFIER and a SOCIAL NETWORK PROCESSOR. All these components use a `NOSQL` database to store the information they hold and expose the output returned by the tool through a `REST` interface as well as through a Web Application. In the following, a description of the single modules that compose the workflow is provided.

2.1 Social Data Extractor

The whole pipeline implemented in the HATECHECKER tool needs some *textual content* posted by the target user to label the user as a *hater* or not. In absence of textual content, it is not possible provide such a classification. To this end, the first and mandatory step carried out by the tool is the extraction of the Tweets posted by the user we want to analyze. In this case, we used the official Twitter APIs to gather the available Tweets and to forward it to the next modules of the workflow.

Given that the *real-time execution* of the workflow is one of the constraints of the project, we limited the extraction to the 200 most recent Tweets posted by the user. This is a reasonable choice, since we aim to detect users who *recently* showed an intolerant behavior, rather than users who posted hate speeches one or two years ago.

2.2 Sentiment Analyzer

Once the Tweets have been collected, it is necessary to provide the tool with the ability to go

³<http://www.acmos.net>

⁴<http://www.controlodio.it>

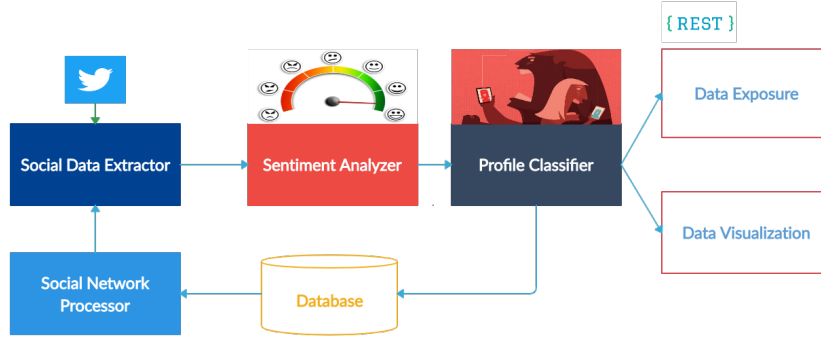


Figure 1: The workflow carried out by the HATECHECKER tool

through the content posted by the target and to automatically identify the *hate speeches*.

To this end, the SENTIMENT ANALYZER module exploits Sentiment Analysis techniques (Pang et al., 2008) to basically classify each Tweet as *positive* or *negative* (that is to say, conveying hate speeches or not). To get this output we integrated and compared two different implementations of sentiment analysis algorithms:

- **SentiPolC**: (Basile and Novielli, 2014) a sentiment analysis algorithm that resulted as the best-performing one in EVALITA 2014 in carrying out the task of associating the correct *sentiment* to Tweets;
- **HanSEL**: an algorithm based on a deep neural network C-BiLSTN (Zhou et al., 2015) with an input layer of word embeddings. This strategy is based on the work proposed by Polignano et al. (Polignano and Basile, 2018) and it has been improved within the activities of the 'Contro l'Odio' research project. In particular, the whole net has been trained for 20 epochs with early stopping criteria, Adam loss function, and binary cross-entropy as optimization function.

A complete overview of the algorithms is out of the scope of this paper and we suggest to go through the references for a thorough discussion. For the sake of simplicity, we can state that the output of both the algorithms is a *binary* classification of each Tweet posted by the target user as *negative* (that is to say, conveying hate speeches) or *positive*. Such an output is then passed to the PROFILE CLASSIFIER module whose goal is to assign a more precise label to the user, on the ground of the nature of the hate speeches she posted (if any).

2.3 Profile Classifier

In such a specific setting, the simple exploitation of sentiment analysis techniques that provide a *rough* binary classification of the single Tweets (*conveying/not conveying hate*) is not enough. Indeed, the answers to two fundamental questions are still lacking:

- How can we label the user *as hater or non-hater* on the ground of the Tweets she posted?
- How can we return a more fine-grained classification of the user (*e.g.*, racist, homofobe, etc.) on the ground of the Tweets she posted?

Both these issues are tackled by the PROFILE CLASSIFIER module. As for the first question, a very simple strategy based on *thresholding* is implemented. In particular, we defined a parameter ϵ , and whether the user posted a number of Tweets labeled as *hate speeches* higher than ϵ , the user herself is labeled as an *hater*. Of course, several values for the parameter ϵ can be taken into account to run the tool.

As for the second question, we used a *lexicon-based* approach to provide a fine-grained classification of users' profiles. The intuition behind our methodology is that for each category a specific lexicon can be defined, and whether a Tweet posted by the user contains one of the terms in the lexicon, the user is labeled with the name of the category.

Formally, let $C = \{c_1, c_2 \dots c_n\}$ be the set of the categories (*e.g.*, racism, homophobia, sexism, etc.) and let $V_{C_i} = \{t_1, t_2 \dots t_m\}$ be the vocabulary of the category C_i . Given a Tweet T written by a user u , if one of the terms in V_{C_i} is contained in T , the user u is labeled with the category C_i .

To define the lexicon for each category, we relied on the research results of the Italian Hate Map

(Lingiardi et al., 2019). In particular, we exploited the categories as well as the lexicon used in the Italian Hate Map Project, which consists of 6 different categories (*racism, homophobia, islamophobia, xenophobia, anti-semitism, sexism, abuse against people with disabilities*) and 76 different terms in total.

In order to (hopefully) enrich and improve the lexicon used in the Italian Hate Map project, we exploited Hurltlex, a multilingual lexicon of hate words (Bassignana et al., 2018). Specifically, we manually selected a subset of relevant terms among those contained in Hurltlex and we merged the new terms with those contained in the original lexicon. In total, the complete lexicon contained 100 terms, 76 coming from the original Italian Hate Map lexicon and 24 gathered from Hurltlex.

Obviously, in the experimental session the effectiveness of the tool on varying of different lexicons and on different configuration of the workflow will be evaluated.

2.4 Social Network Processor

At the end of the previous step, the target user is labeled with a set of categories describing the *facets* of her intolerant behavior.

However, one of the goals of the project was also to investigate the role and the impact of the social network of the users in the dynamics of online *haters*. Accordingly, the SOCIAL NETWORK PROCESSOR gathers the entire social network of the target user and runs again (in background, of course) the whole pipeline on all the *following* and *followers* of the target user, in order to detect whether other people in the social network of the target user can be labeled as *haters* as well. The goal of this step is to further enhance the comprehension of network dynamics and to understand whether online *haters* tend to follow and be followed by *other haters*.

Unfortunately, due to space reasons, the discussion of this part of the workflow is out of the scope of this paper and is left for future discussions.

2.5 Data Exposure and Data Visualization

Finally, the output of the platform is made available to third-party services and to the user itself. In the first case, a REST web service makes available the output of the tool (that is so say, the hate categories and the number of haters in her own social network), while in the latter the same data are shown through an interactive user interface.

A screenshot of the working prototype of the platform is reported in Figure 2. As shown in the Figure, a user interacting with the platform can query the system by interactively providing a Twitter user name. In a few seconds, the interface shows a report of the target user containing a set of emojis reporting the behavior of the user for each of the categories we analyzed, a snapshot of her own Tweets labeled as hate speeches and some information about the percentage of hater profiles that are in the social network of the target user.

It is worth to note that such a web application is very useful for both monitoring tasks (e.g., to verify whether a third-party account is an online hater) as well as for *Quantified Self* scenarios (Swan, 2013), that is to say, to improve the self-awareness and the self-consciousness of the user towards the dynamics of her social network. Our intuition is that a user who is aware of not being an hater, can use the system to identify (if any) the haters that are still in her own social network, and maybe decide to unfollow them.

3 Experimental Evaluation

The goal of the experimental session was to evaluate the effectiveness of the tool on varying of different configurations of the pipeline.

To this end, due to the lack of a dataset of *hater profiles*, we manually crawled and annotated a set of 200 Twitter users, which we made available⁵ for the sake of reproducibility and to foster the research in the area.

In particular, we compared four different strategies to run our tool, on varying on two different parameters, such as the lexicon and the sentiment analysis algorithm. In particular, we exploited the following combinations of parameters:

- **Sentiment Analysis:** SentipolC and HanSEL, as previously explained
- **Lexicons:** HateMap lexicon and complete lexicon (HateMap+Hurltlex)

As for the parameters, the threshold ϵ was set equal to 3 and both the sentiment analysis algorithms were run with the standard parameters introduced in the original papers. To evaluate the effectiveness of the approaches, we calculate the number of correctly classified user profiles over the total of hater users in the dataset.

⁵<https://tinyurl.com/uniba-haters-dataset>

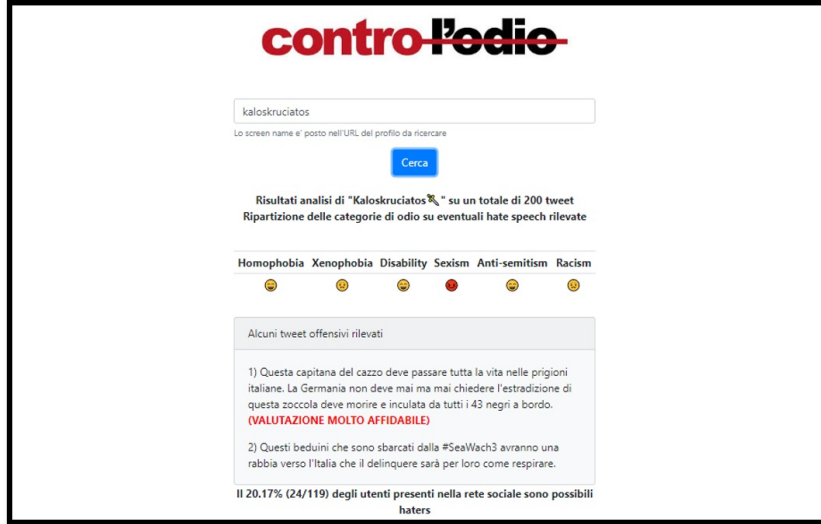


Figure 2: A screenshot of HATECHECKER at work

		Facets					
Lexicon	Algorithm	Racism	Anti-semitism	Disability	Sexism	Homophobia	Xenophobia
HateMap	SentiPolC	71.5	92.0	82.0	77.5	84.0	75.5
HateMap	HanSEL	73.0	95.5	88.5	79.0	84.0	79.0
Complete	SentiPolC	78.0	95.0	86.5	78.0	84.0	78.0
Complete	HanSEL	75.0	97.0	88.5	78.0	84.0	79.0

Table 1: Results of the Experiment. The best-performing configuration for each facet is reported in **bold**.

The results of the experiments are reported in Table 1. In general, we can state that our approach to automatically detect hater users in online social network provided us with encouraging results, since more a percentage between 78% and 97% of the online haters were correctly detected by the algorithm, regardless of the specific category.

It is worth to note that the *worse* results (both of them are beyond 70%, through) were obtained for *racism* and *xenophobia*, that is to say, two facets characterized by a lexicon that quickly evolves and often adopts terms that are *not conventional* and not necessarily conveying *hate* (e.g., expressions as '*Aiutiamoli a casa loro*' or terms as '*clandestini*'). However, even for these categories the results we obtained were encouraging.

Conversely, results were particularly outstanding for facets such as *anti-semitism* and *homophobia*, that have a quite fixed *lexicon* of terms that can be used to hurt or offend such minorities.

As for the different configurations, we noted that HANSEL tended to obtain better results than SENTIPOLC. This is a *quite* expected outcome, since it exploits more novel and effective meth-

ods as those based on word embeddings and deep learning techniques. Moreover, we can state that the results can be further improved since no particular tuning of the parameters was carried out in this work.

As for the lexicons, the extension of the original Italian Hate Map lexicons with new terms led to an improvement of the results for all the facets (except for *homophobia*) for at least one of the comparisons. Such improvement are often tiny, but this is an expected outcome since just a few terms coming from Hurtlex were added. However, even these preliminary results provided us with encouraging findings, since they showed that the integration and the extension of sensible terms with the information coming from recently developed lexical resources can lead to a further improvement of the accuracy of the system.

4 Conclusions and Future Work

In this work we have presented HATECHECKER, a tool that exploits sentiment analysis and natural language processing techniques to automatically detect *hater users* in online social networks.

Given a target user, the workflow we implemented in our system uses sentiment analysis techniques to identify hate speeches posted by the user and exploits a lexicon that extends that of the Italian Hate Map project to assign to the person one or more labels that describe the nature of the hate speeches she posted.

As future work, we plan to arrange a user study, specifically designed for *young people*, to evaluate the effectiveness of the system as a Quantified Self tool (Musto et al., 2018), that is to say, to improve the awareness of the users towards the behavior of other people in their social network.

References

- Maria Anzovino, Elisabetta Fersini, and Paolo Rosso. 2018. Automatic identification and classification of misogynistic language on Twitter. In *International Conference on Applications of Natural Language to Information Systems*, pages 57–64. Springer.
- Pinkesh Badjatiya, Shashank Gupta, Manish Gupta, and Vasudeva Varma. 2017. Deep learning for hate speech detection in tweets. In *Proceedings of the 26th International Conference on World Wide Web Companion*, pages 759–760. International World Wide Web Conferences Steering Committee.
- Pierpaolo Basile and Nicole Novielli. 2014. Uniba at evalita 2014-sentipolc task: Predicting tweet sentiment polarity combining micro-blogging, lexicon and semantic features. *Proceedings of EVALITA*, pages 58–63.
- Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. 2019. Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 54–63.
- Elisa Bassignana, Valerio Basile, and Viviana Patti. 2018. Hurllex: A multilingual lexicon of words to hurt. In *5th Italian Conference on Computational Linguistics, CLiC-it 2018*, volume 2253, pages 1–6. CEUR-WS.
- Cristina Bosco, Viviana Patti, Marcello Bogetti, Michelangelo Conoscenti, Giancarlo Francesco Ruffo, Rossano Schifanella, and Marco Stranisci. 2017. Tools and resources for detecting hate and prejudice against immigrants in social media. In *Symposium III. Social Interactions in Complex Intelligent Systems (SICIS) at AISB 2017*, pages 79–84. AISB.
- Cristina Bosco, Dell’Orletta Felice, Fabio Poletto, Manuela Sanguinetti, and Tesconi Maurizio. 2018. Overview of the evalita 2018 hate speech detection task. In *EVALITA 2018-Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian*, volume 2263, pages 1–9. CEUR.
- Nemanja Djuric, Jing Zhou, Robin Morris, Mihajlo Grbovic, Vladan Radosavljevic, and Narayan Bhamidipati. 2015. Hate speech detection with comment embeddings. In *Proceedings of the 24th international conference on world wide web*, pages 29–30. ACM.
- Elisabetta Fersini, Debora Nozza, and Paolo Rosso. 2018. Overview of the evalita 2018 task on automatic misogyny identification (ami). In *EVALITA@CLiC-it*.
- Njagi Dennis Gitari, Zhang Zuping, Hanyurwimfura Damien, and Jun Long. 2015. A lexicon-based approach for hate speech detection. *International Journal of Multimedia and Ubiquitous Engineering*, 10(4):215–230.
- Vittorio Lingiardi, Nicola Carone, Giovanni Semeraro, Cataldo Musto, Marilisa D’Amico, and Silvia Brena. 2019. Mapping twitter hate speech towards social and sexual minorities: a lexicon-based approach to semantic content analysis. *Behaviour & Information Technology*, 0(0):1–11.
- Binny Mathew, Ritam Dutt, Pawan Goyal, and Animesh Mukherjee. 2018. Spread of hate speech in online social media. *arXiv preprint arXiv:1812.01693*.
- Cataldo Musto, Giovanni Semeraro, Marco de Gemmis, and Pasquale Lops. 2016. Modeling community behavior through semantic analysis of social data: The italian hate map experience. In *Proceedings of the 2016 Conference on User Modeling Adaptation and Personalization*, pages 307–308. ACM.
- Cataldo Musto, Giovanni Semeraro, Cosimo Lovascio, Marco de Gemmis, and Pasquale Lops. 2018. A framework for holistic user modeling merging heterogeneous digital footprints. In *Adjunct Publication of the 26th Conference on User Modeling, Adaptation and Personalization*, pages 97–101. ACM.
- Bo Pang, Lillian Lee, et al. 2008. Opinion mining and sentiment analysis. *Foundations and Trends® in Information Retrieval*, 2(1–2):1–135.
- Marco Polignano and Pierpaolo Basile. 2018. Hansel: Italian hate speech detection through ensemble learning and deep neural networks. In *EVALITA@CLiC-it*.
- Melanie Swan. 2013. The quantified self: Fundamental disruption in big data science and biological discovery. *Big data*, 1(2):85–99.
- Chunting Zhou, Chonglin Sun, Zhiyuan Liu, and Francis Lau. 2015. A c-lstm neural network for text classification. *arXiv preprint arXiv:1511.08630*.

The Contribution of Embeddings to Sentiment Analysis on YouTube

Moniek Nieuwenhuis

CLCG, University of Groningen
The Netherlands

m.l.nieuwenhuis@student.rug.nl

Malvina Nissim

CLCG, University of Groningen
The Netherlands

m.nissim@rug.nl

Abstract

We train a variety of embeddings on a large corpus of YouTube comments, and test them on three different tasks on both the English and the Italian portions of the SenTube corpus. We show that in-domain (YouTube) embeddings perform better than previously used generic embeddings, achieving state-of-the-art performance on most of the tasks. We also show that a simple method for creating sentiment-aware embeddings outperforms previous strategies, and that sentiment embeddings are more informative than plain embeddings for the SenTube tasks.

1 Introduction and Background

Sentiment analysis, or opinion mining, on social media is by now a well established task, though surely not solved (Liu et al., 2005; Barnes et al., 2017). Part of the difficulty comes from its intrinsic subjective nature, which makes creating reliable resources hard (Kiritchenko and Mohammad, 2017). Another part comes from its heavy interaction with pragmatic phenomena such as irony and world knowledge (Nissim and Patti, 2017; Basile et al., 2018; Cignarella et al., 2018; Van Hee et al., 2018). And another difficulty comes from the fact that given a piece of text, be it a tweet, or a review, it isn't always clear what exactly the expressed sentiment (should there be any) is about. In commercial reviews, for example, the target of a user's evaluation could be a specific aspect or part of a given product. Aspect-based sentiment analysis has developed as a subfield to address this problem (Thet et al., 2010; Pontiki et al., 2014).

The SenTube corpus (Uryupina et al., 2014) has been created along these lines. It contains English and Italian commercial or review videos about some product, and annotated comments. The annotations specify both the polarity (positive, negative, neutral) and the target (the video itself or the product in the video). In Figure 1 we show two positive comments with different targets.

The SenTube's tasks have been firstly addressed by Severyn et al. (2016) with an SVM based on topic and shallow syntactic information, later outperformed by a convolutional N-gram BiLSTM word embedding model (Nguyen and Le Nguyen, 2018). The corpus has also served as testbed for multiple state-of-the-art sentiment analysis methods (Barnes et al., 2017), with best results obtained using sentiment-specific word embeddings (Tang et al., 2014). On the English sentiment task of SenTube though this method does not outperform corpus-specific approaches (Severyn et al., 2016; Nguyen and Le Nguyen, 2018).

We further explore the potential of (sentiment) embeddings, using the model developed by Nguyen and Le Nguyen (2018). We believe that training in-domain (YouTube) embeddings rather than using generic ones might yield improvements, and that additional gains might come from sentiment-aware embeddings. In this context, we propose a simple new semi-supervised method to train sentiment embeddings and show that it performs better than two other existing ones. We run all experiments on English and Italian data.

Contributions We show that in-domain embeddings outperform generic embeddings on most task of the SenTube corpus for both Italian and English. We also show that sentiment embeddings obtained through a simple semi-supervised strategy that we newly introduce in this paper add a boost to performance. We make all developed Italian and English embeddings avail-

Copyright © 2019 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0)

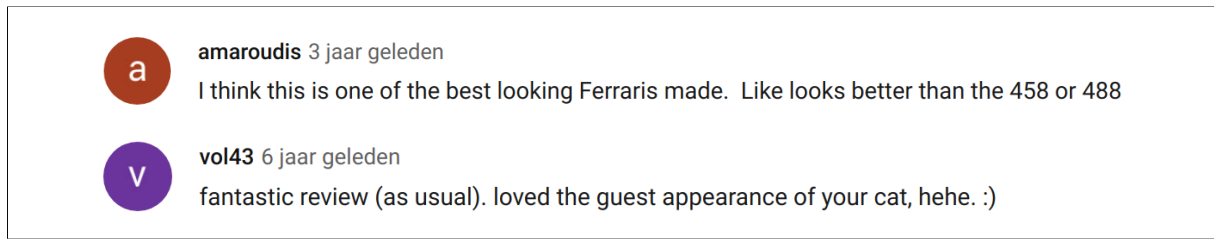


Figure 1: Two sample comments on a video about a Ferrari car. Top: positive comment about the product. Bottom: positive comment about the video.

able at this link: <https://github.com/malvinanissim/youtube-embeds>.

2 Data and Task

We use two different datasets of YouTube comments. The first is the existing SenTube corpus (Uryupina et al., 2014). The other dataset is collected from YouTube to create a big semi-supervised corpus for making the embeddings.

2.1 SenTube corpus

The SenTube corpus contains 217 videos in English and 198 in Italian (Uryupina et al., 2014). All videos are a review or commercial about a product in the category “automobile” or “tablet”.

All comments from the videos are annotated according to their target (whether they are about the video or about the product) and their sentiment polarity (positive, negative, neutral). Some of the comments were discarded because of spam, because they were written in a language other than the intended one (Italian for the Italian corpus, English for the English one), or just off topic. Sentiment is type-specific, and the following labels are used: *positive-product*, *negative-product*, *positive-video* and *negative-video*. If neither positive or negative is annotated, the comment is assumed to be *neutral*.

The corpus lends itself to three different tasks, all of which we tackle in this work:

- the *sentiment task*, namely predicting whether a YouTube comment is written in a positive, negative or a neutral sentiment.
- the *type task*, namely predicting if the comment is written about the product mentioned in the video, about the video itself or if it is not an informative comment (spam or off-topic).
- the *full task*: predicting at the same time the sentiment and the type of each comment.

From SenTube we exclude any comment that is annotated both as product-related and video-related or is both positive and negative. Table 1 shows the label distribution for the three tasks. All comments are further lowercased and tokenised.

2.2 Semi-supervised YouTube corpus

To train in-domain embeddings we collected more data from YouTube. We searched for relevant videos querying the YouTube API with a set of keywords (“car”, “tablet”, “macchina”, “automobile”, ...). For each retrieved video we checked that it was not already included in the SenTube corpus, and verified that its description was in English/Italian using Python’s `langdetect` module. We then retrieved all comments for each video that had more than one comment.

Next, we used the convolutional N-gram BiLSTM word embedding model by (Nguyen and Le Nguyen, 2018), which has state-of-the-art performance on SenTube, to label the data on the sentiment task, as we want to exploit the labels to train sentiment embeddings. Table 2 shows an overview of the collected dataset. A manual check on a randomly chosen test set of 100 comments for each language, revealed a rough accuracy of just under 60% for English, and just under 65% for Italian.

3 Embeddings

We test three different categories of embeddings: some pre-trained models, a variety of models trained on our in-domain dataset, and sentiment-aware embeddings, which we obtain in three different ways. All of the embeddings are tested in the model developed by (Nguyen and Le Nguyen, 2018) to specifically tackle the SenTube tasks.

3.1 Plain Embeddings

Generic models For English we used Google-News vectors¹, which are those used in (Nguyen

¹<https://code.google.com/archive/p/word2vec/>

Table 1: Label distribution for each task in the SenTube corpus

	English				Italian			
	Automobile	%	Tablet	%	Automobile	%	Tablet	%
Product-related	5,834	38.8	11,067	56.2	1,718	40.9	2,976	61.0
Video-related	5,201	34.5	3,665	18.6	1,317	31.4	845	17.3
Uninfo.	4,020	26.7	4,961	25.2	1,161	27.7	1,055	21.6
Positive sentiment	3,284	21.8	3,637	18.5	946	22.5	770	15.8
Negative sentiment	1,988	13.2	3,038	15.4	752	17.9	825	16.9
No sentiment/neutral	9,801	65.0	13,021	66.1	2,499	59.5	3,281	67.3
Product-pos.	1,740	11.5	2,280	11.6	479	11.4	544	11.4
Product-neg.	1,360	9.0	2,473	12.5	538	12.8	711	14.6
Product-neu.	2,744	18.2	6,310	32.0	703	16.8	1,721	35.3
Video-pos.	1,543	10.2	1,357	6.9	467	11.1	226	4.6
Video-neg.	628	4.2	565	2.9	214	5.1	114	2.3
Video-neu.	3,030	20.1	1,743	8.8	635	15.1	505	10.4
Uninfo.	4,028	26.7	4,968	25.2	1,161	27.7	1055	21.6

Table 2: Overview of extra data collected from YouTube

	English			Italian		
	Automobile	Tablet	Total	Automobile	Tablet	Total
Videos	1,592	1,675	3,267	1,622	1,151	2,773
Comments	1,028,136	587,506	1,615,642	99,328	118,274	217,602
Tokens	18,124,184	9,156,324	27,280,508	1,596,190	1,579,591	3,175,781
Unique tokens	754,962	416,835	1,030,574	170,956	155,738	277,114
Positive sentiment	165,725	97,439	263,164 (16.3%)	11,091	13,356	24,447(11.2%)
Negative sentiment	49,490	53,557	103,047 (6.4%)	4,898	4,514	9,412(4.3%)
Neutral sentiment	812,921	436,510	1,249,431 (77.3%)	83,339	100,404	183,743(84.4%)

and Le Nguyen, 2018), and the 200-dimensional GloVe Twitter embeddings². For Italian we used vectors from (Bojanowski et al., 2016) a Fast-Text model trained on the the Italian Wikipedia, and also used by (Nguyen and Le Nguyen, 2018). Furthermore, we tested two models developed at ISTI-CNR, which are trained on Italian Wikipedia with skip-gram’s Word2Vec and with GloVe.³

In-domain trained models We trained three Word2Vec models (Mikolov et al., 2013), all of dimension 300, using Gensim (Řehůřek and Sojka, 2010). Beside a CBOW model with default settings, we trained two different skip-gram models, one with default settings and one with a negative sampling of 10. We also trained a FastText model (Bojanowski et al., 2016), and a 100-dimension GloVe model (Pennington et al., 2014).

3.2 Sentiment-aware Embeddings

We use three methods for adding sentiment to the embeddings, in all cases using the Word2Vec skip-gram models (Mikolov et al., 2013) with and without negative sampling 10. The first two methods are existing methods, namely retrofitting (Faruqui et al., 2015) and the refinement method suggested by Yu et al. (2017), while the third method is newly proposed in this work.

Retrofitting Retrofitting embedding models is a method to refine vector space representations using relational information from semantic lexicons by encouraging linked words to have similar vector representations (Faruqui et al., 2015).⁴ We used two sentiment lexicons to retrofit the skip-gram models. A SentiWordNet-derived lexicon for English (Baccianella et al., 2010), and Sentix for Italian (Basile and Nissim, 2013).⁵

²<https://nlp.stanford.edu/projects/glove/>

³<http://hlt.isti.cnr.it/wordembeddings/>

⁴<https://github.com/mfaruqui/retrofitting>.

⁵<http://valeriobasile.github.io/twita/sentix.html>

Sentiment Embedding refinement We tested the method proposed by Yu et al. (2017) using the provided code⁶ to refine our own skip-gram Word2Vec models. In this method the similar top-k words will be re-ranked by sentiment on the difference in valence scores from a sentiment lexicon. For English we used the E-ANEW sentiment lexicon (Warriner et al., 2013) and for Italian we used Sentix (Basile and Nissim, 2013).

Our Embedding refinement For each language, we use a sentiment lexicon and our YouTube corpus to train sentiment embeddings.

From the sentiment lexicon we create two lists of words: positive words (positive score > 0.6 and negative score < 0.2) and negative words (negative score > 0.6 and positive score < 0.2).

For each word in the positive list, we check if it occurs in a comment with a positive label. We do the same for the negative list and negative labelled comments. If the word occurs in the list we add the affixes "_pos" or "_neg" to the word occurrence in a positive or negative comment. If a word from the positive list is found in a comment with negative or neutral label it isn't touched, and likewise for words in the negative list. An example of this approach is in Table 3.

Example	Label
"I love_pos this review! It's not the technical review that every YouTube vid has bit more of a usable hands on one! makes me really_pos want one even more than before! Thank you!"	positive
"I love being a cheapskate. Please tell me what in the world "gimp" is."	neutral
"I don't understand why people love apple shit [...]"	negative

Table 3: Example of the word “love” changed in the positive comment and not changed in neutral or negative comments.

We then trained the embeddings with skip-gram Word2Vec (Mikolov et al., 2013), with therein the two separate appearances of words, i.e. with and without affixes. This of course poses a problem at test time, since two vectors are now available for some of the words (great_pos and great for “great”, for example, or brutto_neg and brutto for “brutto” [en: ugly]), but one must eventually choose one for representing the encountered word “great”, or “brutto”.

Instead of devising a strategy for choosing one of the two vectors, we opted for *re-joining* the two

versions of the word into a single one, testing two different methods:

- *averaging*: average the vectors with each other; the two contexts have equal weight;
- *weighting*: weigh each vector by the proportion of times the word is in either context (in the semi-supervised corpus), and sum them.

4 Experiments

We split the SenTube corpus in 50% train and 50% test. We could not exactly replicate the split by Nguyen and Le Nguyen (2018) due to lack of sufficient details in their code. We use their model to test all embeddings, including those used in their implementation (GoogleNews for English, and FastText for Italian), for direct comparison with our embeddings. For completeness, we also include the results reported by Severyn et al. (2016) (with their own split), and a most frequent label baseline for each task. As was done in previous work on this corpus, and for more direct comparison, we report accuracy across all experiments.

Table 4: English embeddings results

Task	Embeddings	AUTO	TABLET
Sentiment	Most frequent label baseline	0.632	0.680
	(Severyn et al., 2016)	0.557	0.705
	(Nguyen and Le Nguyen, 2018)	0.669	0.702
	CBOW	0.725	0.755
	SKIP	0.740	0.750
	in-domain SKIP neg samp	0.730	0.756
in-domain	GloVe	0.709	0.754
	FastText	0.729	0.754
generic	GoogleNews	0.715	0.748
	GLoVe Twitter	0.723	0.742
Type	Most frequent label baseline	0.384	0.565
	(Severyn et al., 2016)	0.594	0.786
	(Nguyen and Le Nguyen, 2018)	0.684	0.795
	CBOW	0.714	0.784
	SKIP	0.733	0.800
	in-domain SKIP neg samp	0.723	0.801
in-domain	GloVe	0.697	0.779
	FastText	0.727	0.779
generic	GoogleNews	0.688	0.773
	GLoVe Twitter	0.690	0.775
Full	Most frequent label baseline	0.243	0.342
	(Severyn et al., 2016)	0.415	0.603
	(Nguyen and Le Nguyen, 2018)	0.538	0.613
	CBOW	0.536	0.618
	SKIP	0.547	0.621
	in-domain SKIP neg samp	0.558	0.629
in-domain	GloVe	0.504	0.596
	FastText	0.540	0.615
generic	GoogleNews	0.504	0.580
	GLoVe Twitter	0.487	0.600

⁶ https://github.com/wangjin0818/word_embedding_refine

4.1 Results with plain embeddings

The results using plain embeddings are shown in Tables 4 and 5. Most of the in-domain embeddings on English outperform the GoogleNews vectors used by Nguyen and Le Nguyen (2018); the results are also higher than those reported in previous work with different splits (Severyn et al., 2016; Nguyen and Le Nguyen, 2018). Only for both full tasks and the tablet type task there are a few of the in-domain embeddings which do not outperform on previous work results. For Italian, not all in-domain embeddings outperform previous work in all tasks, but they mostly do when embeddings used in previous work are tested on the same split. For both languages the skip-gram models are performing best compared to all the other in-domain embedding models. On Italian, the generic Wikipedia SKIP embeddings and the generic FastText embeddings (Bojanowski et al., 2016) are performing slightly better on the sentiment and full task for tablets.

Table 5: Italian embedding results

Task	Embeddings	AUTO	TABLET
Sentiment	Most frequent label baseline	0.601	0.668
	(Severyn et al., 2016)	0.616	0.644
	(Nguyen and Le Nguyen, 2018)	0.614	0.656
	CBOW	0.622	0.700
in-domain	SKIP	0.636	0.687
	SKIP neg samp	0.652	0.697
	GloVe	0.607	0.673
	FastText	0.640	0.645
generic	FastText	0.648	0.682
	Wikipedia SKIP	0.629	0.701
	Wikipedia GloVe	0.613	0.679
Type	Most frequent label baseline	0.415	0.568
	(Severyn et al., 2016)	0.707	0.773
	(Nguyen and Le Nguyen, 2018)	0.748	0.796
	CBOW	0.742	0.710
in-domain	SKIP	0.768	0.695
	SKIP neg samp	0.762	0.722
	GloVe	0.744	0.676
	FastText	0.703	0.703
generic	FastText	0.769	0.716
	Wikipedia SKIP	0.756	0.682
	Wikipedia GloVe	0.725	0.694
Full	Most frequent label baseline	0.320	0.252
	(Severyn et al., 2016)	0.456	0.524
	(Nguyen and Le Nguyen, 2018)	0.511	0.550
	CBOW	0.470	0.484
in-domain	SKIP	0.489	0.487
	SKIP neg samp	0.517	0.485
	GloVe	0.450	0.490
	FastText	0.459	0.484
generic	FastText	0.491	0.497
	Wikipedia SKIP	0.492	0.495
	Wikipedia GloVe	0.441	0.449

4.2 Results with sentiment embeddings

Tables 6 and 7 show the results of the sentiment embeddings. In almost all tasks the sentiment embeddings outperform the plain embeddings. Surprisingly, this is true even for the English type task, while the sentiment automobile task has a slightly lower accuracy. For Italian only in the automobile type task sentiment embeddings do not outperform standard ones. Among the sentiment embeddings, our refinement method seems to work best, while retrofitting does not lead to any improvement.

In terms of weighing versus averaging the vectors in our method, for English averaging yields the best score three times, and weighting two times. For Italian, weighting yields the best result two times on the tablet data set, while for the full task averaging is better. For cars, weighting is better, but does not outperform plain embeddings.

Table 6: English sentiment embedding test

Task	Embeddings	AUTO	TABLET
Sentiment	SKIP neg samp retrofitted	0.701	0.751
	SKIP retrofitted	0.710	0.742
	SKIP sentiment embedding refinement	0.725	0.747
	SKIP neg samp sentiment embedding refinement	0.725	0.753
	SKIP sentiment change average	0.715	0.760
	SKIP sentiment change weight sum	0.737	0.767
	SKIP neg samp sentiment change average	0.729	0.758
	SKIP neg samp sentiment change weight sum	0.734	0.749
Type	SKIP neg samp retrofitted	0.688	0.774
	SKIP retrofitted	0.680	0.781
	SKIP sentiment embedding refinement	0.732	0.794
	SKIP neg samp sentiment embedding refinement	0.735	0.796
	SKIP sentiment change average	0.723	0.806
	SKIP sentiment change weight sum	0.716	0.798
	SKIP neg samp sentiment change average	0.722	0.807
	SKIP neg samp sentiment change weight sum	0.739	0.794
Full	SKIP neg samp retrofitted	0.500	0.600
	SKIP retrofitted	0.501	0.594
	SKIP sentiment embedding refinement	0.537	0.594
	SKIP neg samp sentiment embedding refinement	0.522	0.606
	SKIP sentiment change average	0.560	0.616
	SKIP sentiment change weight sum	0.544	0.623
	SKIP neg samp sentiment change average	0.549	0.631
	SKIP neg samp sentiment change weight sum	0.547	0.618

5 Conclusion

We have explored the contribution of in-domain embeddings on the SenTube corpus, on two domains and two languages. In 10 out of the 12 tasks, in-domain embeddings outperform generic ones. This confirms the experiments on the SENTIPOLC 2016 tasks (Barbieri et al., 2016) reported by Petrolito and Dell’Orletta (2018), who recommend the use of in-domain embeddings for sentiment analysis, especially if trained at the word rather than character level. However, a similar work in the field of sentiment analysis for soft-

Table 7: Italian sentiment embedding test

Task	Embeddings	AUTO	TABLET
Sentiment	SKIP neg samp retrofitted	0.649	0.682
	SKIP retrofitted	0.622	0.686
	SKIP sentiment embedding refinement	0.610	0.682
	SKIP neg samp sentiment embedding refinement	0.632	0.703
	SKIP sentiment change average	0.628	0.690
	SKIP sentiment change weight sum	0.623	0.704
	SKIP neg samp sentiment change average	0.640	0.682
	SKIP neg samp sentiment change weight sum	0.631	0.710
	SKIP neg samp retrofitted	0.730	0.712
	SKIP retrofitted	0.744	0.712
Type	SKIP sentiment embedding refinement	0.761	0.716
	SKIP neg samp sentiment embedding refinement	0.754	0.712
	SKIP sentiment change average	0.763	0.701
	SKIP sentiment change weight sum	0.746	0.729
	SKIP neg samp sentiment change average	0.760	0.732
	SKIP neg samp sentiment change weight sum	0.756	0.739
	SKIP neg samp retrofitted	0.478	0.447
	SKIP retrofitted	0.490	0.469
	SKIP sentiment embedding refinement	0.504	0.497
	SKIP neg samp sentiment embedding refinement	0.466	0.500
Full	SKIP sentiment change average	0.503	0.512
	SKIP sentiment change weight sum	0.505	0.477
	SKIP neg samp sentiment change average	0.497	0.489
	SKIP neg samp sentiment change weight sum	0.485	0.497

ware engineering texts, where in-domain (Stackoverflow) embeddings were compared to generic ones (GoogleNews), did not yield such clearcut results (Biswas et al., 2019).

We have also suggested a simple strategy to train sentiment embeddings, and shown that it outperforms other existing methods for this task. More in general, sentiment embeddings perform consistently better than plain embeddings for both languages in the "tablet" domain, but less evidently so in the automobile domain. The reason for this requires further investigation. Further testing is also necessary to assess the influence of vector size in our experiments. Indeed, not all embeddings are trained with the same dimensions, an aspect that might also affect performance differences, though the true impact of size is not yet fully understood (Yin and Shen, 2018).

In terms of different embeddings types, it would be also interesting to compare our simple embedding refinement method, which takes specific contextual occurrences into account, with the performance of contextual word embeddings (Peters et al., 2018; Devlin et al., 2019), which work directly at the token rather than the type level. More complex training strategies could also be explored (Dong and De Melo, 2018).

Acknowledgments

We would like to thank the Center for Information Technology of the University of Groningen for providing access to the Peregrine high perfor-

mance computing cluster which we used to run the experiments reported in this paper. We are also grateful to the reviewers for helpful comments.

References

- Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. 2010. Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. volume 10, 01.
- Francesco Barbieri, Valerio Basile, Danilo Croce, Malvina Nissim, Nicole Novielli, and Viviana Patti. 2016. Overview of the EVALITA 2016 sentiment polarity classification task (SENTIPOLC). In *Proceedings of the 5th evaluation campaign of natural language processing and speech tools for Italian (EVALITA 2016)*.
- Jeremy Barnes, Roman Klinger, and Sabine Schulte im Walde. 2017. Assessing state-of-the-art sentiment models on state-of-the-art sentiment datasets. *arXiv preprint arXiv:1709.04219*.
- Valerio Basile and Malvina Nissim. 2013. Sentiment analysis on italian tweets. In *Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 100–107.
- Valerio Basile, Nicole Novielli, Danilo Croce, Francesco Barbieri, Malvina Nissim, and Viviana Patti. 2018. Sentiment polarity classification at evalita: Lessons learned and open challenges. *IEEE Transactions on Affective Computing*.
- Eeshita Biswas, K Vijay-Shanker, and Lori Pollock. 2019. Exploring word embedding techniques to improve sentiment analysis of software engineering texts. In *Proceedings of the 16th International Conference on Mining Software Repositories*, pages 68–78. IEEE Press.
- Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2016. Enriching word vectors with subword information. *CoRR*, abs/1607.04606.
- Alessandra Teresa Cignarella, Simona Frenda, Valerio Basile, Cristina Bosco, Viviana Patti, Paolo Rosso, et al. 2018. Overview of the evalita 2018 task on irony detection in italian tweets (ironita). In *Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian (EVALITA 2018)*, volume 2263, pages 1–6. CEUR-WS.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June. Association for Computational Linguistics.

- Xin Dong and Gerard De Melo. 2018. A helping hand: Transfer learning for deep sentiment analysis. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2524–2534.
- Manaal Faruqui, Jesse Dodge, Sujay K. Jauhar, Chris Dyer, Eduard Hovy, and Noah A. Smith. 2015. Retrofitting word vectors to semantic lexicons. In *Proceedings of NAACL*.
- Svetlana Kiritchenko and Saif M Mohammad. 2017. Capturing reliable fine-grained sentiment associations by crowdsourcing and best-worst scaling. *arXiv preprint arXiv:1712.01741*.
- Bing Liu, Minqing Hu, and Junsheng Cheng. 2005. Opinion observer: analyzing and comparing opinions on the web. In *Proceedings of the 14th international conference on World Wide Web*, pages 342–351. ACM.
- Tomas Mikolov, Kai Chen, G.s Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *Proceedings of Workshop at ICLR*, 2013, 01.
- Huy Tien Nguyen and Minh Le Nguyen. 2018. Multi-lingual opinion mining on youtube—a convolutional n-gram bilstm word embedding. *Information Processing & Management*, 54(3):451–462.
- Malvina Nissim and Viviana Patti. 2017. Semantic aspects in sentiment analysis. In *Sentiment analysis in social networks*, pages 31–48. Elsevier.
- Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.
- Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, volume 1, pages 2227–2237.
- Ruggero Petrolito and Felice Dell’Orletta. 2018. Word embeddings in sentiment analysis. In *CLiC-it*.
- Maria Pontiki, Dimitris Galanis, John Pavlopoulos, Harris Papageorgiou, Ion Androutsopoulos, and Suresh Manandhar. 2014. SemEval-2014 task 4: Aspect based sentiment analysis. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 27–35, Dublin, Ireland. Association for Computational Linguistics.
- Radim Řehůřek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta, May. ELRA. <http://is.muni.cz/publication/884893/en>.
- Aliaksei Severyn, Alessandro Moschitti, Olga Uryupina, Barbara Plank, and Katja Filippova. 2016. Multi-lingual opinion mining on youtube. *Information Processing & Management*, 52(1):46–60.
- Duyu Tang, Furu Wei, Nan Yang, Ming Zhou, Ting Liu, and Bing Qin. 2014. Learning sentiment-specific word embedding for twitter sentiment classification. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1555–1565. Association for Computational Linguistics.
- Tun Thura Thet, Jin-Cheon Na, and Christopher SG Khoo. 2010. Aspect-based sentiment analysis of movie reviews on discussion boards. *Journal of information science*, 36(6):823–848.
- Olga Uryupina, Barbara Plank, Aliaksei Severyn, Agata Rotondi, and Alessandro Moschitti. 2014. Sentube: A corpus for sentiment analysis on youtube social media. In *LREC*, pages 4244–4249.
- Cynthia Van Hee, Els Lefever, and Véronique Hoste. 2018. Semeval-2018 task 3: Irony detection in english tweets. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 39–50.
- Amy Beth Warriner, Victor Kuperman, and Marc Brysbaert. 2013. Norms of valence, arousal, and dominance for 13,915 english lemmas. *Behavior Research Methods*, 45(4):1191–1207, Dec.
- Zi Yin and Yuanyuan Shen. 2018. On the dimensionality of word embedding. In *Advances in Neural Information Processing Systems*, pages 887–898.
- Liang-Chih Yu, Jin Wang, K Lai, and Xuejie Zhang. 2017. Refining word embeddings for sentiment analysis. pages 534–539, 01.

A Novel Integrated Industrial Approach with Cobots in the Age of Industry 4.0 through Conversational Interaction and Computer Vision

Andrea Pazienza and Nicola Macchiarulo and Felice Vitulano and
Antonio Fiorentini and Marco Cammisa

Innovation Lab, Exprivia S.p.A.

{andrea.pazienza, nicola.macchiarulo, felice.vitulano,
antonio.fiorentini, marco.cammisa}@exprivia.com

Leonardo Rigutini and Ernesto Di Iorio and Achille Globo and Antonio Trevisi
QuestIT S.r.l.

{rigutini, diiorio, globo, trevisi}@quest-it.com

Abstract

English. From robots that replace workers to robots that serve as helpful colleagues, the field of robotic automation is experiencing a new trend that represents a huge challenge for component manufacturers. The contribution starts from an innovative vision that sees an ever closer collaboration between Cobot, able to do a specific physical job with precision, the AI world, able to analyze information and support the decision-making process, and the man able to have a strategic vision of the future.

1 Introduction

In the last century, the manufacturing world has adopted solutions for the advanced automation of production systems. Today, thanks to the evolution and maturity of new technologies such as Artificial Intelligence (AI), Machine Learning (ML), new generation networks, and the growing adoption of the Internet of Things (IoT) approach, a new paradigm emerges, aiming at integrating the Cyber-Physical System (CPS) with business processes, thus opening the doors to the fourth industrial revolution (Industry 4.0) and that will allow us to join in the era driven by information and further handled with cognitive computing techniques (Wenger, 2014).

Robots and humans have been co-workers for years, but rarely have we been truly working together. This may be about to change with the rise of Collaborative Robotics (Colgate et al.,

1996). Collaborative Robots (better known as **Cobots**) are specifically designed for direct interaction with a human within a defined collaborative work-space, i.e., a safeguard space where the robot and a human can perform tasks simultaneously during an automatic operation. Then, human-robot collaboration fosters various levels of automation and human intervention. Tasks can be partially automated if a fully automated solution is not economical or too complex. Therefore, manufacturers may benefit from the rising of AI-driven automation, and the progress of **Adaptable End Effectors** devices, mounted at the end of Cobot's arms, may help to perform specific intelligent tasks (Dubey and Crowder, 2002).

The way in which Cobots and humans interact, exchanging and conveying information, is fundamental. The key role in this landscape would be addressed by **Conversational Interfaces** (Zue and Glass, 2000), which exploit and take advantages from the recent achievements in the field of Natural Language Processing (NLP), to understand user need and generate the right answer or action. In this scenario, **Computer Vision** also plays an important role in the process of creating collaborative environments between humans and robots. Systems of this type are already introduced into the industry to facilitate tasks of product quality control or component assembly inspection. By giving vision to a robot, it can make it able to understand the industrial environment that surrounds it and can improve the execution of tasks in support to other people.

Improving robots software with AI will be key to making robots more collaborative. The work starts from an innovative vision that beholds, in the future, an ever closer collaboration between

Copyright © 2019 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

Cobot, able to do a specific physical job with precision and without alienation, the AI world, able to analyze, process, and learn from information and support the decision-making process, and the employee able to have a strategic vision of the future. To validate its effectiveness, a collaborative environment between employee, Cobot and AI systems has been crafted to make possible the three subjects communicate in a simple way and without requiring the employee to have specific skills to interact with the Cobot and Enterprise Resource Planning (ERP) systems.

Our contribution is indeed placed in this scenario where the convergence of multiple technologies allows us to define a new approach related to the management of a core business process (e.g. shipments) which tends to ensure more and more flexibility of the process thanks to a simplification of human interaction with Cyber-Physical Systems, with a better coordination between the physical world (the packaging line), and that of IT processes (the ERP model). In the belief that the complexity of new industrial production systems requires interdisciplinary skills, our intents are to bring together knowledge from related disciplines such as computational linguistics, cognitive science, machine learning, computer vision, human-machine interaction, and collaborative robotics automation towards an integrated novel approach specifically designed for the smart management of a manufacturing process line by fostering and strengthening the synergy and the interaction between robot and human.

Our research is broadly situated in Human-Robot Collaboration (HRC), a promising robotics discipline focusing on enabling robots and humans to operate jointly to complete collaborative tasks. Recent works tried to figure out in which way Cobots may help humans in collaborative industrial tasks (El Zaatari et al., 2019) or in participatory design in fablabs (Ionescu and Schlunda, 2019). An initial study centered cobots in advanced manufacturing systems (Djuric et al., 2016). No or little work (Ivorra et al., 2018) is done to endow Cobots with cognitive intelligence like conversational interaction and computer vision.

This paper is organized as follows. Section 2 introduces the functionalities and the architecture of our approach, focusing on the main four technological aspects: cobots, adaptable end effectors, conversational interfaces, and computer vi-

sion. Section 3 describes the possible scenarios of application specifically designed for our approach, such as Smart Manufacturing. Finally, Section 4 discusses the proposed framework and concludes the paper, outlining future works.

2 Architecture Proposal

In this Section we introduce our main proposal taking into account all the requirements coming from different technologies. The leading idea is to develop and validate a general framework concerning an Intelligent Cyber-Physical System made up of four crucial components: (i) a Cobot, equipped with (ii) an adaptable end effector, which may change according to a specific scenario, and two major components coming from the AI world, i.e. (iii) a Computer Vision module to allow the cobot detecting an object, and (iv) one or more Conversational Interfaces to facilitate the human-machine interaction and keep the man in the loop. Figure 1 depicts the prototypical architecture of our framework proposal.

In order to integrate different technologies, from the high-level voice command to low-level execute command, we developed a web application, powered by *Spring Boot* framework², able to receive commands from user interfaces and transform them into machine commands. We consider in this framework the possibility to give vocal commands to the Cobot. In this perspective, the mechanical arm is controlled through a series of connected conversational devices like chat-bots, powered by *Cisco WebEx Teams*³ and *QuestIT Algho*⁴, and a virtual assistant such as *Amazon Alexa*⁵. In particular, Cisco WebEx Teams is an all-in-one solution for messaging, file sharing, white boarding, video meetings, and calling, while Amazon Alexa is a voice interaction device capable of a large set of human-interacting functions. The Cobot, through the use of a camera, is able to acquire images and process through the use of Computer Vision algorithms, recognizing exactly the object to be selected without knowing its position in advance. Hence, the vocal commands sent via Alexa are managed by lambda functions using the *AWS Lambda* service⁶, which is a serverless event-driven computing platform. It permits to ex-

²<https://spring.io/projects/spring-boot>

³<https://www.webex.com/team-collaboration.html>

⁴<https://www.alghoncloud.com/it/>

⁵<https://developer.amazon.com/it/alexa>

⁶<https://aws.amazon.com/it/lambda/>

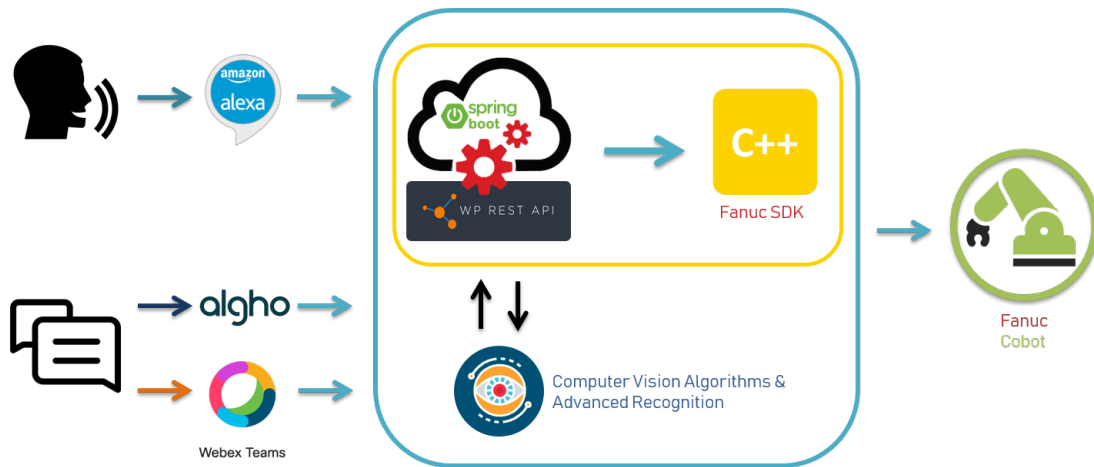


Figure 1: Framework Architecture including Conversational Interfaces, Computer Vision and Cobot

ecute code in response to particular events, automatically managing also the resources required by the programming code. Indeed, lambda's goal is to simplify the construction of on-demand applications that respond to events and new information.

Therefore, all commands are sent via HTTP calls to the web application using the Spring Boot framework, receiving also calls from the one or more chat-bots with which the user can interact. Once a command has been received, the web application executes a C# application, based on Fanuc SDK, that sends to the Cobot the request to execute a particular script written in Teach Pendant (TP) language.

2.1 Cobot with Fanuc

The right choice of a cobot comes with the fulfillment of various safety requirements, such as a collision stop protection, a function to restart them easily and quickly after a stop, and anti-trap features for additional protection. For our purposes, we used a Cobot from Fanuc, in particular the *CR-4iA* model⁷. It is endowed with six axis in its arm, and its maximum payload is 4 kg. Also, it handles lightweight tasks that are tedious, highly manual. Since it can take over these dull jobs, the operator hands are free to focus on more intelligent work or even more pressing matters. This cobot can also work side-by-side on tasks that are more complex, and require more interactive approaches.

⁷<https://www.fanuc.eu/it/it/robot/robot-filter-page/robot-collaborativi/collaborative-cr4ia>

2.2 Adaptable End Effector with Schunk

This category includes grippers, which hold and manipulate objects, and end-of-arm tools (EOATs), which are complex systems of grippers designed to handle large or delicate components. Handling tasks mainly include pick and place, sorting, packaging, and palletizing. As gripping tool we used the *Schunk Co-act EGP-C* gripper⁸. It is an Electric 2-finger parallel gripper certified for collaborative operation with actuation via 24 V and digital I/O. It is used for gripping and moving small and medium-sized workpieces with flexible force in collaborative operation in the areas of assembly, electronics and machine tool loading. We chose this model due to its certified and pre-assembled gripping unit with functional safety, and its "plug & work" mode with Fanuc cobots.

2.3 Conversational Interfaces with Algho

The achievements in the field of Artificial Intelligence (AI) in the recent years have led to the birth of a new paradigm of human-machine interaction: the conversational agents. This new way of interacting with a computer is based on the use of natural language and is getting closer to the way humans communicate with each other. Conversational agents take advantage of recent achievements in the field of Natural Language Processing (NLP-U) to understand user requests and behave accordingly, providing appropriate answers or performing required actions. The design of an innovative Cobot cannot fail to consider the use of a such straightforward human-machine interaction.

⁸https://schunk.com/it_it/co-act/pinza-co-act-egp-c/

The conversational functionalities for the Cobot described in this paper have been provided by using Algho⁴, a proprietary conversational-agent building tool developed by QuestIT⁹ and based on NLP and AI techniques. In particular, Algho is a suite designed to facilitate the creation of personal conversational agents and the subsequent deploy on several proprietary channels. The user of Algho can create his own chat-bot simply by entering the personal knowledge base and the system, after a few minutes, is able to handle conversations about it. The natural language understanding functionalities of Algho are based on a proprietary NLP Platform developed by QuestIT⁹ consisting in more than 25 layers of morphological, syntactic and semantic analysis based on Machine Learning (ML) and Artificial Intelligence techniques: tokenization, lemmatization, Part-Of-Speech (POS), Collocation Detection, Word Sense Disambiguation, Dependency Tree Parsing, Sentiment and Emotional Analysis, Intent Recognition, and many others. The NLP Platform exploits the most recent techniques in the field of NLP and Machine Learning to enrich the input raw text with a set of high-level cognitive information (Melacci et al., 2018; Bongini et al., 2018). The Word Sense Disambiguation (WSD) layer is one of the main levels of the NLP Platform and it follows a Deep Neural Network approach based on RNN and word embedding. It provides state-of-the-art performances with regard to the disambiguation accuracy (Melacci et al., 2018; Bongini et al., 2018).

The enriched text is subsequently exploited by the conversational engine to understand the user request, to identify the “intent” and to behave accordingly to the knowledge base provided by the creator of the conversational agent. The intent of a request is defined as the hidden desire that underlies the user’s request.

During the construction of the conversational agent, the Algho suite allows the user to define specific objects called “Conversational Form” which can be used to collect structured information from the user. In particular, a “conversational form” consists in a typical form for collecting data which is linked to a set of intent defined in the knowledge base. During the conversation, when an input user request triggers an intent having a linked conversational form, the system: (i) tries to fill the form fields by extracting the information

from the NLP analysis of the request (Auto-Form-Filling procedure); and (ii) proposes sequentially to the user the fields that have not been filled by the automatic procedure. When an user input request trigger a conversational form, the returned NLP information are used to automatically fill the fields of the structured form without requesting further data from the user. Furthermore, Algho allows to specify an URL to which the collected information can be sent via the call to a web-service. In this case, the system uses the field’s values as parameters for the call to the service.

2.4 Computer Vision

The computer vision functionalities for the described work have been implemented with two open source libraries, OpenCV and TensorFlow. OpenCV (Laganière, 2014) provides the state-of-the-art algorithms in this field and, starting from version 4.0, has introduced more advanced features for deep learning. TensorFlow (Abadi et al., 2016) is a library to develop and train machine learning models, in particular its used to create deep neural networks. Our approach follows a general pipeline composed of three main steps:

- Dataset creation: several images of the objects of interest are collected and their position is annotated manually by specifying their coordinates;
- Training the model: a model is trained in order to recognize the objects of our interest and its coordinates within the image. For this purpose, we decided to fine tune the model Faster R-CNN (Ren et al., 2015) with Inception V2 (Szegedy et al., 2016) pre-trained on the COCO dataset (Lin et al., 2014);
- Using the model: the detection of the requested object through the conversational interface is performed in real time by analysing a video stream received from a video camera.

3 Exprivia’s Use Case Scenarios

Exprivia prototyped this general framework in two different use case scenarios, with the main target of enabling communication between all the machines and ICT systems located in a factory in a capillary way, ranging from supply chain systems to administrative ones. The ultimate goal is to manage of the entire production life-cycle to

⁹<https://www.quest-it.com>

a cost saving optimization of each resource that turns into an advantage, not only economical but also competitive, allowing company to play a leading role in the challenge of the future.

Food Supply Chain. An interesting example of the application of our framework has been made within the food supply chain, in particular referring to the *pasta creation chain*, presented at the DevNet Create 2019 conference in Mountain View (California) in April. The purpose of the project was to automate a series of activities typical of daily operations, specifically to medium-high difficulty activities that are the cause of most problems in the production life-cycle. Pasta creation process is very complex and requires a concatenation of different work steps. Many of these are performed manually (e.g. quality control) and typically the machines are not able to communicate with each other: this means that operators and the management cannot have information on the operating status. Thanks to our framework that includes a chat-bot to communicate with the machinery and computer vision algorithms able to automate the pasta quality control, the communication with management systems enables a two-way exchange of information that automates activities, improving overall operating efficiency.

Coffee Pod Selection with Nuccio. The following solution provides the possibility to use a Fanuc Cobot to select a *coffee pod*. This prototype has been presented at Mobile World Congress 2019, in Barcelona, in February. The Cobot “Nuccio” is controlled through the Algho conversational interface. In particular, the idea was to create a conversational agent focused on a specific knowledge base about coffee. The resulting bot was able to handle conversations about coffee and about many aspects related to this topic. Afterwards, a specific “conversational form” was developed for collecting a set of information useful for preparing a coffee (taste, aroma, sugar, short or long) and required by the actuator system. Finally, the form has been connected with the web-service of the actuator system and linked to the set of intents for which activation was desired. Thus, the resulting bot was able to handle conversation concerning coffee and if the user request deals with the intent to have a coffee, the linked conversational form allows to collect all the information required by the actuator system to prepare the coffee and to notify via a web service call. More-

over, Nuccio, through the use of a camera, was able to acquire images and process through the use of Computer Vision algorithms, recognizing exactly the pod to be selected without knowing the position in advance. Through the Algho conversational interface, the user is helped and guided in the choice of the most suitable coffee pod, according to his/her tastes.

4 Conclusion

In line with the main objectives, we contributed to the development and validation of a framework in an operational environment of intelligent robotic systems and HRC. In particular, we dealt with conversational interaction technologies useful to perform: (i) high-performance linguistic analysis services based on NLP technologies; (ii) models for the symbiotic human-robot interaction management; (iii) services and tools for the adaptation of linguistic interfaces with respect to user characteristics. The Cobots are close to operating in environments where the presence of man plays a key role. A fundamental characteristic is therefore the Cobot’s ability in reacting to textual and vocal commands to properly understand the user’s commands. The Cobot’s perception is leveraged with its ability to detect object and understand what there is around him; computer vision processing becomes crucial to the extent of giving Cobots a cognitive profile. We therefore envision our framework to be fully operable in complex manufacturing systems, in which the collaboration between robot and man is facilitated by advanced AI and cognitive techniques.

We showed how, already today, it is possible to “humanize” highly automated processes through a Cobot, collecting and integrating the operational information in the corporate knowledge base. In fact, we believe that in the long term there will be a convergence between automation, AI and IoT, allowing the market to create a full “Digital Twin” with an organization that will lead to a strong automation of organizational choices driven by data collected in the field. The digitized organization can then be equipped with its own “Company Brain”, an AI able to make autonomous complex decisions aimed at maximizing a business goal that, working in a cooperative manner with the company management, will be able to respond much more precisely and quickly to changes in an increasingly unstable and fluid market.

References

- Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. 2016. Tensorflow: A system for large-scale machine learning. In *12th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 16)*, pages 265–283.
- Marco Bongini, Leonardo Rigutini, and Edmondo Trentin. 2018. Recursive neural networks for density estimation over generalized random graphs. *IEEE transactions on neural networks and learning systems*, (99):1–18.
- James Edward Colgate, Michael A. Peshkin, and Witaya Wannasuphoprasit. 1996. Cobots: Robots for collaboration with human operators. In *Proceedings of the International Mechanical Engineering Congress and Exhibition, Atlanta, GA*, volume 58, pages 433–439. Citeseer.
- Ana M. Djuric, R.J. Urbanic, and J.L. Rickli. 2016. A framework for collaborative robot (cobot) integration in advanced manufacturing systems. *SAE International Journal of Materials and Manufacturing*, 9(2):457–464.
- Venketesh N. Dubey and Richard M. Crowder. 2002. A finger mechanism for adaptive end effectors. In *ASME 2002 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*, pages 995–1001. American Society of Mechanical Engineers.
- Shirine El Zaatari, Mohamed Marei, Weidong Li, and Zahid Usman. 2019. Cobot programming for collaborative industrial tasks: An overview. *Robotics and Autonomous Systems*, 116:162–180.
- Tudor B. Ionescua and Sebastian Schlunda. 2019. A participatory programming model for democratizing cobot technology in public and industrial fablabs. *Procedia CIRP*, 81:93–98.
- Eugenio Ivorra, Mario Ortega, Mariano Alcañiz, and Nicolás García-Aracil. 2018. Multimodal computer vision framework for human assistive robotics. In *2018 Workshop on Metrology for Industry 4.0 and IoT*, pages 1–5. IEEE.
- Robert Laganière. 2014. *OpenCV Computer Vision Application Programming Cookbook Second Edition*. Packt Publishing Ltd.
- Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer.
- Stefano Melacci, Achille Globo, and Leonardo Rigutini. 2018. Enhancing modern supervised word sense disambiguation models by semantic lexical resources. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*, Miyazaki, Japan, May. European Languages Resources Association (ELRA).
- Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. FasterR-CNN: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99.
- Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826.
- Etienne Wenger. 2014. *Artificial intelligence and tutoring systems: computational and cognitive approaches to the communication of knowledge*. Morgan Kaufmann.
- Victor W Zue and James R Glass. 2000. Conversational interfaces: Advances and challenges. *Proceedings of the IEEE*, 88(8):1166–1180.

Annotating Hate Speech: Three Schemes at Comparison

**Fabio Poletto, Valerio Basile,
Cristina Bosco, Viviana Patti**

Dipartimento di Informatica
University of Turin
{poletto,basile,
bosco,patti}@di.unito.it

Marco Stranisci

Acmos
marco.stranisci@acmos.net

Abstract

Annotated data are essential to train and benchmark NLP systems. The reliability of the annotation, i.e. low inter-annotator disagreement, is a key factor, especially when dealing with highly subjective phenomena occurring in human language. Hate speech (HS), in particular, is intrinsically nuanced and hard to fit in any fixed scale, therefore crisp classification schemes for its annotation often show their limits. We test three annotation schemes on a corpus of HS, in order to produce more reliable data. While rating scales and best-worst-scaling are more expensive strategies for annotation, our experimental results suggest that they are worth implementing in a HS detection perspective.¹

1 Introduction

Automated detection of hateful language and similar phenomena — such as offensive or abusive language, slurs, threats and so on — is being investigated by a fast-growing number of researchers. Modern approaches to Hate Speech (HS) detection are based on supervised classification, and therefore require large amounts of manually annotated data. Reaching acceptable levels of inter-annotator agreement on phenomena as subjective as HS is notoriously difficult. Poletto et al. (2017), for instance, report a “very low agreement” in the HS annotation of a corpus of Italian tweets, and similar annotation efforts showed similar results (Del Vigna et al., 2017; Waseem, 2016; Gitari et al., 2015; Ross et al., 2017). In an attempt to tackle the agreement issue, annotation schemes have been proposed based on numeric

scales, rather than strict judgments (Kiritchenko and Mohammad, 2017). *Ranking*, rather than *rating*, has also proved to be a viable strategy to produce high-quality annotation of subjective aspects in natural language (Yannakakis et al., 2018). Our hypothesis is that binary schemes may oversimplify the target phenomenon, leaving it uniquely to the judges’ subjectivity to sort less prototypical cases and likely causing higher disagreement. Rating or ranking schemes, on the other hand, are typically more complex to implement, but they could provide higher quality annotation.

A framework is first tested by annotators: inter-annotator agreement, number of missed test questions and overall opinion are some common standards against which the quality of the task can be tested. A certain degree of subjectivity and bias is intrinsic to the task, but an effective scheme should be able to channel individual interpretations into unambiguous categories.

A second reliability test involves the use of annotated data to train a classifier that assigns the same labels used by humans to previously unseen data. This process, jointly with a thorough error analysis, may help spot bias in the annotation or flaws in the dataset construction.

We aim to explore whether and how different frameworks differ in modeling HS, what problems do they pose to human annotators and how suitable they are for training. In particular, we apply a binary annotation scheme, as well as a rating scale scheme and a best-worst scale scheme, to a corpus of HS. We set up experiments in order to assess whether such schemes help achieve a lower disagreement and, ultimately, a higher quality dataset for benchmarking and for supervised learning.

The experiment we set up involves two stages: after having the same dataset annotated with three different schemes on the crowdsourcing platform Figure Eight², we first compare their agreement

¹Copyright © 2019 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

²<https://www.figure-eight.com/>.

rates and label distributions, then we map all schemes to a “yes/no” structure to perform a cross-validation test with a SVM classifier. We launched three separate tasks on the platform: Task 1 with a binary scheme, Task 2 with an asymmetric rating scale, and Task 3 with a best-worst scale. For each task, a subset has been previously annotated by experts within the research team, to be used as gold standard against which to evaluate contributors’ trustworthiness on Figure Eight.

2 Related Work

Several frameworks have been proposed and tested so far for HS annotation, ranging from straightforward binary schemes to complex, multi-layered ones and including a variety of linguistic features. Dichotomous schemes are used, for example, by Alfina et al. (2017), Ross et al. (2017) and Gao et al. (2017) for HS, by Nithyanand et al. (2017) for offensiveness and by Hammer (2016) for violent threats. Slightly more nuanced frameworks try to highlight particular features. Davidson et al. (2017) distinguish between *hateful*, *offensive but not hateful* and *not offensive*, as do Mathur et al. (2018) who for the second type use the label *abusive* instead; similarly, Mubarak et al. (2017) use the labels *obscene*, *offensive* and *clean*. Waseem (2016) differentiate hate according to its target, using the labels *sexism*, *racism*, *both* and *none*. Nobata et al. (2016) uses a two-layer scheme, where a content can be first labeled either as *abusive* or *clean* and, if abusive, as *hate speech*, *derogatory* or *profanity*. Del Vigna et al. (2017) uses a simple scale that distinguishes between *no hate*, *weak hate* and *strong hate*.

Where to draw the line between weak and strong hate is still highly subjective but, if nothing else, the scheme avoids feebly hateful comments to be classified as not hateful (thus potentially neutral or positive) just because, strictly speaking, they can not be called HS. Other authors, such as Olteanu et al. (2018) and Fišer et al. (2017), use heavier and more elaborated schemes. Olteanu et al. (2018), in particular, experimented with a rating-based annotation scheme, reporting low agreement. Sanguinetti et al. (2018) also uses a complex scheme in which HS is annotated both for its presence (binary value) and for its intensity (1–4 rating scale). Such frameworks potentially provide valuable insights into the investigated issue, but as a downside they make the whole

annotation process very time-consuming. More recently, a ranking scheme has been applied to the annotation of a small dataset of German hate speech messages (Wojatzki et al., 2018).

3 Annotation Schemes

In this section, we introduce the three annotation schemes tested in our study.

Binary. Binary annotation implies assigning a binary label to each instance. Beside HS, binary classification is common in a variety of NLP tasks and beyond. Its simplicity allows a quick manual annotation and an easy computational data processing. As a downside, such a dichotomous choice presupposes that is always possible to clearly and objectively determine what answer is true. This may be acceptable in some tasks, but it is not always the case with human language, especially for more subjective and nuanced phenomena.

Rating Scales. Rating Scales (RS) are widely used for annotation and evaluation in a variety of tasks. Likert scale is the best known (Likert, 1932): values are arranged at regular intervals on a symmetric scale, from the most to the least typical of a given concept. It is suitable for measuring subjective opinion or perception about a given topic with a variable number of options. Compared to binary scheme, scales are better for managing subjectivity and intermediate nuances of a concept. On the other hand, as pointed out by (Kiritchenko and Mohammad, 2017), they present some flaws: high inter-annotator disagreement (the more fine-grained the scale, the higher the chance of disagreement), individual inconsistencies (judges may express different values for similar items, or the same value for different items), scale region bias (judges may tend to prefer values in one part of the scale, often the middle) and fixed granularity (which may not represent the actual nuances of a concept).

Best-Worst Scaling. The Best-Worst Scaling model (BWS) is a comparative annotation process developed by Louviere and Woodworth (1991). In a nutshell, a BWS model presents annotators with n items at a time (where $n > 1$ and normally $n = 4$) and asks them to pick the best and worst ones with regard to a given property. The model has been used in particular by Kiritchenko

ethnic group	religion	Roma
immigrat*, immigrazione	terrorismo	rom
migrant*, profug*	terrorist*, islam	nomad*
stranier*	mus[s]ulman*	
	corano	

Table 1: List of keywords used to filter our dataset.

and Mohammad (2017) and Mohammad and Kiritchenko (2018), who proved it to be particularly effective for subjective tasks such as sentiment intensity annotation, which are intrinsically nuanced and hardly fit in any fixed scale.

4 Dataset and task description

For our experiment, we employ a dataset of 4,000 Italian tweets, extracted from a larger corpus collected within the project *Contro l’odio*³. For the purpose of this research, we filtered all the tweets written between November 1st and December 31st with a list of keywords. This list, reported in Table 1, is the same proposed in Poletto et al. (2017) for collecting a dataset focused on three typical targets of discrimination — namely Immigrants, Muslims and Roma.

The concept of HS underlying all three annotation tasks includes any expression based on intolerance and promoting or justifying hatred towards a given target. For each task we explicitly asked the annotators to consider only HS directed towards one of the three above-mentioned targets, ignoring other targets if present. Each message is annotated by at least three contributors. Figure Eight also report a measure of agreement computed as a Fleiss’ κ weighted by a score indicating the trustworthiness of each contributor on the platform. We note, however, that the agreement measured on the three tasks is not directly comparable, since they follow different annotation schemes.

4.1 Task 1: Binary Scheme.

The first scheme is very straightforward and simply asks judges to tell whether a tweet contains HS or not. Each line will thus receive the label *HS_yes* or *HS_no*. The definition of HS is drawn by (Poletto et al., 2017). In order to be labeled as hateful, a tweet must:

- address one of above-mentioned targets;
- either incite, promote or justify hatred, violence or intolerance towards the target, or de-

³<https://controlodio.it/>.

label	tweet
yes	<i>Allora dobbiamo stringere la corda: pena capitale per tutti i musulmani in Europa immediatamente!</i> Then we have to adopt stricter measures: death penalty for all Muslims in Europe now!
no	<i>I migranti hanno sempre il posto e non pagano.</i> Migrants always get a seat and never pay.

Table 2: Annotation examples for Task 1 (gold labels).

mean, dehumanise or threaten it.

We also provided a list of expressions that are not to be considered HS although they may seem so: for example, these include slurs and offensive expressions, slanders, and blasphemy. An example of annotation for this task is presented in Table 2.

4.2 Task 2: Unbalanced Rating Scale

This task requires judges to assign a label to each tweet on a 5-degree asymmetric scale (from 1 to -3) that encompasses the content and tone of the message as well as the writer’s intention. Again, the target of the message must be one of three mentioned above. The scheme structure is reported in Table 3, while Table 4 shows an example for each label.

label	meaning
+1	positive
0	neutral, ambiguous or unclear
-1	negative and polite, dialogue-oriented attitude
-2	negative and insulting/abusive, aggressive attitude
-3	strongly negative with overt incitement to hatred, violence or discrimination, attitude oriented at attacking or demeaning the target

Table 3: Annotation scheme for Task 2: evaluate the stance or opinion expressed in each tweet.

This scale was designed with a twofold aim: to avoid a binary choice that could leave too many doubtful cases, and to split up negative contents in more precise categories, in order to distinguish different degrees of “hatefulness”.

We tried not to influence annotators by matching the grades of our scale in Task 2 to widespread concepts such as stereotypes, abusive language or hateful language, which people might tend to apply by intuition rather than by following strict rules. Instead, we provided definitions as neutral and objective as possible, in order to differentiate this task from the others and avoid biases. An asymmetric scale, although unusual, fits our purpose of an in-depth investigation of negative language very well. A possible downturn of this

label	tweet
+1	<i>Gorino Alla fine questi profughi l'hanno scampata bella. Vi immaginate avere tali soggetti come vicini di casa?</i> These asylum-seekers had a narrow escape. Can you imagine having such folks (TN: racist Gorino inhabitants) as neighbours?
0	<i>Bellissimo post sulle cause e conseguenze dell'immigrazione, da leggere!</i> Great post on causes and consequences of immigration, recommended!
-1	<i>I migranti hanno sempre il posto e non pagano.</i>
-2	<i>Con tutti i soldi elargiti ai rom, vedere il degrado nel quali si crogiolano, non meritano di rimanere in un paese civile!</i> Seeing the decay Roma people wallow in, despite all the money lavished on them, they don't deserve to stay in a civilized country!
-3	<i>Allora dobbiamo stringere la corda: pena capitale per tutti i musulmani in Europa immediatamente!</i>

Table 4: Examples of annotation for Task 2 (gold labels).

scheme is that grades in the scale are supposed to be evenly spread, while the real phenomena they represent may not be so.

4.3 Task 3: Best-Worst Scaling

The structure of this task differs from the previous two. We created a set of tuples made up by four tweets (4-tuples), grouped so that each tweet is repeated four times in the dataset, combined with three different tweet each time. Then we provided contributors with a set of 4-tuples: for each 4-tuple they were asked to point out the most hateful and the least hateful of the four. Judges have thus seen a given tweet four times, but have had to compare it with different tweets every time⁴. This method avoids assigning a discrete value to each tweet and gathers information on their “hatefulness” by comparing them to other tweets. An example of annotation, with the least and most hateful tweets marked in a set of four, is provided in Table 5.

5 Task annotation results

In Task 1, the distribution of the labels *yes* and *no*, referred to the presence of HS, conforms to that of other similar annotated HS datasets, such as Burnap and Williams (2015) in English and Sanguinetti et al. (2018) in Italian. After applying a majority criterion to non-unanimous cases, tweets labeled as HS are around 16% of the dataset (see Figure 1). Figure Eight measures the agreement in terms of *confidence*, with a κ -like func-

⁴The details of the tuple generation process are explained in this blog post: <http://valeriobasile.github.io/Best-worst-scaling-and-the-clock-of-Gauss/>

label	tweet
least	<i>Roma, ondata di controlli anti-borseggio in centro: arrestati 8 nomadi, 6 sono minorenni.</i> Rome, anti-pickpocketing patrolling in the centre: 8 nomads arrested, 6 of them are minor. <i>Tutti i muslims presenti in Europa rappresentano un pericolo mortale latente. L'islam è incompatibile con i valori occidentali.</i> All Muslims in Europe are a dormant deadly danger. Islam is incompatible with Western values. <i>Trieste, profughi cacciano disabile dal bus: arrivano le pattuglie di Forza Nuova sui mezzi pubblici.</i> Trieste, asylum-seekers throw disabled person off the bus: Forza Nuova (TN: far-right, nationalist fringe party) to patrol public transport.
most	<i>Unica soluzione è cacciare TUTTI i musulmani NON integrati fino alla 3a gen che si ammazzassero nei loro paesi come fanno da secoli MALATI!</i> Only way is to oust EVERY NON-integrated Muslim down to 3rd generation let them kill each other in their own countries as they've done for centuries INSANE!

Table 5: Examples of annotation for Task 3: 4-tuple with marks for the least hateful and the most hateful tweets.

tion weighted by the *trust* of each contributor, i.e., a measure of their reliability across their history on the platform. On task 1, about 70% of the tweets were associated with a confidence score of 1, while the remaining 30% follow a low-variance normal distribution around .66.

As for Task 2, label distribution tells a different story. When measuring inter-annotator agreement, the mean value between all annotations has been computed instead of using the majority criterion. Therefore, results are grouped in intervals rather than in discrete values, but we can still easily map these intervals to the original labels. As shown in Figure 1, tweets labeled as having a neutral or positive content (in green) are only around 27%, less than one third of the tweets labeled as non-hateful in Task 1. Exactly half of the whole dataset is labeled as negative but oriented to dialogue (in yellow), while 20% is labeled as negative and somewhat abusive (orange) and only less than 3% is labeled as an open incitement to hatred, violence or discrimination (red). With respect to the inter-annotator agreement, only 25% of the instances are associated with the maximum confidence score of 1, while the distribution of confidence presents a high peak around .66 and a minor peak around 0.5. Note that this confidence distribution is not directly comparable to Task 1, since the schemes are different.

In Task 3, similarly to Task 2, the result of the annotation is a real value. More precisely, we

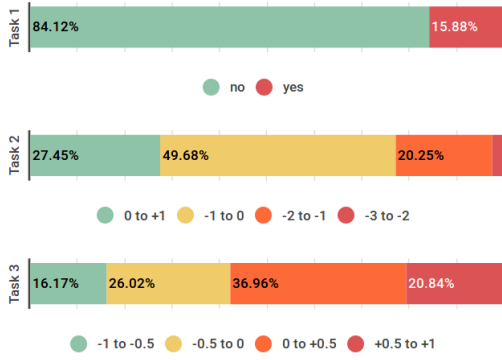


Figure 1: Label distribution for Tasks 1, 2 and 3 (red portion of Task 2 bar corresponds to 2.63%).

compute for each tweet the percentage of times it has been indicated as *best* (more indicative of HS in its tuple) and *worst* (least indicative of HS in its tuple), and compute the difference between these two values, resulting in a value between -1 (non-hateful end of the spectrum) and 1 (hateful end of the spectrum). The bottom chart in Figure 1 shows that the distribution of values given by the BWS annotation has a higher variance than the scalar case, and is skewed slightly towards the hateful side. The confidence score for Task 3 follows a similar pattern to Task 2, while being slightly higher on average, with about 40% of the tweets having confidence 1.

A last consideration concerns the cost of annotation tasks in terms of time and resources. We measured the cost of our three tasks: T1 and T2 had almost the same cost in terms of contributors retribution, but T2 required about twice the time to be completed; T3 resulted the most expensive in terms of both money and time. With nearly equal results, a strategy could be chosen instead of others for being quicker or cheaper: therefore, when designing a research strategy, we deem important not to forget this factor.

6 Classification tests with different schemes at comparison

Having described the process and results for each task, we will now observe how they affect the quality of resulting datasets. Our running hypothesis is that a better quality dataset provides better training material for a supervised classifier, thus leading to higher predictive capabilities.

Assuming that the final goal is to develop an effective system for recognizing HS, we opted to test the three schemes against the same binary classi-

fier. In order to do so, it was necessary to make our schemes comparable without losing the information each of them gives: we mapped Task 2 and Task 3 schemes down to a binary structure, directly comparable to Task 1 scheme. For Task 2, this was done by drawing an arbitrary line that would split the scale in two. We tested different thresholds, mapping the judgements above each threshold to the label *HS_no* from Task 1 and all judgements below the threshold to the label *HS_yes*. We experimented with three values: -0.5 , -1.0 and -1.5 . For Task 3, similarly, we tried setting different thresholds along the hateful end of the answers distribution spectrum (see Section 5), respectively at 0 , 0.25 , 0.5 and 0.75 . We mapped all judgements below each threshold to the label *HS_no* from Task 1 and all judgements above the threshold to the label *HS_yes*.

When considering as *HS_yes* all tweets whose average value for Task 2 is above 0.5 , the number of hateful tweets increases (25.35%); when the value is set at -1.0 , slightly decreases (10.22%); but as soon as the threshold is moved up to -1.5 , the number drops dramatically. A possible explanation for this is that a binary scheme is not adequate to depict the complexity of HS and forces judges to squeeze contents into a narrow black-or-white frame. Conversely, thresholds for Task 3 return different results (however partial). The threshold 0.5 is the closest to the Task 1 partition, with a similar percentage of HS (16.90%), while lower thresholds allow for much higher percentages of tweets classified as hateful — setting the value at 0 , for example, results in 40.52% of tweets classified as HS.

To better understand the impact of the different annotation strategies on the quality of the resulting datasets, we performed a cross-validation experiment. We implemented a SVM classifier using n -grams ($1 \leq N \leq 4$) as features and measuring its precision, recall and F1 score in a stratified 10-fold fashion. Results are shown in Table 6.

From the results of this cross-validation experiment, we draw some observations. When mapping the non-binary classification to a binary one, choosing an appropriate threshold has a key impact on the classifier performance. For both RS and BWS, the strictness of the threshold (i.e., how close it is to the hateful end of the spectrum) is directly proportional to the performance on the negative class (0) and inversely proportional to the

Dataset	Threshold	support (0)	support (1)	P (0)	R (0)	F1 (0)	P (1)	R (1)	F1 (1)	F1 (macro)
binary		3365	635	.878	.923	.899	.450	.316	.354	.627
RS	-0.5	2976	1014	.785	.841	.812	.408	.322	.359	.585
RS	-1.0	3581	409	.912	.966	.938	.391	.186	.250	.594
RS	-1.5	3845	145	.964	.991	.978	.200	.028	.047	.512
BWS	0.0	2206	1782	.677	.703	.690	.614	.585	.599	.644
BWS	0.25	2968	1020	.806	.860	.832	.492	.398	.439	.635
BWS	0.5	3480	508	.893	.949	.920	.390	.222	.281	.601
BWS	0.75	3835	153	.963	.992	.977	.147	.039	.060	.518

Table 6: Result of 10-fold cross-validation on datasets obtained with different annotation strategies.

performance on the positive class (1). This may be explained by different amounts of training data available: as we set a stricter threshold, we will have fewer examples for the positive class, resulting in a poorer performance, but more examples for the negative class, resulting in a more accurate classification. Yet, looking at the rightmost column, we observe how permissive thresholds return a higher overall F1-score for both RS and BWS.

Regardless of the threshold, RS appears to produce the worst performance, suggesting that reducing continuous values to crisp labels is not the best way to model the phenomenon, however accurate and pondered the labels are. Conversely, compared to the binary annotation, BWS returns higher F1-scores with permissive threshold (0.0 and 0.25), thus resulting in the best method to obtain a stable dataset. Furthermore, performances with BWS are consistently higher for the positive class (HS): considering that the task is typically framed as a *detection* task (as opposed to a *classification* task, this result confirms the potential of ranking annotation (as opposed to rating) to generate better training material for HS detection.

7 Conclusion and Future Work

We performed annotation tasks with three annotation schemes on a HS corpus, and computed inter-annotator agreement rate and label distribution for each task. We also performed cross-validation tests with the three annotated datasets, to verify the impact of the annotation schemes on the quality of the produced data.

We observed that the RS we designed seems easier to use for contributors, but its results are more complex to understand, and it returns the worst overall performance in a cross-validation test. It is especially difficult to compare it with a binary scheme, since merging labels together and mapping them down to a dichotomous choice is in contrast with the nature of the scheme itself.

Furthermore, such scale necessarily oversimplifies a complex natural phenomenon, because it uses equidistant points to represent shades of meaning that may not be as evenly arranged.

Conversely, our experiment with BWS applied to HS annotation gave encouraging results. Unlike Wojatzki et al. (2018), we find that a ranking scheme is slightly better than a rating scheme, be it binary or scalar, in terms of prediction performance. As future work, we plan to investigate the extent to which such variations depend on circumstantial factors, such as how the annotation process is designed and carried out, as opposed to intrinsic properties of the annotation procedure.

The fact that similar distributions are observed when the dividing line for RS and BWS is drawn in a permissive fashion suggests that annotators tend to overuse the label *HS.yes* when they work with a binary scheme, probably because they have no milder choice. This confirms that, whatever framework is used, the issue of hateful language requires a nuanced approach that goes beyond the binary classification, being aware that an increase in complexity and resources will likely pay off in terms of more accurate and stable performances.

Acknowledgments

The work of V. Basile, C. Bosco, V. Patti is partially funded by Progetto di Ateneo/CSP 2016 *Immigrants, Hate and Prejudice in Social Media* (S1618.L2.BOSC.01) and by Italian Ministry of Labor (*Contro l'odio: tecnologie informatiche, percorsi formativi e storytelling partecipativo per combattere l'intolleranza*, avviso n.1/2017 per il finanziamento di iniziative e progetti di rilevanza nazionale ai sensi dell'art. 72 del decreto legislativo 3 luglio 2017, n. 117 - anno 2017). The work of F. Poletto is funded by Fondazione Giovanni Gorla and Fondazione CRT (*Bando Talenti della Società Civile 2018*).

References

- Ika Alfina, Rio Mulia, Mohamad Ivan Fanany, and Yudo Ekanata. 2017. Hate speech detection in the indonesian language: A dataset and preliminary study. In *2017 International Conference on Advanced Computer Science and Information Systems (ICACSIS)*, pages 233–238. IEEE.
- Pete Burnap and Matthew L. Williams. 2015. Cyber Hate Speech on Twitter: An Application of Machine Classification and Statistical Modeling for Policy and Decision Making. *Policy & Internet*, 7(2):223–242.
- Thomas Davidson, Dana Warmley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Eleventh International AAAI Conference on Web and Social Media*, pages 368 – 371.
- Fabio Del Vigna, Andrea Cimino, Felice Dell’Orletta, Marinella Petrocchi, and Maurizio Tesconi. 2017. Hate Me, Hate Me Not: Hate Speech Detection on Facebook. In *Proceedings of the First Italian Conference on Cybersecurity (ITASEC17)*, pages 86–95.
- Darja Fišer, Tomaž Erjavec, and Nikola Ljubešić. 2017. Legal framework, dataset and annotation schema for socially unacceptable online discourse practices in slovene. In *Proceedings of the first workshop on abusive language online*, pages 46–51.
- Lei Gao, Alexis Kuppersmith, and Ruihong Huang. 2017. Recognizing explicit and implicit hate speech using a weakly supervised two-path bootstrapping approach. *arXiv preprint arXiv:1710.07394*.
- Njagi Dennis Gitari, Zhang Zuping, Hanyurwimfura Damien, and Jun Long. 2015. A lexicon-based approach for hate speech detection. *International Journal of Multimedia and Ubiquitous Engineering*, 10(4):215–230.
- Hugo Lewi Hammer. 2016. Automatic detection of hateful comments in online discussion. In *International Conference on Industrial Networks and Intelligent Systems*, pages 164–173. Springer.
- Svetlana Kiritchenko and Saif Mohammad. 2017. Best-worst scaling more reliable than rating scales: A case study on sentiment intensity annotation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 465–470. ACL.
- Rensis Likert. 1932. A technique for the measurement of attitudes. *Archives of psychology*, 22(140).
- Jordan J Louviere and George G Woodworth. 1991. Best-worst scaling: A model for the largest difference judgments. *University of Alberta: Working Paper*.
- Puneet Mathur, Rajiv Shah, Ramit Sawhney, and Debanjan Mahata. 2018. Detecting offensive tweets in hindi-english code-switched language. In *Proceedings of the Sixth International Workshop on Natural Language Processing for Social Media*, pages 18–26.
- Saif Mohammad and Svetlana Kiritchenko. 2018. Understanding emotions: A dataset of tweets to study interactions between affect categories. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*, pages 198–209.
- Hamdy Mubarak, Kareem Darwish, and Walid Magdy. 2017. Abusive language detection on arabic social media. In *Proceedings of the First Workshop on Abusive Language Online*, pages 52–56.
- Rishab Nithyanand, Brian Schaffner, and Phillipa Gill. 2017. Measuring offensive speech in online political discourse. In *7th {USENIX} Workshop on Free and Open Communications on the Internet ({FOCI} 17)*.
- Chikashi Nobata, Joel Tetreault, Achint Thomas, Yashar Mehdad, and Yi Chang. 2016. Abusive language detection in online user content. In *Proceedings of the 25th international conference on world wide web*, pages 145–153. International World Wide Web Conferences Steering Committee.
- Alexandra Olteanu, Carlos Castillo, Jeremy Boy, and Kush R Varshney. 2018. The effect of extremist violence on hateful speech online. In *Twelfth International AAAI Conference on Web and Social Media*, pages 221–230.
- Fabio Poletto, Marco Stranisci, Manuela Sanguinetti, Viviana Patti, and Cristina Bosco. 2017. Hate Speech Annotation: Analysis of an Italian Twitter Corpus. In *Proceedings of the Fourth Italian Conference on Computational Linguistics (CLiC-it 2017)*. CEUR.
- Björn Ross, Michael Rist, Guillermo Carbonell, Benjamin Cabrera, Nils Kurowsky, and Michael Wojatzki. 2017. Measuring the reliability of hate speech annotations: The case of the European refugee crisis. *arXiv preprint arXiv:1701.08118*.
- Manuela Sanguinetti, Fabio Poletto, Cristina Bosco, Viviana Patti, and Marco Stranisci. 2018. An Italian Twitter Corpus of Hate Speech against Immigrants. In *Proceedings of the 11th Language Resources and Evaluation Conference 2018*, pages 2798–2805.
- Zeera Waseem. 2016. Are you a racist or am i seeing things? annotator influence on hate speech detection on twitter. In *Proceedings of the first workshop on NLP and computational social science*, pages 138–142.
- Michael Wojatzki, Tobias Horsmann, Darina Gold, and Torsten Zesch. 2018. Do Women Perceive Hate Differently: Examining the Relationship Between Hate Speech, Gender, and Agreement Judgments. In *Proceedings of the Conference on Natural Language Processing (KONVENS)*, pages 110–120, Vienna, Austria.

Georgios Yannakakis, Roddy Cowie, and Carlos Busso. 2018. The ordinal nature of emotions: An emerging approach. *IEEE Transactions on Affective Computing*, pages 1–20, 11. Early Access.

ALBERTo: Italian BERT Language Understanding Model for NLP Challenging Tasks Based on Tweets

Marco Polignano

University of Bari A. Moro
Dept. Computer Science
E.Orabona 4, Italy

marco.polignano@uniba.it

Pierpaolo Basile

University of Bari A. Moro
Dept. Computer Science
E.Orabona 4, Italy

pierpaolo.basile@uniba.it

Marco de Gemmis

University of Bari A. Moro
Dept. Computer Science
E.Orabona 4, Italy

marco.degemmis@uniba.it

Giovanni Semeraro

University of Bari A. Moro
Dept. Computer Science
E.Orabona 4, Italy

giovanni.semeraro@uniba.it

Valerio Basile

University of Turin
Dept. Computer Science
Via Verdi 8, Italy

valerio.basile@unito.it

Abstract

English. Recent scientific studies on natural language processing (NLP) report the outstanding effectiveness observed in the use of context-dependent and task-free language understanding models such as ELMo, GPT, and BERT. Specifically, they have proved to achieve state of the art performance in numerous complex NLP tasks such as question answering and sentiment analysis in the English language. Following the great popularity and effectiveness that these models are gaining in the scientific community, we trained a BERT language understanding model for the Italian language (**ALBERTo**). In particular, **ALBERTo** is focused on the language used in social networks, specifically on Twitter. To demonstrate its robustness, we evaluated **ALBERTo** on the EVALITA 2016 task SENTIPOLC (SENTiment POLarity Classification) obtaining state of the art results in subjectivity, polarity and irony detection on Italian tweets. The pre-trained **ALBERTo** model will be publicly distributed through the GitHub platform at the following web address: <https://github.com/marcopoli/ALBERTo-it> in order to facilitate future research.

1 Introduction

The recent spread of pre-trained text representation models has enabled important progress in

Natural Language Processing. In particular, numerous tasks such as part of speech tagging, question answering, machine translation, and text classification have obtained significant contributions in terms of performance through the use of distributional semantics techniques such as word embedding. Mikolov et al. (2013) notably contributed to the genesis of numerous strategies for representing terms based on the idea that semantically related terms have a similar vector representations. Such technologies as Word2Vec (Mikolov et al., 2013), Glove (Pennington et al., 2014), and FastText (Bojanowski et al., 2017) suffer from a problem that multiple concepts, associated with the same term, are not represented by different wordembedding vectors in the distributional space (context-free). New strategies such as ELMo (Peters et al., 2018), GPT/GPT-2 (Radford et al., 2019), and BERT (Devlin et al., 2019) overcome this limit by learning a language understanding model for a contextual and task-independent representation of terms. In their multilingual version, they mainly use a mix of text obtained from large corpora in different languages to build a general language model to be reused for every application in any language. As reported by the BERT documentation "the Multilingual model is somewhat worse than a single-language model. However, it is not feasible for us to train and maintain dozens of single-language model." This entails significant limitations related to the type of language learned (with respect to the document style) and the size of the vocabulary. These reasons have led us to create the equivalent of the BERT model for the Italian language and specifically on the language style used on Twitter: **ALBERTo**. This idea was supported by the intuition that many of the NLP

Copyright © 2019 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

tasks for the Italian language are carried out for the analysis of social media data, both in business and research contexts.

2 Related Work

A Task-Independent Sentence Understanding Model is based on the idea of creating a deep learning architecture, particularly an encoder and a decoder, so that the encoding level can be used in more than one NLP task. In this way, it is possible to obtain a decoding level with weights optimized for the specific task (fine-tuning). A general-purpose encoder should, therefore, be able to provide an efficient representation of the terms, their position in the sentence, context, grammatical structure of the sentence, semantics of the terms. One of the first systems able to satisfy these requirements was ELMo (Peters et al., 2018) based on a large neural network biLSTM (2 biLSTM layers with 4096 units and 512 dimension projections and a residual connection from the first to the second layer) trained for 10 epochs on the 1B WordBenchmark (Chelba et al., 2013). The goal of the network was to predict the same starting sentence in the same initial language (like an autoencoder). It has guaranteed the correct management of polysemy of terms by demonstrating its efficacy on six different NLP tasks for which it obtained state-of-the-art results: Question Answering, Textual Entailment, Semantic Role labeling, Coreference Resolution, Name Entity Extraction, and Sentiment Analysis. Following the basic idea of ELMo, another language model called GPT has been developed in order to improve the performance on the tasks included in the GLUE benchmark (Wang et al., 2018). GPT replaces the biLSTM network with a Transformer architecture (Vaswani et al., 2017). A Transformer is an encoder-decoder architecture that is mainly based on feed-forward and multi-head attention layers. Moreover, in Transformers terms are provided as input without a specific order and consequently a positional vector is added to the term embeddings. Unlike ELMo, in GPT, for each new task, the weights of all levels of the network are optimized, and the complexity of the network (in terms of parameters) remains almost constant. Moreover, during the learning phase, the network does not limit itself to work on a single sentence but it splits the text into spans to improve the predictive capacity and the general-

ization power of the network. The deep neural network used is a 12-layer decoder-only transformer with masked self-attention heads (768 dimensional states and 12 attention heads) trained for 100 epochs on the BooksCorpus dataset (Zhu et al., 2015). This strategy proved to be successful compared to the results obtained by ELMo on the same NLP tasks. BERT (Bidirectional Encoder Representations from Transformers) (Devlin et al., 2019) was developed to work with a strategy very similar to GPT. In its basic version, it is also trained on a Transformer network with 12 levels, 768 dimensional states and 12 heads of attention for a total of 110M of parameters and trained on BooksCorpus (Zhu et al., 2015) and Wikipedia English for 1M of steps. The main difference is that the learning phase is performed by scanning the span of text in both directions, from left to right and from right to left, as was already done in biLSTMs. Moreover, BERT uses a “masked language model”: during the training, random terms are masked in order to be predicted by the net. Jointly, the network is also designed to potentially learn the next span of text from the one given in input. These variations on the GPT model allow BERT to be the current state of the art language understanding model. Larger versions of BERT (BERT large) and GPT (GPT-2) have been released and are scoring better results than the normal scale models but require much more computational power. The base BERT model for English language is exactly the same used for learning the Italian Language Understanding Model (AIBERTO) but we are considering the possibility to develop a large version of it soon.

3 AIBERTO

As pointed out in the previous sections, the aim of this work is to create a linguistic resource for Italian that would follow the most recent strategies used to address NLP problems in English. It is well known that the language used on social networks is different from the formal one, also as a consequence of the presence of mentions, uncommon terms, links, and hashtags that are not present elsewhere. Moreover multiple language models in their multilingual version, are not performing well in every specific language, especially with a writing style different from that of books and encyclopedic descriptions (Polignano et al., 2019). AIBERTO aims to be the first Italian language under-

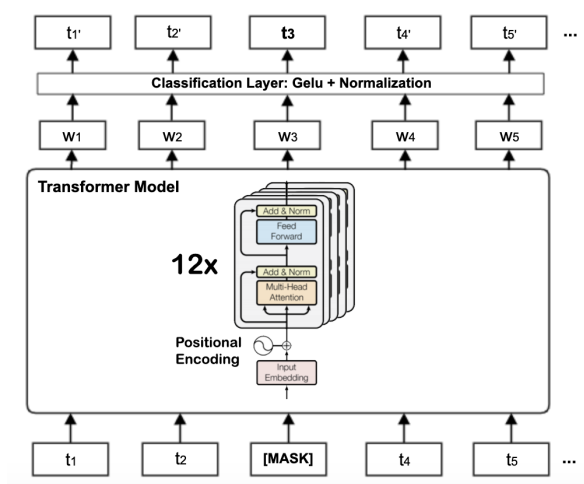


Figure 1: BERT and ALBERTo learning strategy

standing model to represent the social media language, Twitter in particular, written in Italian. The model proposed in this work is based on the software distributed through GitHub by Devlin et al. (2019)¹ with the endorsement of Google. It has been trained, without consequences, on text spans containing typical social media characters including emojis, links hashtags and mentions.

Figure 1 shows the BERT and ALBERTo strategy of learning. The “masked learning” is applied on a 12x Transformer Encoder, where, for each input, a percentage of terms is hidden and then predicted for optimizing network weights in back-propagation. In ALBERTo, we implement only the “masked learning” strategy, excluding the step based on “next following sentence”. This is a crucial aspect to be aware of because, in the case of tweets, we do not have cognition of a flow of tweets as it happens in a dialog. For this reason, we are aware that ALBERTo is not suitable for the task of question answering, where this property is essential. On the contrary, the model is well suited for classification and prediction tasks. The decision to train ALBERTo, excluding the “next following sentence” strategy, makes the model similar in purposes to ELMo. Differently from it, BERT and ALBERTo use transformer architecture instead on biLSTM which have been demonstrated to perform better in natural language processing tasks. In any case, we are considering the possibility to learn an Italian ELMo model and to compare it with the here proposed model.

¹<https://github.com/google-research/bert/>

Original tweet: #labuonascuola Eccolo, il rapporto on line qui <http://t.co/U5AXNySaJu>

Preprocessed: <hashtag> la buona scuola </hashtag> eccolo il rapporto on line qui <url>

Figure 2: Example of preprocessed Tweet

3.1 Text Preprocessing

In order to tailor the tweet text to BERT’s input structure, it is necessary to carry out preprocessing operations. More specifically, using Python as the programming language, two libraries were mainly adopted: Ekphrasis (Baziotis et al., 2017) and SentencePiece² (Kudo, 2018). Ekphrasis is a popular tool comprising an NLP pipeline for text extracted from Twitter. It has been used for:

- Normalizing URL, emails, mentions, percents, money, time, date, phone numbers, numbers, emoticons;
- Tagging and unpacking hashtags.

The normalization phase consists in replacing each term with a fixed tuple $\langle [entity\ type] \rangle$. The tagging phase consists of enclosing hashtags with two tags $\langle hashtag \rangle \dots \langle /hashtag \rangle$ representing their beginning and end in the sentence. Whenever possible, the hashtag has been unpacked into known words. The text is cleaned and made easily readable by the network by converting it to its lowercase form and all characters except emojis, !, ? and accented characters have been deleted. An example of preprocessed tweet is shown in Figure 2.

SentencePiece is a segmentation algorithm used for learning the best strategy for splitting text into terms in an unsupervised and language-independent way. It can process up to 50k sentences per seconds and generate an extensive vocabulary. It includes the most common terms in the training set and the subwords which occur in the middle of words, annotating them with ‘##’ in order to be able to encode also slang, incomplete or uncommon words. An example of a piece of the vocabulary generated for ALBERTo is shown in Figure 3. SentencePiece also produced a tokenizer, used to generate a list of tokens for each tweet further processed by BERT’s *create_pretraining_data.py* module.

²<https://github.com/google/sentencepiece>

```

[PAD] [UNK] [CLS] [SEP] [MASK]
##> < ##hashtag ##user </ ##url
! di e a che il la ##number
non ? è per anche in un della
l ma mi i grazie tutti alla
con si sono una tutto le ho
se ##👉 ##👈 ##😄 ##🙌 ##👄 ##😞
fare io da ti bene fatto italia

```

Figure 3: An extract of the vocabulary created by SentencePiece for ALBERTo

3.2 Dataset

The dataset used for the learning phase of ALBERTo is TWITA (Basile et al., 2018) a huge corpus of Tweets in the Italian language collected from February 2012 to the present day from Twitter’s official streaming API. In our configuration, we randomly selected 200 million Tweets removing re-tweets, and processed them with the pre-processing pipeline described previously. In total, we obtained 191GB of raw data.

3.3 Learning Configuration

The ALBERTo model has been trained using the following configuration:

```

bert_base_config = {
    "attention_probs_dropout_prob": 0.1,
    "directionality": "bidi",
    "hidden_act": "gelu",
    "hidden_dropout_prob": 0.1,
    "hidden_size": 768,
    "initializer_range": 0.02,
    "intermediate_size": 3072,
    "max_position_embeddings": 512,
    "num_attention_heads": 12,
    "num_hidden_layers": 12,
    "pooler_fc_size": 768,
    "pooler_num_attention_heads": 12,
    "pooler_num_fc_layers": 3,
    "pooler_size_per_head": 128,
    "pooler_type": "first_token_transform",
    "type_vocab_size": 2,
    "vocab_size": 128000
}

# Input data pipeline config
TRAIN_BATCH_SIZE = 128
MAX_PREDICTIONS = 20
MAX_SEQ_LENGTH = 128
MASKED_LM_PROB = 0.15

# Training procedure config
EVAL_BATCH_SIZE = 64
LEARNING_RATE = 2e-5
TRAIN_STEPS = 1000000
SAVE_CHECKPOINTS_STEPS = 2500
NUM_TPU_CORES = 8

```

The training has been performed over the Google Collaborative Environment (Colab)³, Using a 8 core Google TPU-V2⁴ and a Google Cloud Storage Bucket⁵. In total, it took ~ 50 hours to create a complete ALBERTo model. More technical details are available in the Notebook *"Italian Pre-training BERT from scratch with cloud TPU"* into the project repository.

4 Evaluation and Discussion of Results

We evaluate ALBERTo on a task of sentiment analysis for the Italian language. In particular, we decided to use the data released for the SENTIPOLC (SENTiment Polarity Classification) shared task (Barbieri et al., 2016) carried out at EVALITA 2016 (Basile et al., 2016) whose tweets comes from a distribution different from them used for training ALBERTo. It includes three subtasks:

- **Subjectivity Classification:** “a system must decide whether a given message is subjective or objective”;
- **Polarity Classification:** “a system must decide whether a given message is of positive, negative, neutral or mixed sentiment”;
- **Irony Detection:** “a system must decide whether a given message is ironic or not”.

Data provided for training and test are tagged with six fields containing values related to manual annotation: subj, opos, oneg, iro, lpos, lneg. These labels describe consequently if the sentence is subjective, positive, negative, ironical, literal positive, literal negative. For each of these classes, there is a 1 where the sentence satisfy the label, a 0 instead.

The last two labels “lpos” and “lneg” that describe the literal polarity of the tweet have not been considered in the current evaluation (nor in the official shared task evaluation). In total, 7410 tweets have been released for training and 2000 for testing. We do not used any validation set because we do not performed any phase of model selection during the fine-tuning of ALBERTo. The evaluation was performed considering precision (p), recall (r) and F1-score (F1) for each class and for each classification task.

³<https://colab.research.google.com>

⁴<https://cloud.google.com/tpu/>

⁵<https://cloud.google.com/storage/>

	Prec. 0	Rec. 0	F1. 0
Subjectivity	0.6838	0.8058	0.7398
Polarity Pos.	0.9262	0.8301	0.8755
Polarity Neg.	0.7537	0.9179	0.8277
Irony	0.9001	0.9853	0.9408
	Prec. 1	Rec. 1	F1. 1
Subjectivity	0.8857	0.8015	0.8415
Polarity Pos.	0.5818	0.5314	0.5554
Polarity Neg.	0.7988	0.5208	0.6305
Irony	0.6176	0.1787	0.2772

Table 1: Results obtained using the official evaluation script of SENTIPOLC 2016

<i>System</i>	<i>Obj</i>	<i>Subj</i>	<i>F</i>
AIBERTo	0.7398	0.8415	0.7906
Unitor.1.u	0.6784	0.8105	0.7444
Unitor.2.u	0.6723	0.7979	0.7351
samskara.1.c	0.6555	0.7814	0.7184
ItaliaNLP.2.c	0.6733	0.7535	0.7134

System	Pos	Neg	F
AIBERTo	0.7155	0.7291	0.7223
UniPI.2.c	0.6850	0.6426	0.6638
Unitor.1.u	0.6354	0.6885	0.6620
Unitor.2.u	0.6312	0.6838	0.6575
ItaliaNLP.1.c	0.6265	0.6743	0.6504

System	Non-Iro	Iro	F
AIBERTo	0.9408	0.2772	0.6090
tweet2check16.c	0.9115	0.1710	0.5412
CoMoDI.c	0.8993	0.1509	0.5251
tweet2check14.c	0.9166	0.1159	0.5162
IRADABE.2.c	0.9241	0.1026	0.5133

Table 2: Comparison of results with the best systems of SENTIPOLC for each classification task

AIBERTo fine-tuning. We fine-tuned AIBERTo four different times, in order to obtain one classifier for each task except for the polarity where we have two of them. In particular, we created one classifier for the Subjectivity Classification, one for Polarity Positive, one for Polarity Negative and one for the Irony Detection. Each time we have re-trained the model for three epochs, using a learning rate of $2e-5$ with 1000 steps per loops on batches of 512 example from the training set of the specific task. For the fine-tuning of the Irony Detection classifier, we increased the number of epochs of training to ten observing low performances using only three epochs as for the other classification tasks. The fine-tuning process lasted ~ 4 minutes every time.

Discussion of the results. The results reported in Table 1 show the output obtained from the official evaluation script of SENTIPOLC 2016. It is important to note that the values on the individual classes of precision, recall and, F1 are not compared with them of the systems that participated in the competition because they are not reported in the overview paper of the task. Nevertheless, some considerations can be drawn. The classifier based on AIBERTo achieves, on average, high recall on class 0 and low values on class 1. The opposite situation is instead observed on the precision, where for the class 1 it is on average superior to the recall values. This note suggests that the system is very good at classifying a phenomenon and when it does, it is sure of the prediction made even at the cost of generating false negatives.

On each of the sub-tasks of SENTIPOLC, it can be observed that AIBERTo has obtained state of the art results without any heuristic tuning of learning parameters (model as it is after fine-tuning training) except in the case of irony detection where it was necessary to increase the number of epochs of the learning phase of fine-tuning. Comparing AIBERTo with the best system of each subtask, we observe an increase in results between 7% and 11%. The results obtained are exciting, from our point of view, for further future work.

5 Conclusion

In this work, we described AIBERTo, the first Italian language understanding model based on social media writing style. The model has been trained using the official BERT source code on a Google TPU-V2 on 200M tweets in the Italian language. The pre-trained model has been fine-tuned on the data available for the classification task SENTIPOLC 2016, showing SOTA results. The results allow us to promote AIBERTo as the starting point for future research in this direction. Model repository: <https://github.com/marcopoli/AIBERTo-it>

Acknowledgment

The work of Marco Polignano is funded by project "DECISION" codice raggruppamento: BQS5153, under the Apulian INNONETWORK programme, Italy. The work of Valerio Basile is partially funded by Progetto di Ateneo/CSP 2016 (*Immigrants, Hate and Prejudice in Social Media*, S1618_L2_BOSC_01).

References

- Francesco Barbieri, Valerio Basile, Danilo Croce, Malvina Nissim, Nicole Novielli, and Viviana Patti. 2016. Overview of the evalita 2016 sentiment polarity classification task. In *Proceedings of third Italian conference on computational linguistics (CLiC-it 2016) & fifth evaluation campaign of natural language processing and speech tools for Italian. Final Workshop (EVALITA 2016)*.
- Pierpaolo Basile, Franco Cutugno, Malvina Nissim, Viviana Patti, Rachele Sprugnoli, et al. 2016. Evalita 2016: Overview of the 5th evaluation campaign of natural language processing and speech tools for Italian. In *3rd Italian Conference on Computational Linguistics, CLiC-it 2016 and 5th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian, EVALITA 2016*, volume 1749, pages 1–4. CEUR-WS.
- Valerio Basile, Mirko Lai, and Manuela Sanguinetti. 2018. Long-term social media data collection at the university of turin. In *Fifth Italian Conference on Computational Linguistics (CLiC-it 2018)*, pages 1–6. CEUR-WS.
- Christos Baziotis, Nikos Pelekis, and Christos Doulkeridis. 2017. Datastories at semeval-2017 task 4: Deep lstm with attention for message-level and topic-based sentiment analysis. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 747–754, Vancouver, Canada, August. Association for Computational Linguistics.
- Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.
- Ciprian Chelba, Tomas Mikolov, Mike Schuster, Qi Ge, Thorsten Brants, Phillipp Koehn, and Tony Robinson. 2013. One billion word benchmark for measuring progress in statistical language modeling. *arXiv preprint arXiv:1312.3005*.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June. Association for Computational Linguistics.
- Taku Kudo. 2018. Subword regularization: Improving neural network translation models with multiple subword candidates. *arXiv preprint arXiv:1804.10959*.
- Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.
- Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.
- Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. pages 2227–2237, June.
- Marco Polignano, Pierpaolo Basile, Marco de Gemmis, and Giovanni Semeraro. 2019. A comparison of word-embeddings in emotion detection from text using bilstm, cnn and self-attention. In *Adjunct Publication of the 27th Conference on User Modeling, Adaptation and Personalization*, pages 63–68. ACM.
- Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.
- Alex Wang, Amapreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2018. Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*.
- Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the IEEE international conference on computer vision*, pages 19–27.

<https://github.com/marcopoli/AIBERTO-it>



Evaluating the MuMe Dialogue System with the IDIAL Protocol

Aureliano Porporato

Università degli Studi di Torino
aureliano.porporato@unito.it

Alessandro Mazzei

Università degli Studi di Torino
alessandro.mazzei@unito.it

Rosa Meo

Università degli Studi di Torino
rosa.meo@unito.it

Daniele P. Radicioni

Università degli Studi di Torino
daniele.radicioni@unito.it

Abstract

English. In this paper we describe the implementation of the MuMe dialogue system, a task-based dialogue system for a car sharing service, and its evaluation through the IDIAL protocol. Finally we report some comments on this novel dialogue system evaluation method.¹

Italiano. *In questo lavoro descriviamo l'implementazione del sistema di dialogo MuMe, realizzato per un sistema di car sharing, e la sua valutazione attraverso il protocollo IDIAL. Infine, offriamo alcuni commenti su questo nuovo metodo per la valutazione di sistemi di dialogo.*

1 Introduction

The interest in dialogue systems is on the rise in the NLP community (McTear et al., 2016), under the strong demand for the introduction of a natural and effective user interaction in applications, like in the customer care domain (Hu et al., 2018). A related and central issue is the evaluation of such systems. In this setting, it is largely known that most evaluation metrics that come from machine translation and compare a model generated response to a single target response, exhibit a poor correlation with the human judgement (Liu et al., 2016).

In this paper we briefly illustrate a task-oriented dialogue system called MuMe (from “MUoversi MEglio”, “travelling better” in English language), and examine how far the evaluation protocol IDIAL (Cutugno et al., 2018) is helpful in its assessment. IDIAL is composed by a usability evaluation (done by a group of users) and by an evaluation of the robustness of the dialog model based

on the linguistic variations of the successful interactions with the users. The application being tested is a prototype dialogue system that we developed for the reservation of electric vehicles in the context of a car sharing service. A user must be able to interact with the system, to specify when and where s/he wants to leave and which sort of vehicle is needed. While there are some services and frameworks dedicated to the development of machine-learning-based dialogue systems, like Google Dialogflow² or the open source Rasa³ frameworks, the lack of Italian dialogue corpora in the specific domain of car sharing reservations (see, e.g., Serban et al. (2018)) and the impossibility on our part to recruit a number of people large enough for the creation of such a corpus, forced us to choose a different solution: we developed a simpler and less data-reliant rule-based system, based on slot-filling semantics. Moreover, the decisions made by this kind of systems can be tracked throughout the computation, thereby resulting in the advantage of being quite explainable. This is a desirable feature, since it simplifies the debugging and the maintenance of the routines, and allows an easier extension of the system to meet additional requirements.

This paper is mostly concerned with the evaluation of the MuMe system. The structure of the paper is as follows. After surveying on related work (Section 2), we briefly introduce the overall architecture and the main components of the MuMe dialogue system (see Section 3); we evaluate MuMe by using the IDIAL protocol, and employ MuMe experimentation as a case study for giving feedback on the IDIAL protocol itself (Section 4); finally, in the final Section we briefly recap the main contributions of the paper, and point to ongoing and future work.

¹Copyright © 2019 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

²<https://dialogflow.com/>

³<https://rasa.com/>

2 Related Work

The pioneering work of (Bobrow et al., 1977) proposed the frame-based architecture that most of task-based dialogue systems implement. The basic idea is to abandon the demanding goal to have a genuine logic representation of the dialog meaning and adopt a simpler slot-filling semantics. In some sense, the event-entities representation of the modern neural-based dialogue system frameworks can be seen as an ultimate evolution of that simplification idea. Aust et al. (1995) presented a rule-based system to some extents similar to ours in its purpose and structure, created for a train-seat reservation project. This system has to grasp the names of cities, train stations, dates and times, and it is able to perform quite sophisticated temporal information processing. Further rule-based systems are reviewed in the survey by (Abdul-Kader and Woods, 2015).

A different class of dialogue systems are based on neural networks. A survey on this class of systems can be found in (Mathur and Singh, 2018).

Regarding the evaluation of dialogue systems, the work by (Bohlin et al., 1999) proposes the Trindi Tick-list, a wish list of the desired dialogue behaviour and features specified as a checklist of "yes-no" questions. As regards this approach, Braunger and Maier (2017) argue that standardised evaluation models do not enable a complete evaluation of a dialogue system. Rather, they suggest that such evaluation must take into account the *natural flow* of the interaction between the user and the system itself; such measure involves many language- and user-dependant factors, such as the length of the user utterances. Such principles were tested in human-computer vocal interactions occurring on board of vehicles. Further information on dialogue systems evaluation methods can be found in the survey by Deriu et al. (2019).

3 The MuMe system architecture

In Figure 1 we depicted the basic architecture of the MuME dialogue system. The information flow starts from a sentence typed by the user: this sentence is handled by the OpenDial system (see Section 3.1) which plays both the role of the dialogue manager and of the system orchestrator. So, the sentence is syntactically parsed and semantically analyzed by an IE module (see Section 3.2). At this point, the result of the processing is converted

into a slot-filling form. When control returns to OpenDial, it generates an answer and returns it to the user on the basis of a dialogue control strategy (see Section 3.3).

3.1 The OpenDial Dialogue Manager

The main component of our software architecture is the OpenDial open source framework for dialogue management (Lison, 2015). The system, that was designed for speech interaction, adopts the *information state* approach for modelling the state of the dialogue (Traum and Larsson, 2003), that is a collection of variables representing the actual state of the system. The transition between states, i.e. the change of the variables values, is governed by the activation of a set of "if-then-else" rules on input values as well as on the variation of some variables. Indeed, OpenDial uses these rules when it models the sub-tasks of user utterance understanding, the dialogue management and the response generation. Moreover the integration of the system with external tools is simple. We exploited this capability in MuMe since for language understanding we used a module based on an external parser (see below). Additionally, the OpenDial framework implements some statistical-based techniques to deal with uncertainty. This is a way to learn interaction models from existing dialogues. This feature is particularly important for speech based dialogue systems where uncertain information arises from automatic speech recognition. However, at this stage of the MuMe project, we did not use this feature since we were working on written texts only.

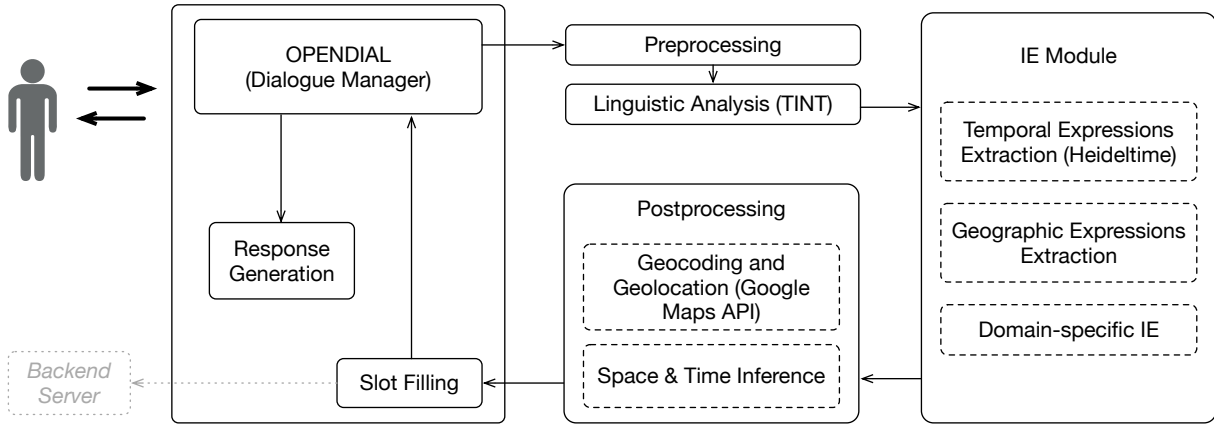
3.2 Parsing and Information Extraction

In order to assign semantic roles to the entities in the dialogues, we decided to use a syntactic parser on the text inserted by the user.

As our main parsing module we used Tint (The Italian NLP Tool) (Palmero Aprosio and Moretti, 2016), a framework modeled on Stanford CoreNLP (Manning et al., 2014). Tint performs some fundamental processing of user utterances, such as dependency parsing, Named Entities Recognition and the extraction of Temporal Expressions. In particular, the tasks are executed by interfacing with external tools.

For the recognition of temporal expressions (such as dates and times), Tint integrates the services provided by HeidelTime (Strötgen and Gertz, 2013). HeidelTime allows the extraction

Figure 1: The schematic architecture of the MuMe dialogue system.



of various sorts of temporal expressions in various languages, including the Italian language, and represents them in the standard TIMEX3 format.

For the treatment of geographic expressions, Tint is interfacing with the Nominatim wrapper.⁴ However, this (free and open source) service performs poorly in geocoding (i.e., in searching the GPS coordinates of a given address). As a consequence we decided to use the Google Maps API⁵, which provides for better performances. Indeed, Maps offers an API for address autocomplete, once this information piece has been isolated from the rest of the sentence, and for geolocation (i.e., searching the coordinates of the user), too.

3.3 Dialogue Control Strategy

The simple control strategy implemented, that governs the *moves* of the dialogue, is based on the fulfillment of a number of mandatory slots in the domain-specific slot-filling semantics adopted for the car reservation domain.

In particular, the mandatory slots are the *start date*, the *start time* and *start stall* (which encodes the start position). Indeed, the simplest reservation in MuMe needs only of these pieces of information: a person reserves a standard car, starting at a specific time of a specific day from a specific stall, and will return the car in the same stall without the need to specify the return date and time.

However, more complex reservations need more information, that are encoded in the non-mandatory slots of *end date*, *end time*, *end stall*

and *vehicle type*. For example, the user can choose between three types of vehicles, but if the kind of vehicle is not specified, the system assigns a default ‘economy car’ to the *vehicle type* slot.

The MuMe system adopts a mixed initiative for dialogue handling. Although the dialogue is overall system-driven, the user starts the conversation by possibly providing some initial information. A richer initial information is expected to result in a shorter dialogue interaction. Indeed, a design goal of the MuMe system is to produce a dialogue as short as possible. For this reason, also in the subsequent interactions, if the user gives various pieces of information in a single utterance, the system can extract all such information and is able to assign each filler to the corresponding slot, thus avoiding further unnecessary questions.

When the user begins the interaction with the MuMe system, the system replies with a welcome message, and with a general question aiming at encouraging the user to start the interaction in the most natural way.

In order to give more details on the control strategy, we consider now the following running example and its processing in MuMe (see Figure 1):

(it) “**User:** Ho bisogno di un’auto domani per andare in via Pessinetto”

(en) “**User:** I need a car tomorrow to go in Pessinetto street”⁶

The Information Extraction phase detects a date (through HeidelTime) and an address (extracted through a basic set of custom rules) in the user

⁴<http://nominatim.org/>.

⁵<https://cloud.google.com/maps-platform/>.

⁶The English version of the user and system sentences are given for clarity. The system is available in Italian language only.

sentence. By means of other rules that check the shape of the dependency tree (obtained through Tint), date and address are labelled as *start date* and *end address*. Particularly relevant in this case is the verb “andare” (“to go”), that signals that the following address is where the user wants to arrive and not a starting point. In the post-processing phase some additional information can be inferred, like the value of the *start address*, left unspecified by the user: it can be selected by retrieving the GPS coordinates of the address by means of the Google Maps API. Once the user’s current location has been identified, the nearest stall is selected as the *start stall*.

At the end of this processing, the system successfully filled the *start address*, *start stall*, *end address*, *end stall* and *start date* slots. Some mandatory slots are still left unfilled, such as the *start time*, so that the system will ask the user to provide the missing information. As a consequence, the response of the system will be a question selected from a fixed list based on unfilled slots: in this specific example, the system will continue asking for the departure time.

At the end of the filling-phase of the mandatory-slots, the systems gives the user the possibility to modify the request and to correct possible errors and misunderstandings. The slot-filling values will be sent to a dedicated server for the finalization of the reservation.

4 Evaluation

In order to have a first preliminary evaluation of the MuMe system, we applied the Trindi Tick-list protocols, that is a set of “yes-no” questions concerning specific capabilities of the developed system (Bohlin et al., 1999). While this simple questionnaire is helpful in the development phase, since it is able to give a measure of the system limits, it is not suitable to completely evaluate the actual experience of the user. At this stage of development, the MuMe system has a Trindi score of six over twelve with respect to the (original) list. Among the six features not yet implemented, there are complex tasks, such as the management of the *help* and *non-help* sub-dialogues, dealing with negative information, and dealing with noisy input.

In the rest of the Section, we report the results obtained by applying the IDIAL evaluation protocol to the current version of the MuMe system,

which is split in a questionnaire concerning the user experience (Section 4.1), and a number of *stress tests* concerning the linguistic robustness of the system (Section 4.2).

4.1 IDIAL User Evaluation

A group of 5 subjects (3 males, 2 females, 19, 22, 25, 26 and 61 years old) were recruited for the evaluation task by personal invitation and without rewards. After a brief oral description of the domain and of the basic mechanisms of interaction with the system, each user was asked to generate 7 complete dialogues with the system in a controlled environment. We asked the users to simulate the process of reserving a car without other specific constraints.

In Table 1 we report the ten questions of the IDIAL user test with the average score, obtained by using a Likert scale based on five points.⁷ Note that the questions 3, 4, 7 and 10 have been designed to evaluate the effectiveness of the dialogue system, while questions 1 and 2 regard the system efficiency.⁸

4.2 IDIAL Stress Tests

The second evaluation stage in the IDIAL protocol consists in a set of linguistic stress tests. We selected 5 dialogues (one for each user) among those successfully completed⁹ during the user evaluation stage. Following the IDIAL protocol, we modified one sentence in each dialogue, once for each test, as illustrated in (Cutugno et al., 2018), and repeated the dialogue with the modified sentence. The results are reported in Table 2.

Note that we could not perform three stress tests for distinct reasons. We could not perform the ST-8 test, regarding active-passive alternation, because the users almost always used intransitive verbs (like “andare” [“to go”] and “partire” [“to depart”]). We could not perform the ST-9 test, concerning adjective-noun alternation, since the users used a very few adjectives (like vehicle types modifiers “lussuosa” [“luxurious”]), and no adjectives have been used in a successful dialogue. Fi-

⁷We used the Italian version of the questionnaire, found in the Appendix A of <https://tinyurl.com/yxngqkx4>, but for sake of readability in Table 1 we report the English version.

⁸The answers of each subjects are available at <https://tinyurl.com/y6nruwon>

⁹We considered an interaction as ‘successfully completed’ if the system recognized and processed correctly all the data given by the user.

N	Sentence	Evaluation
1	The system was efficient in accomplishing the task.	3.2 (0.45)
2	The system quickly provided all the information that the user needed.	3.6 (0.55)
3	The system is easy to use.	3.6 (1.52)
4	The system is awkward when the user interacts with a non-standard or unexpected input.	2.8 (0.84)
5	The user is satisfied by his/her experience.	3.0 (0.00)
6	The user would recommend the system.	3.2 (0.84)
7	The system has a fluent dialogue.	2.8 (0.84)
8	The system is charming.	3.4 (0.90)
9	The user enjoyed the time s/he spent using the software.	3.8 (0.84)
10	The system is flexible to the user's needs.	3.6 (0.55)

Table 1: IDIAL user ratings of their experience: the average scores are provided on a 1-5 Likert scale with standard deviation, in parentheses.

nally, we could not perform the ST-10 test, concerning anaphora resolution, since at the actual stage of development the system never asks the user to pick an answer from a set of options.

4.3 Discussion

With respect to the user evaluation test, a number of considerations arise from scores. The main issue pointed out by the users during the evaluation phase is the difficulty in grasping when and why the system misunderstood (or lost) some pieces of information, thereby resulting in a relatively poor evaluation score for the fluency of the system (average score of 2.8). The lack of feedback due to the too simple way we used to generate system responses has even worsened this problem, leading the user to repeat the same mistake more than once. The standard deviation of the evaluations given to question 3 shows the high subjectivity of the user experiences with the system, and points out the necessity to equip the system with some form of *user model* to account for the expectation of different kinds of users. It is worth noting that

Stress Test		Passed
Spelling Substitutions		
ST-1	Confused words	60%
ST-2	Misspelled words	40%
ST-3	Character replacement	80%
ST-4	Character swapping	60%
Lexical Substitutions		
ST-5	Less frequent synonyms	60%
ST-6	Change of register	40%
ST-7	Coreference	100%
Syntactic Substitutions		
ST-8	Active-Passive alternation	—
ST-9	Nouns-adjectives inversion	—
ST-10	Anaphora resolution	0%
ST-11	Verbal-modifier inversion	80%

Table 2: IDIAL stress test results.

4 out of 5 users explicitly stated (in private conversations after the evaluation phase) that they expected longer interactions. Also, they expected to receive more questions by the system, challenging our assumption on the length of dialogues. However, two of the same users added that 7 interactions are enough to evaluate the system.

With respect to the evaluation of the stress tests, we can say that the sentences provided by the users during the interaction with the system, were often very short and scarcely usable from the viewpoint of the IDIAL stress tests (especially those concerned with lexical and syntactic aspects). Another source of problems are *typos*, in particular in expressions regarding time and addresses. While our system seems quite robust to this kind of errors (see the first 4 rows of Table 2), it is difficult to automatically deal with them without some domain specific knowledge on their occurrence and some correction strategies.

As a final note, we want to report some comments given by the users about the questionnaire. Two users expressed some doubts on the interpretation of question 8 and in general all of them found difficult to assign a meaningful evaluation to it. For example, some of the users interpreted the question as regarding the lack of a GUI, absent in our prototype. We think that the ambiguity of the sentence explains the slightly higher standard deviation for that question in respect to others. Other comments include the lack of diversity between some sentences (like questions 1 and 5, often judged as redundant), and the inade-

quacy of this Likert scale to evaluate some questions, like 5 and 9: they consider a more subjective scale (“poco” [“few”] - “molto” [“a lot”]) more appropriate, perceiving the whole process as a single experience.

While the linguistic stress test can be a valuable tool for the improvement of the system, the questionnaire concerning the user experience should be revised for addressing some critics that we collected. In particular, the questionnaire should be augmented with more specific questions.

5 Conclusion and Future Work

We presented the MuMe system, a prototype of a rule-based dialogue system and its evaluation through the IDIAL method.

Since the MuMe project is still in development, there is much room for improvement. The most pressing problem to be addressed in future development is the generation of a response more meaningful to the user. The application of a natural language generation pipeline for Italian (e.g. (Mazzei et al., 2016; Mazzei, 2016; Conte et al., 2017; Ghezzi et al., 2018)) could help to these ends.

Acknowledgments

This project has been partially supported by the MuMe Project (Muoversi Meglio), funded by the Piedmont Region and EU in the frame of the F.E.S.R. 2014/2020.

References

- [Abdul-Kader and Woods2015] Sameera A Abdul-Kader and JC Woods. 2015. Survey on chatbot design techniques in speech conversation systems. International Journal of Advanced Computer Science and Applications, 6(7).
- [Aust et al.1995] Harald Aust, Martin Oerder, Frank Seide, and Volker Steinbiss. 1995. The philips automatic train timetable information system. Speech Communication, 17(3-4):249–262.
- [Bobrow et al.1977] Daniel G. Bobrow, Ronald M. Kaplan, Martin Kay, Donald A. Norman, Henry Thompson, and Terry Winograd. 1977. Gus, a frame-driven dialog system. Artif. Intell., 8(2):155–173, April.
- [Bohlin et al.1999] Peter Bohlin, Johan Bos, Staffan Larsson, Ian Lewin, Colin Matheson, and David Milward. 1999. Survey of existing interactive systems. Deliverable D1, 3:1–23.
- [Braunger and Maier2017] Patricia Braunger and Wolfgang Maier. 2017. Natural language input for in-car spoken dialog systems: How natural is natural? In Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue, pages 137–146.
- [Conte et al.2017] Giorgia Conte, Cristina Bosco, and Alessandro Mazzei. 2017. Dealing with italian adjectives in noun phrase: a study oriented to natural language generation. In Proceedings of the Fourth Italian Conference on Computational Linguistics (CLiC-it 2017), Rome, Italy, December 11-13, 2017., December.
- [Cutugno et al.2018] Francesco Cutugno, Maria Di Maro, Sara Falcone, Marco Guerini, Bernardo Magnini, and Antonio Origlia. 2018. Overview of the evalita 2018 evaluation of italian dialogue systems (idial) task. In EVALITA@ CLiC-it.
- [Deriu et al.2019] Jan Deriu, Alvaro Rodrigo, Arantxa Otegi, Guillermo Echevoyen, Sophie Rosset, Eneko Agirre, and Mark Cieliebak. 2019. Survey on evaluation methods for dialogue systems. arXiv preprint arXiv:1905.04071.
- [Ghezzi et al.2018] Ilaria Ghezzi, Cristina Bosco, and Alessandro Mazzei. 2018. Auxiliary selection in italian intransitive verbs: A computational investigation based on annotated corpora. In Proceedings of the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018), pages 1–6, Berlin. CEUR.
- [Hu et al.2018] Tianran Hu, Anbang Xu, Zhe Liu, Quanzeng You, Yufan Guo, Vibha Sinha, Jiebo Luo, and Rama Akkiraju. 2018. Touch your heart: A tone-aware chatbot for customer care on social media. In Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, page 415. ACM.
- [Lison2015] Pierre Lison. 2015. A hybrid approach to dialogue management based on probabilistic rules. Computer Speech & Language, 34(1):232 – 255.
- [Liu et al.2016] Chia-Wei Liu, Ryan Lowe, Iulian V Serban, Michael Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How not to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. arXiv preprint arXiv:1603.08023.
- [Manning et al.2014] Christopher Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, and David McClosky. 2014. The stanford corenlp natural language processing toolkit. In Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations, pages 55–60.
- [Mathur and Singh2018] Vinayak Mathur and Arpit Singh. 2018. The rapidly changing landscape of conversational agents. arXiv preprint arXiv:1803.08419.

- [Mazzei et al.2016] Alessandro Mazzei, Cristina Battaglino, and Cristina Bosco. 2016. Simplenlg-it: adapting simplenlg to italian. In Proceedings of the 9th International Natural Language Generation conference, pages 184–192, Edinburgh, UK, September 5-8. Association for Computational Linguistics.
- [Mazzei2016] Alessandro Mazzei. 2016. Building a computational lexicon by using SQL. In Pierpaolo Basile, Anna Corazza, Francesco Cugugno, Simonetta Montemagni, Malvina Nissim, Viviana Patti, Giovanni Semeraro, and Rachele Sprugnoli, editors, Proceedings of Third Italian Conference on Computational Linguistics (CLiC-it 2016) & Fifth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2016), Napoli, Italy, December 5-7, 2016., volume 1749, pages 1–5. CEUR-WS.org, December.
- [McTear et al.2016] Michael McTear, Zoraida Callejas, and David Griol. 2016. The Conversational Interface: Talking to Smart Devices. Springer Publishing Company, Incorporated, 1st edition.
- [Palmero Aprosio and Moretti2016] A. Palmero Aprosio and G. Moretti. 2016. Italy goes to Stanford: a collection of CoreNLP modules for Italian. ArXiv e-prints, September.
- [Serban et al.2018] Iulian Vlad Serban, Ryan Lowe, Peter Henderson, Laurent Charlin, and Joelle Pineau. 2018. A survey of available corpora for building data-driven dialogue systems: The journal version. Dialogue & Discourse, 9(1):1–49.
- [Strötgen and Gertz2013] Jannik Strötgen and Michael Gertz. 2013. Multilingual and cross-domain temporal tagging. Language Resources and Evaluation, 47(2):269–298.
- [Traum and Larsson2003] David Traum and Staffan Larsson. 2003. The Information State Approach to Dialogue Management. In Current and New Directions in Discourse and Dialogue, pages 325–353. Springer.

To be Fair: a Case for Cognitively-Inspired Models of Meaning

Simon Preissner

Center for Mind/Brain Sciences
University of Trento

simon.preissner@gmx.de

Aurélie Herbelot

Center for Mind/Brain Sciences &
Dept. of Information Engineering
and Computer Science

University of Trento

aurelie.herbelot@unitn.it

Abstract

In the last years, the cost of Natural Language Processing algorithms has become more and more evident. That cost has many facets, including training times, storage, replicability, interpretability, equality of access to experimental paradigms, and even environmental impact. In this paper, we review the requirements of a ‘good’ model and argue that a move is needed towards lightweight and interpretable implementations, which promote scientific fairness, paradigmatic diversity, and ultimately foster applications available to all, regardless of financial prosperity. We propose that the community still has much to learn from cognitively-inspired algorithms, which often show extreme efficiency and can ‘run’ on very simple organisms. As a case study, we investigate the fruit fly’s olfactory system as a distributional semantics model. We show that, even in its rawest form, it provides many of the features that we might require from an ideal model of meaning acquisition.¹

1 Introduction

In recent years, the Natural Language Processing (NLP) community has seen an increase in the popularity of expensive models requiring enormous computational resources to train and run. The cost of such models is multi-faceted. From the point of view of shaping the scientific community, they create a huge gap between researchers in wealthy institutions and those with less resources and they often make replication prohibitive. From the point of view of applicability, they make the end-user dependent on high-tech hardware which they may not afford, or on cloud services which may have problematic privacy side-effects (and

are not available to those with poor Internet access). Training such models can often take a long time and extraordinary amounts of energy, generating CO₂ emissions disproportionate to the models’ improvements (Strubell et al., 2019). From a pure modelling point of view, finally, complexity often comes with a loss of interpretability, which weakens theoretical insights. Whilst we appreciate that a part of NLP is focused on engineering applications rather than modelling natural language proper, the linguists and cognitive scientists in the community have a duty to provide transparent, explanatory simulations of particular phenomena.

Such considerations call for smaller and more interpretable systems. In this paper, we offer an example investigation into one of the most widely used techniques in NLP: the vectorial representation of word meanings. Our starting point is the set of requirements that should be fulfilled by an ideal model of lexical acquisition, which is expressed in QasemiZadeh et al. (2017): (A) high performance on fundamental lexical tasks, (B) efficiency, (C) low dimensionality for compact storage, (D) amenability to incremental learning, (E) interpretability. As we will show in §2, state-of-the-art systems still fail to integrate all those points. (A-D) are however basic features of humans and animal cognition. It seems, therefore, that we should find inspiration in algorithms from cognitive science, which in turn would allow us to derive interpretability (E) from the clear underpinnings of biological or psychological theories.

We propose that a good place to find appropriate algorithms is the natural world, as many organisms display core cognitive abilities such as incremental learning, generalization or classification, which many NLP systems need to emulate. Such faculties develop in extremely simple systems, which are good contenders for the type of models we advocate here. One success

¹Copyright ©2019 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

story from ‘algorithmic’ cognitive science is based on the neural architecture of the fruit fly’s olfactory system, which clusters patterns of chemicals into categories of smells (Stevens, 2015), and has inspired the so-called *Fruit Fly Optimization Algorithm* (Pan, 2011; here: Fruit Fly Algorithm or ‘FFA’). The FFA has been implemented as a lightweight neural algorithm that performs random indexing for locality-sensitive hashing (LSH) (Dasgupta et al., 2017). This LSH algorithm has successfully been applied to various tasks, particularly in information retrieval and for data compression (Andoni and Indyk, 2008). As a simple LSH algorithm, the FFA compresses data while preserving the notion of similarity of the original data, which is one of the core mechanisms involved in constructing vector representations of word meaning. To our knowledge, it has however never been taken as the basis for building distributional semantic models from scratch, even though it seems to naturally fulfill a number of requirements of those models.

In the following, we present the FFA and show how it can be adapted to create vector spaces of word meaning (§4). We then apply the FFA in an incremental setup (§5) and assess its worth as a *model*, according to the various criteria highlighted above (§6), including a possible interpretation of the FFA’s output.

2 Related work

In Distributional Semantics (DS: Turney and Pantel, 2010; Erk, 2012), the meaning of words is represented by points in a multidimensional space, derived from word co-occurrence statistics. The quality of models usually correlates with the amount of data that is used. With increasing processing resources and larger corpora available, a variety of approaches have been developed in that area (e.g., Bengio et al., 2003; Pennington et al., 2014; Mikolov et al., 2013). State-of-the-art models perform remarkably well and are often a core component of NLP applications. Recent work on DS (e.g., ELMo (Peters et al., 2018) and BERT (Devlin et al., 2018) shifts the scope of representations from word meaning to sentence meaning, pushing performance, but also model complexity, even further.

The latest DS techniques yield high performance, but they have multiple shortcomings. First, they require massive amounts of text, followed

by computationally intensive procedures involving weighting, dimensionality reduction, complex attention mechanisms etc. The high complexity of most current architectures often comes at the cost of flexibility: once a language model is constructed, any new data requires a re-run of the complete system in order to be incorporated. This makes incrementality unsatisfiable in those frameworks (Sahlgren, 2005; Baroni et al., 2007). Further, architectures themselves have become increasingly complex, at the expense of transparency. We recall that even Word2Vec (W2V: Mikolov et al., 2013), which is a comparatively simple system by today’s standards, has attracted a large amount of literature which attempts to explain the effects of various hyperparameters in the model (Levy and Goldberg, 2014; Levy et al., 2015; Gittens et al., 2017). Finally, high-performance DS representations are hardly or not at all interpretable. As a result, much research has been dedicated to producing representations that are intuitively interpretable by humans (Murphy et al., 2012; Luo et al., 2015; Fyshe et al., 2015; Shin et al., 2018). These approaches typically attempt to preserve or reconstruct word labels for the basis of the dimensionality-reduced representations, but they can themselves require intensive procedures. In summary, it becomes apparent that the ideal vector-based semantics model that fulfills all requirements highlighted in our introduction has not yet been found.

The Fruit Fly Algorithm we present here can be related to two existing techniques in computer science: Random Indexing and Locality-Sensitive Hashing. Random Indexing (RI) is a simple and efficient method for dimensionality reduction (cf. Sahlgren, 2005), originally used to solve clustering problems (Kaski, 1998). It is also a less-travelled technique in distributional semantics (Kanerva et al., 2000; QasemiZadeh et al., 2017; QasemiZadeh and Kallmeyer, 2016). Its advocates argue that it fulfills a number of requirements of an ideal vector space construction method, in particular incrementality. As for Locality-Sensitive Hashing (LSH: Slaney and Casey, 2008), it is a way to produce hashes that preserve a notion of distance between points in a space, thus satisfying storage efficiency whilst maintaining the spatial configuration of a representation. A comparison of various hash functions for LSH, including RI, is provided by Paulevé

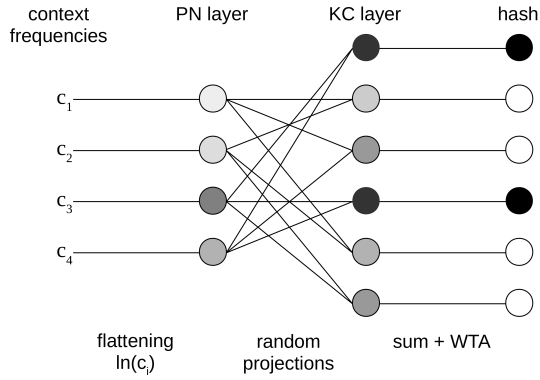


Figure 1: Schematic of the adapted FFA, with input size $m = 4$ and output size $n = 6$ (dense representation: 2). Darker cells correspond to higher activation.

et al. (2010).

3 Data

In the spirit of ‘training small’, the corpus used for our experiments is a subset of 100M words from the ukWaC corpus (Ferraresi et al., 2008), minimally pre-processed (tokenized and stripped of punctuation signs); this results in a corpus of 87.8M words. Following common practice, we quantitatively evaluate the FFA as a lexical acquisition algorithm by testing it over the MEN similarity dataset (Bruni et al., 2014), which consists of 3000 word pairs (751 unique English words), human-annotated for semantic relatedness.

For our experiments, we compute two co-occurrence count spaces over our corpus, with different context sizes (± 2 and ± 5 around the target). We only consider the 10k most frequent words in the data, ensuring we cover all 751 words in MEN.

4 Model

The Fruitfly Algorithm mimics the olfactory system of the fruit fly, which assigns a pattern of binary activations to a particular smell (i.e., a combination of multiple chemicals), using sparse connections between just two neuronal layers. This mechanism allows the fly to ‘conceptualize’ its environment and to appropriately react to new smells by relating them to previous experiences. Our implementation of the FFA is an extension of the work of Dasgupta et al. (2017) which allows us to generate a semantic space by hashing each word – as represented by its co-occurrences in a corpus – to a pattern of binary activations.

As in the original implementation, our FFA is a simple feedforward architecture consisting of two layers connected by random projections (Fig. 1). The input layer, the *projection neuron layer* or *PN layer*, consists of m nodes $\{x_1 \dots x_m\}$ which encode the raw co-occurrence counts of a target word with a particular context. To satisfy incrementality, m is variable and can grow as the algorithm encounters new data. If a new context is observed, then a node x_{m+1} is recruited to encode that context. A logarithmic function is applied to the input in order to diminish frequency effects of natural languages (Zipf, 1932). This ‘flattens’ activation across the PN layer, reducing the impact of very frequent words (e.g., stopwords). The second layer (*Kenyon Cell layer* or *KC layer*) consists of n nodes $\{y_1 \dots y_n\}$. It is larger than the PN layer and fixed at a constant size (n does not grow). PN and KC are *not* fully connected. Instead, each KC cell receives a constant number of connections from the PN layer, randomly and uniformly allocated. In other words, the mapping from *PN* to *KC* is a bipartite connection matrix M so that $M_{ji} = 1$ if x_i is connected to y_j and 0 otherwise. The connectivity of each PN is thus variable, albeit uniformly distributed. The activation function on each KC is simply the sum of the activations of its connected PNs. In the end, hashing is carried out via a winner-takes-all (WTA) procedure that ‘remembers’ the IDs of a small percentage of the most activated KCs as a compact representation of the word’s meaning. So $WTA(y_i) = 1$ if y_i is one of the k top values in y and 0 otherwise.

The FFA’s hyperparameters are expressed as a 5-tuple (f, m, n, c, h) , where f is the flattening function, m is the size of the PN layer (initially 0), n is the size of the KC layer, c is the number of connections leading to any one KC, and h is the percentage of activated KCs to be hashed.

Note that, since both the connectivity per KC and the size of the KC layer are constant, the overall number of connections is constant. Thus, the expansion mechanism (which increments m) does not create new connections: it randomly selects existing PNs and reallocates connections from those PNs to the new PN. In the reallocation process, we encode a bias towards taking connections from those PNs with the most outgoing connections in order to ensure even connectivity of the PN layer. For example, in a setup with parameters $(f = \ln, m = 300, n = 10000, c = 14, h = 8)$,

the average number of connections going out from each PN is $(n \times c)/m = 466.67$: some PNs have 466 connections, some have 467 or more. The next newly encountered word will lead to the creation of x_{301} and the expansion process will reallocate $\lfloor (n \times c)/301 \rfloor = 465$ already existing connections to x_{301} . For this, it will choose PNs with 467 or more connections with a higher probability than those with 466 connections. The parameters after the expansion process are ($f = ln, m = 301, n = 10000, c = 14, h = 8$).

The expansion of dimensions from the PN layer to the KC layer in combination with random projections can be interpreted as a form of ‘zooming’ into a concept for a particular target word: multiple context words are randomly projected onto a single KC. If several of these context words are important for the target (i.e., their PNs have high activation), the corresponding KC will be activated in the final hash. We can imagine this process as aggregating dimensions of the original co-occurrence space, thus generating latent features which give different ‘views’ into the raw data. For example, one might imagine that a random projection from the PNs *beak*, *bill*, *bank*, *wing*, and *feather*, have one KC in common. This KC might be somewhat activated by the PNs *bank* and *bill* in finance contexts, but more crucially, it will consistently be strongly activated for target words related to birds and thus selected for the final hashes of those words. Note that this behaviour lets us backtrack from a dimensionality-reduced representation to the most characteristic contexts for a particular target word, and gives interpretability to the KCs. We will come back to that feature in §6.

5 Experiments and results

In order to characterize the behavior and performance of our incremental FFA, we evaluate the quality of its output vectors against the MEN test set by means of the non-parametric Spearman rank correlation ρ . In order to run the experiments with a sound configuration of the hyperparameters f , n , c , and h , we first perform a grid search, applying various configurations of the FFA to the counts (window size: ± 5) of the 10k most frequent words of a held-out corpus.² For this setting, the grid search yields the following optimal configuration:

²we restricted the grid search and the subsequent experiment setup to a vocabulary of 10k words for more convenient experimentation. The actual FFA potentially has no such limit

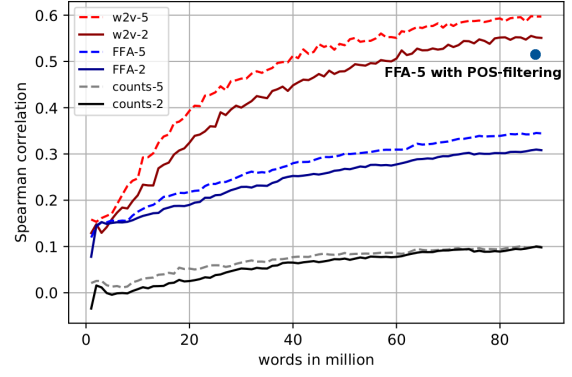


Figure 2: ρ -values of co-occurrence counts, hashed spaces, and Word2Vec models (window sizes ± 2 (lines) and ± 5 (dotted)). The blue dot shows the performance on POS-tagged data with FFA-5.

($f = ln, n = 40000, c = 20, h = 0.08$); we use this for all further experiments.³ (The grid search revealed in fact that the factor of expansion $\frac{n}{m}$ is minimally important.)

Next, we *incrementally* generate a raw frequency-count model of the 10k most frequent words of our corpus, parallelly expanding the FFA with every newly encountered word. Every 1M processed words, the aggregated co-occurrences are hashed by the FFA and the corresponding word vectors (i.e., binary hashes) are stored for evaluation. We compare a) the raw frequency space (input to the FFA); b) the final hashes (output of the FFA); c) a separate Word2Vec (W2V) model trained on exactly the same data, using standard hyperparameters and a minimum count set to match the 10k target words of our co-occurrence space. We repeat this experiment for window sizes ± 2 and ± 5 .

Figure 2 shows the results of our incremental simulation. For the window size ± 5 , we reach $\rho = 0.100$ for raw counts, $\rho = 0.345$ for the FFA output, and $\rho = 0.600$ for W2V. The 2-word-context setup yields very similar results. The FFA hashing thus has a clear and positive effect (+0.245 from 80M words on for the ± 5 setup). The amount of improvement is already large at the beginning of training (+0.136 at 5M words) and slowly increases with corpus size. Results are comparable to W2V for very small corpus sizes, but start lagging behind after around 10M words.

³The source code of this implementation of the FFA will be released for public use on [git@github.com:SimonPreissner/semantic-fruitfly.git](https://github.com/SimonPreissner/semantic-fruitfly)

6 Discussion

Investigating cognitive algorithm from scratch requires a clear stance on evaluation: we cannot expect a very simple model to beat the performance of heavily-trained systems, but we can require it to give satisfactory results whilst also being a good *model* in the strong sense of the term, that is, simulating all observable features of a given real-world phenomenon. Our discussion keeps this in mind, as we focus on the ‘wish list’ highlighted in §1.

Performance: hashing increases performance over the raw co-occurrence space by over 20 points overall. The system is however outperformed by W2V after seeing around 10M words. In the spirit of providing a comprehensive evaluation of the modelling power of the FFA, we attempt to pull apart aspects of the learning process that are captured by its very simple algorithm, and those that are not. In other words, which feature results in the large increase over baseline performance? What does the FFA fail to model with respect to W2V? We know that the algorithm generates latent features out of the original space dimensions, encapsulated in each KC. We have tuned the size of the KC layer, so the number of features captured by the FFA should be optimal for our task. We assume that the performance displayed by the algorithm is due to correctly generalizing over contexts. As for its *lack* of performance, we can make hypotheses based on what we know from other DS models. The FFA does not perform any subsampling or weighting of its input data, and the log function we use to minimize the impact of very frequent items is probably too crude to fulfill that purpose. When we informally inspect the performance of the algorithm on a POS-tagged version of our corpus, keeping only verbs, nouns and adjectives in the input and filtering some highly frequent stopwords (punctuation, auxiliaries), we obtain $\rho \approx 0.51$ over the whole corpus,⁴ coming close to W2V’s performance and thus indicating that indeed, a higher-level ‘attention’ mechanism could be added to the input layer. (Note that the olfactory system of actual fruit flies only has ≈ 50 odorant receptors, which makes it potentially less crucial to successfully suppress large parts of the input.)

Dimensionality: The size of the hashes produced by the FFA is variable; in the experiments, it

was set to 3200,⁵ which is much larger than the optimal 300-400 dimensions of W2V. However, the hash corresponds to a sparse vector of *integers* and is thus efficiently stored and manipulated. The hyperparameter grid search revealed that the factor of expansion from PN layer to KC layer is much less important than expected, although the expansion is a core characteristic of the FFA and intuitively, its factor should have an effect on performance. This suggests that the FFA does not require inconveniently high-dimensional hash signatures to reach its performance. However, it will take further experiments, especially with larger vocabularies, to fully characterize this behaviour.

Incrementality: the FFA is fully incremental. Note that in our experiments, the W2V space is retrained from scratch after each addition of 1M words to the corpus while the FFA simply increments counts in its stored co-occurrence space. It is also in stark contrast with weighted count-based distributional models which require some global PMI (re-)computation to outperform the raw co-occurrence count vectors.

Time efficiency: our FFA runs without costly learning mechanisms; its two most costly operations are (1) the expansion of the PN layer along with new vocabulary and (2) the projection from PN layer to KC layer. Following Zipf’s Law, most new words are encountered within the first few millions of words. As a consequence, the frequency of expansion operations on the PN layer is high at first, but decreases rapidly, resulting in fast scaling to large amounts of text. Hashing is solely dependent on the number of connections per KC and the size of the KC layer (both constant).

Interpretability: the FFA’s two-layer architecture allows for uncomplicated backtracking. Each of the activated nodes in a word’s hash represents a single KC. The connections of these ‘winner’ KCs with the PN layer let us reconstruct which context words originally contributed to the largest activations in the KC layer. To illustrate this, we use the hashes obtained at the last iteration of our incremental experiment (based on window ± 5) and identify the $k = 50$ most characteristic PNs for each hash, ignoring stopwords. Table 1 reports the characteristic PNs shared by various sets of input words. For example, for the words *hawk*, *pi-*

⁴We use the top 4000 dimensions of the co-occurrence matrix, with $n = 16000$, $c = 20$ and $h = 0.08$.

⁵This results from expressing the ($n=40k$ -dimensional) binary vector as the positions of its 1s, which make up $h = 8\%$ of the vector. This yields a much smaller representation of length $n \times h = 3200$.

Hashed Words	Mutual Important Words
hawk, pigeon, parrot	tailed, breasted, black, red, dove
library, collection, museum	collection, national, new, art
beard, wig	man, wearing, long, like, hair
cold, dirty	get, said, war, mind

Table 1: Top PNs for selected sets of words. The importance of a PN for a word is estimated by the number of connections to KCs that are activated in the word’s hash (window size ± 5).

geon, and *parrot* the *tailed*, *black*, *breasted*, *red*, and *dove* PNs are among the most influential, contributing to many of the activated KCs. Similarly, we can connect *beard* to *wig* and *cold* to *dirty*; the shared important words of the latter seem to encode shared collocates (*cold/dirty war*, *cold/dirty mind*, *get cold/dirty*).

7 Conclusion

We started this paper suggesting that NLP should explore a different class of algorithms for its most fundamental tasks. We argued that it is worth investigating cognitively-inspired architectures, which may not (yet) perform at state-of-the-art level, but give us insights into potentially more plausible ways to model linguistic faculties in the mind. We also made a case for ‘small’ and ‘fair’ systems, in reach of all researchers and end-users.

As illustration, we have explored what the olfactory system of a fruit fly can do for the representation of word meanings. The algorithm is certainly ‘fair’ in terms of complexity and required resources. Being based on an actual cognitive mechanism, it naturally encodes requirements such as (processing and storage) efficiency. Its simplicity lends itself to incremental learning and interpretability. Performance on a relatedness data set highlights that the raw model successfully captures latent concepts in the data but would probably require an extra attention layer, as indicated by the stronger results obtained on additionally pre-processed data.

We hope to have demonstrated that such study is accessible to all, and actually sheds insights into the minimal components of a model in a way that more complex systems do not achieve. We particularly draw attention to the fact that the inter-

esting behaviour of the fruit fly with respect to interpretability and incrementality makes it a worthy competitor for other distributional models – or at the very least, a source of inspiration.

References

- Alexandr Andoni and Piotr Indyk. 2008. Near-optimal hashing algorithms for approximate nearest neighbor in high dimensions. *Communications of the ACM*, 51(1):117.
- Marco Baroni, Alessandro Lenci, and Luca Onnis. 2007. Isa meets lara: An incremental word space model for cognitively plausible simulations of semantic learning. In *Proceedings of the workshop on cognitive aspects of computational language acquisition*, pages 49–56.
- Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. 2003. A neural probabilistic language model. *Journal of machine learning research*, 3(Feb):1137–1155.
- Elia Bruni, Nam-Khanh Tran, and Marco Baroni. 2014. Multimodal distributional semantics. *Journal of Artificial Intelligence Research*, 49:1–47.
- Sanjoy Dasgupta, Charles F Stevens, and Saket Navlakha. 2017. A neural algorithm for a fundamental computing problem. *Science*, 358(6364):793–796.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
- Katrin Erk. 2012. Vector space models of word meaning and phrase meaning: A survey. *Language and Linguistics Compass*, 6(10):635–653.
- Adriano Ferraresi, Eros Zanchetta, Marco Baroni, and Silvia Bernardini. 2008. Introducing and evaluating ukwac, a very large web-derived corpus of english. In *Proceedings of the 4th Web as Corpus Workshop (WAC-4) Can we beat Google*, pages 47–54.
- Alona Fyshe, Leila Wehbe, Partha P Talukdar, Brian Murphy, and Tom M Mitchell. 2015. A compositional and interpretable semantic space. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 32–41.
- Alex Gittens, Dimitris Achlioptas, and Michael W Mahoney. 2017. Skip-gram- zipf+ uniform= vector additivity. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 69–76.

- Pentti Kanerva, Jan Kristoferson, and Anders Holst. 2000. Random indexing of text samples for latent semantic analysis. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 22.
- Samuel Kaski. 1998. Dimensionality reduction by random mapping: Fast similarity computation for clustering. In *1998 IEEE International Joint Conference on Neural Networks Proceedings. IEEE World Congress on Computational Intelligence (Cat. No. 98CH36227)*, volume 1, pages 413–418. IEEE.
- Omer Levy and Yoav Goldberg. 2014. Neural word embedding as implicit matrix factorization. In *Advances in neural information processing systems*, pages 2177–2185.
- Omer Levy, Yoav Goldberg, and Ido Dagan. 2015. Improving distributional similarity with lessons learned from word embeddings. *Transactions of the Association for Computational Linguistics*, 3:211–225.
- Hongyin Luo, Zhiyuan Liu, Huanbo Luan, and Maosong Sun. 2015. Online learning of interpretable word embeddings. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1687–1692.
- Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.
- Brian Murphy, Partha Talukdar, and Tom Mitchell. 2012. Learning effective and interpretable semantic models using non-negative sparse embedding. *Proceedings of COLING 2012*, pages 1933–1950.
- Wen-Tsao Pan. 2011. A new evolutionary computation approach: fruit fly optimization algorithm. In *Proceedings of the conference on digital technology and innovation management*.
- Loïc Paulevé, Hervé Jégou, and Laurent Amsaleg. 2010. Locality sensitive hashing: A comparison of hash function types and querying mechanisms. *Pattern Recognition Letters*, 31(11):1348–1358.
- Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.
- Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*.
- Behrang QasemiZadeh and Laura Kallmeyer. 2016. Random positive-only projections: Ppmi-enabled incremental semantic space construction. In *Proceedings of the Fifth Joint Conference on Lexical and Computational Semantics*, pages 189–198.
- Behrang QasemiZadeh, Laura Kallmeyer, and Aurélie Herbelot. 2017. Projection aléatoire non-négative pour le calcul de word embedding. In *24e Conférence sur le Traitement Automatique des Langues Naturelles (TALN)*, pages 109–122.
- Magnus Sahlgren. 2005. An introduction to random indexing. In *Proceedings of the Methods and Applications of Semantic Indexing Workshop at the 7th International Conference on Terminology and Knowledge Engineering (TKE)*.
- Jamin Shin, Andrea Madotto, and Pascale Fung. 2018. Interpreting word embeddings with eigenvector analysis. *32nd Conference on Neural Information Processing Systems (NIPS 2018), IRASL workshop*.
- Malcolm Slaney and Michael Casey. 2008. Locality-sensitive hashing for finding nearest neighbors [lecture notes]. *IEEE Signal processing magazine*, 25(2):128–131.
- Charles F Stevens. 2015. What the fly's nose tells the fly's brain. *Proceedings of the National Academy of Sciences*, 112(30):9460–9465.
- Emma Strubell, Ananya Ganesh, and Andrew McCallum. 2019. Energy and policy considerations for deep learning in nlp. *arXiv preprint arXiv:1906.02243*.
- Peter D Turney and Patrick Pantel. 2010. From frequency to meaning: Vector space models of semantics. *Journal of artificial intelligence research*, 37:141–188.
- George Kingsley Zipf. 1932. *Selected studies of the principle of relative frequency in language*. Harvard university press.

The Impact of Self-Interaction Attention on the Extraction of Drug-Drug Interactions

Luca Putelli^{1,2}, Alfonso E. Gerevini¹, Alberto Lavelli², Ivan Serina¹

¹Università degli Studi di Brescia, ²Fondazione Bruno Kessler
{alfonso.gerevini, ivan.serina}@unibs.it, {l.putelli, lavelli}@fbk.eu

Abstract

Since a large amount of medical treatments requires the assumption of multiple drugs, the discovery of how these interact with each other, potentially causing health problems to the patients, is the subject of a huge quantity of documents. In order to obtain this information from free text, several methods involving deep learning have been proposed over the years. In this paper we introduce a Recurrent Neural Network-based method combined with the Self-Interaction Attention Mechanism. Such a method is applied to the DDI2013-Extraction task, a popular challenge concerning the extraction and the classification of drug-drug interactions. Our focus is to show its effect over the tendency to predict the majority class and how it differs from the other types of attention mechanisms.

1 Introduction

Given the increase of publications regarding side effects, adverse drug reactions and, more in general, how the assumption of drugs can cause risks of health issues that may affect patients, a large quantity of free-text containing crucial information has become available. For doctors and researchers, accessing this information is a very demanding task, given the number and the complexity of such documents.

Hence, the automatic extraction of Drug-Drug Interactions (DDI), i.e. situations where the simultaneous assumption of drugs can cause adverse drug reactions, is the goal of the DDIExtraction-2013 task (Segura-Bedmar et al., 2014). DDIs

have to be extracted from a corpus of free-text sentences, combining machine learning with natural language processing (NLP).

Starting from the introduction of word embedding techniques like Word2Vec (Mikolov et al., 2013) and GloVe (Pennington et al., 2014) for word representation, Recurrent Neural Networks (RNN) and in particular Long Short Term Memory networks (LSTM) have become the state-of-the-art technology for most of natural language processing tasks like text classification or relation extraction.

The main idea behind the attention mechanism (Bahdanau et al., 2014) is that the model “pays attention” only to the parts of the input where the most relevant information is present. In our case, this mechanism assigns a higher weight to the most influential words, i.e. the ones which describe an interaction between drugs.

Several attention mechanisms have been proposed in the last few years (Hu, 2018), in particular self-interaction mechanism (Zheng et al., 2018) applies attention with a different weight vector for each word in the sequence, producing a matrix that represents the influence between all word pairs. We consider this information very meaningful, especially in a task like this one where we need to discover connections between pairs of words.

In this paper we show how self-interaction attention improves the results in the DDI-2013 task, comparing it to other types of attention mechanisms. Given that this dataset is strongly unbalanced, the main focus of the analysis is how each attention mechanism deals with the tendency to predict the majority class.

2 Related work

The best performing teams in the DDI-2013 original challenge (Segura-Bedmar et al., 2014) used SVM (Björne et al., 2013) but, more recently, Convolutional Neural Networks (CNN) (Liu et al.,

Copyright © 2019 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0)

2016), (Quan et al., 2016) and mostly Recurrent Neural Networks (RNN) have proved to be the new state of the art.

Kumar and Anand (2017) propose a double LSTM. The sentences are processed by two different bidirectional LSTM layers: one followed by a max-pooling layer and the other one by a custom made attention-pooling layer that assign weights to words. Furthermore Zhang et al. (2018) design a multi-path LSTM neural network. Three parallel bidirectional LSTM layers process the sentence sequence and a fourth one processes the shortest dependency path between the two candidate drugs in the dependency tree. The output of these four layers is merged and handled by another bidirectional LSTM layer.

Zheng et al. (2017) apply attention directly to word vectors, creating a “candidate-drugs-oriented” input which is processed by a single LSTM layer.

Yi et al. (2017) use a RNN with Gated Recurrent Units (GRU) (Cho et al., 2014) instead of LSTM units, followed by a standard attention mechanism, and exploits information contained in other sentences with a custom made sentence attention mechanism.

Putelli et al. (2019) introduce an LSTM model followed by a self-interaction attention mechanism which computes, for each pair of words, a vector representing how much it is related to the other. These vectors are concatenated into a single one which is passed to a classification layer. In this paper, starting from the results reported in Putelli et al. (2019), we improve the input representation, the negative filtering and extend the analysis of self-interaction attention, comparing it to more standard attention mechanisms.

3 Dataset description

This dataset was released for the shared challenge SemEval 2013 - Task 9 (Segura-Bedmar et al., 2014) and contains annotated documents from the biomedical literature. In particular, there are two different sources: abstracts from MEDLINE research articles and texts from DrugBank.

Every document is divided into sentences and, for each sentence, the dataset provides annotations of every drug mentioned. The task requires to classify all the possible $\binom{n}{2}$ pairs of n drugs mentioned in the given sentences. The dataset provides the instances with their classification value.

There are five different classes: **unrelated**: there is no relation between the two drugs mentioned; **effect**: the text describes the effect of the drug-drug interaction; **advise**: the text recommends to avoid the simultaneous assumption of two drugs; **mechanism**: the text describes an anomaly of the absorption of a drug, if assumed simultaneously with another one; **int**: the text states a generic interaction between the drugs.

4 Pre-processing

The pre-processing phase exploits the “en_core_web_sm” model of spaCy¹, a Python tool for Natural Language Processing, and it is composed by these steps:

Substitution: after tokenization and POS-tagging, the drug mention tokens are replaced by the standard terms `PairDrug1` and `PairDrug2`. In the particular case when the pair is composed by two mentions of the same drug, these are replaced by `NoPair`. Every other drug mentioned in the sentence is replaced with the generic name `Drug`.

Shortest dependency path: spaCy produces the dependency tree associated to the sentence, with tokens as nodes and dependency relations between the words as edges. Then, we calculate the shortest path in the dependency tree between `PairDrug1` and `PairDrug2`.

Offset features: given a word w in the sentence, D_1 is calculated as the distance (in terms of words) from the first drug mention, divided by the length of the sentence. Similarly, D_2 is calculated as the distance from the second drug mention.

4.1 Negative instance filtering

The DDI-2013 dataset contains many “negative instances”, i.e. instances that belong to the unrelated class. In an unbalanced dataset, machine learning algorithms are more likely to classify a new instance over the majority class, leading to poor performance for the minority classes (Weiss and Provost, 2001). Given that previous studies (Chowdhury and Lavelli, 2013; Kumar and Anand, 2017; Zheng et al., 2017) have demonstrated a positive effect of reducing the number of negative instances on this dataset, we have filtered out some instances from the training-set relying only on the structure of the sentence, starting from the pairs of drugs with the same name. In

¹<https://spacy.io>

addition to this case, we can filter out a candidate pair if the two drug mentions appear in coordinate structure, checking the shortest dependency path between the two drug mentions. If they are not connected by a path, i.e. there is no grammatical relation between them, the candidate pair is filtered out.

While other works like (Kumar and Anand, 2017) and (Liu et al., 2016) apply custom-made rules for this dataset (such as regular expressions), our choice is to keep the pre-processing phase as general as possible, defining an approach that can be applied for other relation extraction tasks.

5 Model description

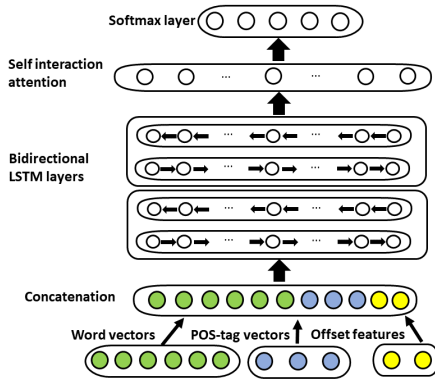


Figure 1: Model architecture

In this section we present the LSTM-based model (Figure 1), the self-attention mechanism and how it is used for relation extraction.

5.1 Embedding

Each word in our corpus is represented with a vector of length 200. These vectors are obtained with a Word2Vec (Mikolov et al., 2013) fine-tuning. We initialized a Word2Vec model with the vectors obtained by the authors of McDonald et al. (2018) the same algorithm over PubMed abstracts and PMC texts, and trained our Word2Vec model using the DDI-2013 corpus.

PoS tags are represented with vectors of length 4. These are obtained applying the Word2Vec method to the sequence of PoS tags in our corpus.

5.2 Bidirectional LSTM layer

A Recurrent neural network is a deep learning model for processing sequential data, like natural language sentences. Its issues with vanishing gradient are avoided using LSTM cells (Hochreiter and Schmidhuber, 1997; Gers et al., 2000),

which allow to process longer and more complex sequences. Given $x_1, x_2 \dots x_m$, h_{t-1} and c_{t-1} where m is the length of the sentence and $x_i \in \mathbb{R}^d$ is the vector obtained by concatenating the embedded features, h_{t-1} and c_{t-1} are the hidden state and the cell state of the previous LSTM cell (h_0 and c_0 are initialized as zero vectors), new hidden state and cell state values are computed as follows:

$$\begin{aligned}\hat{c}_t &= \tanh(W_c[h_{t_i}, x_t] + b_c) \\ i_t &= \sigma(W_i[h_{t_i}, x_t] + b_i) \\ f_t &= \sigma(W_f[h_{t_i}, x_t] + b_f) \\ o_t &= \sigma(W_o[h_{t_i}, x_t] + b_o) \\ c_t &= i_t * \hat{c}_t + f_t * c_{t-1} \\ h_t &= \tanh(c_t) * o_t\end{aligned}$$

with σ being the sigmoid activation function and $*$ denoting the element wise product. $W_f, W_i, W_o, W_c \in \mathbb{R}^{(N+d) \times N}$ are weight matrices and $b_f, b_i, b_o, b_c \in \mathbb{R}^N$ are bias vectors. Weight matrices and bias vectors are randomly initialized and learned by the neural network during the training phase. N is the LSTM layer size and d is the dimension of the feature vector for each input word. The vectors in square brackets are concatenated.

Bidirectional LSTM not only computes the input sequence in the order of the sentence but also backwards (Schuster and Paliwal, 1997). Hence, we can compute h^r using the same equations described earlier but reversing the word sequence. Given h_t computed in the sentence order and h_t^r in the reversed order, the output of the t bidirectional LSTM cell h_t^b is the result of the concatenation of h_t and h_t^r .

5.3 Sentence representation and attention mechanisms

The LSTM layers produce, for each word input w_i , a vector $h_i \in \mathbb{R}^n$ which is the result of computing every word from the start of the sentence to w_i . Hence, given a sentence of length m , h_m can be considered as the sentence representation produced by the LSTM layer. So, for a sentence classification task, h_m can be used as the input to a fully connected layer that provides the classification.

Even if they perform better than simple RNNs, LSTM neural networks have difficulties preserving dependencies between distant words (Raffel and Ellis, 2015) and, especially for long sentences, h_m may not be influenced by the first

words or may be affected by less relevant words. The **Attention** mechanism (Bahdanau et al., 2014; Kadlec et al., 2016) deals with these problems taking into consideration each h_i , computing weights α_i for each word contribution:

$$u_i = \tanh(W_a h_i + b_a) \\ \alpha_i = \text{softmax}(u_i) = \exp(u_i) / \sum_{k=1}^n \exp(u_k)$$

where $W_a \in \mathbb{R}^{N \times N}$ and $b_a \in \mathbb{R}^N$.

The attention mechanism outputs the *sentence representation*

$$s = \sum_{i=1}^m \alpha_i h_i$$

The **Context Attention** mechanism (Yang et al., 2016) is more complex. In order to enhance the importance of the words for the meaning of the sentence, this uses a *word level context vector* u_w of additional weights for the calculation of α_i :

$$\alpha_i = \text{softmax}(u_w^T u_i)$$

As proposed by Zheng et al. (2018), **Self-Interaction Attention** mechanism uses multiple v_i for each word w_i instead of using a single one. This way, we can extract the influence (called *action*) between the *action controller* w_i and the rest of the sentence, i.e. each w_k for $k \in \{1, m\}$. The action of w_i is calculated as follows:

$$s_i = \sum_{k=1}^m \alpha_{i,k} u_i \\ \alpha_{i,k} = \exp(v_k^T u_i) / \sum_{j=1}^m \exp(v_j^T u_i)$$

with u_i defined in the same way as the standard attention mechanism.

5.4 Model architecture

In order to obtain also in this case a *context vector* representing the sentence, in Zheng et al. (2018) each s_i is aggregated into a single vector s as its average, maximum or even applying another standard attention layer. In our model we choose to avoid any pooling operations and to concatenate instead each s_i , creating a *flattened representation* (Du et al., 2018) and passing it to the classification layer.

The model designed (see Figure 1) and tested for the DDI-2013 Relation Extraction task includes the following layers: three parallel **embedding layers**: one with pre-trained word vectors, one with pre-trained PoS tag vectors and one that calculates the embedding of the offset features; two **bidirectional LSTM layers** that process the word sequence; the **self-interaction attention mechanism**; a **fully connected layer** with

5 neurons (one for each class) and **softmax** activation function that provides the classification.

In our experiments, we compare this model with similar configurations obtained substituting the self-interaction attention with the standard attention layer introduced by Bahdanau et al. (2014) and the context-attention of Yang et al. (2016).

6 Results and discussion

Our models are implemented using Keras library with Tensorflow backend. We perform a simple random hyper-parameter search (Bergstra and Bengio, 2012) in order to optimize the learning phase and avoiding overfitting, using a subset of sentences as validation set.

6.1 Evaluation

We have tested our two models with different input configurations: using only word vectors, using word and PoS tag vectors or adding also offset features.

In Table 1 we show the recall measure for each input configuration. The effect of self-interaction is also verified through the Friedman test (Friedman, 1937): for all input configurations, the model with self-interaction attention performs better than the other configurations with a level of confidence equals to 99%. Similarly, the simple Attention Mechanism obtains better performances with respect to the Context Attention with confidence of 99% (see Figure 2).

In Table 2 we show the F-Score for each class of the dataset. The overall performance of the configuration including word vectors, PoS tagging and offset features as input is considered also in Table 3.

In Table 3 we compare our results with other state-of-the-art methods and compare the overall performance of the three attention mechanisms. The Context-Att obtains results similar to those of most of the approaches based on Convolution Neural Networks and worse than most of LSTM-based models.

In terms of F-Score, Word Attention LSTM (Zheng et al., 2017) outperforms our approach and the other LSTM-based models by more than 4%. As we discussed in (Putelli et al., 2019), we have tried to replicate their model but we could not obtain the same results. Furthermore, their attention mechanism aimed to creating a "candidate-drugs-oriented" input did not improve the performance.

Input	No Attention	Context-Att	Attention	Self-Int-Att
Word	64.44	65.32	66.60	69.72
Word+Tag	65.37	65.20	67.57	68.95
Word+Tag+Offset	60.67	65.82	69.47	70.88

Table 1: Overall recall (%) comparison with different attention mechanisms and input configurations. For each input configuration, the best recall is marked in bold.

Input	Effect				Mechanism			
	No Att	C-Att	Att	Self-Int	No Att	C-Att	Att	Self-Int
Word	0.68	0.71	0.72	0.70	0.69	0.72	0.72	0.70
Word+Tag	0.67	0.70	0.70	0.69	0.71	0.73	0.74	0.70
Word+Tag+Offset	0.65	0.70	0.70	0.69	0.68	0.73	0.74	0.76

Input	Advise				Int			
	No Att	C-Att	Att	Self-Int	No Att	C-Att	Att	Self-Int
Word	0.77	0.71	0.74	0.78	0.53	0.49	0.45	0.45
Word+Tag	0.78	0.73	0.77	0.77	0.55	0.50	0.45	0.43
Word+Tag+Offset	0.74	0.75	0.79	0.78	0.50	0.52	0.50	0.49

Table 2: Detailed F-Score comparison with different configurations and attention mechanisms. For each class, the best F-Score is marked in bold.

Method	P(%)	R(%)	F(%)
UTurku (SVM)	73.2	49.9	59.4
FBK-irst (SVM)	64.6	65.6	65.1
Zhao SCNN	72.5	65.1	68.6
Liu CNN	75.7	64.7	69.8
Multi-Channel	76.0	65.3	70.2
Context-Att	75.9	65.8	70.5
Joint-LSTMs	73.4	69.7	71.5
Self-Int	73.0	70.9	71.9
GRU	73.7	70.8	72.2
Attention	75.6	69.5	72.4
SDP-LSTM	74.1	71.8	72.9
Word-Att LSTM	78.4	76.2	77.3

Table 3: Comparison with overall precision (P), recall (R) and F-Score (F) of other state-of-the-art methods: , ordered by F. Our models are marked in bold, results higher than ours are marked in red.

7 Conclusions and future work

We have compared the self-interaction attention model to alternative configurations using the standard attention mechanism introduced by Bahdanau et al. (2014) and the context-attention mechanism of Yang et al. (2016).

Our experiments show that the self-interaction mechanism improves the performance with respect to other versions, in particular reducing the

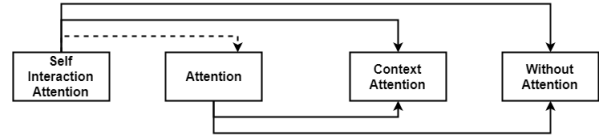


Figure 2: Recall comparison for models with different attention mechanisms for Word+Tag+Offset. The continue arrow means 99% confidence, while the dashed arrow means 95%.

tendency of predicting the majority class, hence decreasing the number of false negatives. The standard attention mechanism produces better results than the context attention.

As future work, our objective is to exploit or adapt the Transformer architecture (Vaswani et al., 2017), which has become quite popular for machine translation tasks and relies almost only on attention mechanisms, and apply it to relation extraction tasks like DDI-2013.

Another direction includes the exploitation of a different pre-trained language modeling. For example, BioBERT (Lee et al., 2019) obtains good results for several NLP tasks like Named Entity Recognition or Question Answering and we plan to apply it to our task.

References

- Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv:1409.0473* Comment: Accepted at ICLR 2015 as oral presentation.
- James Bergstra and Yoshua Bengio. 2012. Random search for hyper-parameter optimization. *J. Mach. Learn. Res.*, 13(1):281–305, February.
- Jari Björne, Suwisa Kaewphan, and Tapio Salakoski. 2013. UTurku: Drug named entity recognition and drug-drug interaction extraction using SVM classification and domain knowledge. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 651–659, Atlanta, Georgia, USA, June. Association for Computational Linguistics.
- Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*.
- Md. Faisal Mahbub Chowdhury and Alberto Lavelli. 2013. FBK-irst : A multi-phase kernel based approach for drug-drug interaction detection and classification that exploits linguistic information. In *Proceedings of the 7th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT 2013, Atlanta, Georgia, USA, June 14-15, 2013*, pages 351–355.
- Jinhua Du, Jingguang Han, Andy Way, and Dadong Wan. 2018. Multi-level structured self-attentions for distantly supervised relation extraction. *CoRR*, abs/1809.00699.
- Milton Friedman. 1937. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the American Statistical Association*, 32(200):675–701.
- Felix A. Gers, Jürgen Schmidhuber, and Fred A. Cummins. 2000. Learning to forget: Continual prediction with LSTM. *Neural Computation*, 12:2451–2471.
- Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9:1735–80, 12.
- Dichao Hu. 2018. An introductory survey on attention mechanisms in NLP problems. *CoRR*, abs/1811.05544.
- Rudolf Kadlec, Martin Schmid, Ondrej Bajgar, and Jan Kleindienst. 2016. Text understanding with the attention sum reader network. *CoRR*, abs/1603.01547.
- Sunil Kumar and Ashish Anand. 2017. Drug-drug interaction extraction from biomedical text using long short term memory network. *CoRR*, abs/1701.08303.
- Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. BioBERT: pre-trained biomedical language representation model for biomedical text mining. *arXiv preprint arXiv:1901.08746*.
- Shengyu Liu, Buzhou Tang, Qingcai Chen, and Xiaolong Wang. 2016. Drug-drug interaction extraction via convolutional neural networks. *Computational and mathematical methods in medicine*, 2016.
- Ryan McDonald, Georgios-Ioannis Brokos, and Ion Androutsopoulos. 2018. Deep relevance ranking using enhanced document-query interactions. *CoRR*, abs/1809.01682.
- Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 3111–3119. Curran Associates, Inc.
- Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.
- Luca Putelli, Alfonso E. Gerevini, Alberto Lavelli, and Ivan Serina. 2019. Applying self-interaction attention for extracting drug-drug interactions. In *Proceedings of 18th International Conference of the Italian Association for Artificial Intelligence*.
- Chanqin Quan, Lei Hua, Xiao Sun, and Wenjun Bai. 2016. Multichannel convolutional neural network for biological relation extraction. *BioMed research international*, 2016.
- Colin Raffel and Daniel P. W. Ellis. 2015. Feed-forward networks with attention can solve some long-term memory problems. *CoRR*, abs/1512.08756.
- Mike Schuster and Kuldip K Paliwal. 1997. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 45(11):2673–2681.
- Isabel Segura-Bedmar, Paloma Martínez, and María Herrero-Zazo. 2014. Lessons learnt from the DDIE Extraction-2013 shared task. *Journal of Biomedical Informatics*, 51:152–164.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *CoRR*, abs/1706.03762.

- Gary Weiss and Foster Provost. 2001. The effect of class distribution on classifier learning: An empirical study. Technical report, Department of Computer Science, Rutgers University.
- Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alexander J. Smola, and Eduard H. Hovy. 2016. Hierarchical attention networks for document classification. In *HLT-NAACL*.
- Zibo Yi, Shasha Li, Jie Yu, Yusong Tan, Qingbo Wu, Hong Yuan, and Ting Wang. 2017. Drug-drug interaction extraction via recurrent neural network with multiple attention layers. In *International Conference on Advanced Data Mining and Applications*, pages 554–566. Springer.
- Yijia Zhang, Wei Zheng, Hongfei Lin, Jian Wang, Zhihao Yang, and Michel Dumontier. 2018. Drug-drug interaction extraction via hierarchical RNNs on sequence and shortest dependency paths. *Bioinformatics*, 34(5):828–835.
- Wei Zheng, Hongfei Lin, Ling Luo, Zhehuan Zhao, Zhengguang Li, Zhang Yijia, Zhihao Yang, and Jian Wang. 2017. An attention-based effective neural model for drug-drug interactions extraction. *BMC Bioinformatics*, 18, 12.
- Jianming Zheng, Fei Cai, Taihua Shao, and Honghui Chen. 2018. Self-interaction attention mechanism-based text representation for document classification. *Applied Sciences*, 8(4).

Enriching Open Multilingual Wordnets with Morphological Features*

Stefania Racioppa

German Research Center
for Artificial Intelligence
Saarbrücken, Germany

stefania.racioppa@dfki.de

Thierry Declerck

German Research Center
for Artificial Intelligence
Saarbrücken, Germany

&

Austrian Centre for Digital Humanities
Austrian Academy of Sciences
Vienna, Austria

thierry.declerck@dfki.de

Abstract

English. In this article, we describe our work on porting Open Multilingual Wordnet resources into the OntoLex-Lemon model, in order to establish an interlinking with corresponding morphological resources, such as the MMorph resource set. For this purpose, the morphological resources were also ported onto OntoLex-Lemon. We show how the “lemmas” contained in the Wordnet resources can be enriched with morphological features using the lexical representation and linking features of OntoLex-Lemon, which support, among others, the formulation of restrictions in the usage of such expressions. Our work will result in an improved lexical resource combining Wordnet senses and full morphological descriptions in a single ontological framework, as specified in the OntoLex-Lemon model.

1 Introduction

WordNets are well-established lexical resources with a wide range of applications. For more than twenty years they have been elaborately set up and maintained by hand, especially the original Princeton WordNet of English (PWN) (Fellbaum, 1998). In recent years, there have been increasing activities in which open WordNets for different languages have been automatically extracted from various resources and enriched with lexical semantics information, building the so-called Open Multilingual Wordnet (OMW) (Bond and Paik, 2012). These WordNets were linked to PWN via

shared synset IDs (Bond and Foster, 2013; Bond et al., 2016). The resources in OMW are of different coverage and contain not always the same amount of information, as for example many resources are lacking definitions (or “glosses”), contrary to the PWN resource, or example sentences.

The work described in the present article is an extension of previous experiments done with English (Gromann and Declerck, 2019) and more recently with German lexical semantics resource, as we wanted to consider languages with a complex morphology.¹ In the present article we focus on Romance languages, especially Italian.

Our current work deals primarily with the morphological enrichment of OMW resources for Italian, i.e. “ItalWordNet”.² The first morphological resource we took into consideration for this purpose is an updated version of the MMorph morphological analyser (Petitpierre and Russell, 1995).

As a representation mean we chose OntoLex-Lemon (Cimiano et al., 2016)³, as this model has proven to be able to represent both a classical lexicographic description (McCrae et al., 2017) as well as lexical semantics networks like WordNet (McCrae et al., 2014).

OntoLex-Lemon is a further development of the “Lexicon Model for Ontologies” (*lemon*) (McCrae et al., 2012). Following the Guidelines⁴ for mapping Global WordNet formats onto *lemon*-based RDF⁵, some WordNets have already been

¹This work will be published soon in the proceedings of the Global Wordnet Conference 2019.

²See (Pianta et al., 2002; Toral et al., 2010). But we also made similar experiments with French and Spanish.

³See also <https://www.w3.org/2016/05/ontolex/> for more details.

⁴See <https://globalwordnet.github.io/schemas/##rdf>.

⁵RDF stands for “Resource Description Framework”. See

mapped onto the former *lemon* model (McCrae et al., 2014). Our present goal is thus to integrate conceptual descriptions, lemmas and morphological descriptions in the extended ontological framework specified by the OntoLex-Lemon model.⁶

In the next sections, we give some background information on OMW and MMorph. We continue with a section on OntoLex-Lemon, followed by sections that describe how OntoLex-Lemon supports the linking of lemmas in the OMW resources to full morphological descriptions. Doing so, morphological descriptions can be associated with the conceptual entries of WordNet.

2 Open Multilingual WordNet

OMW is an initiative that brings together Wordnets in different languages, linking them to the original Princeton WordNet (PWN). As stated on the web page of OMW, those Wordnets were of different quality, and some of those were in fact extracted from different types of language resources. We are dealing with three OMW WordNet resources.⁷ OMW provided for an harmonization of such resources, and published them in a uniform format, which is displayed just below, showing here a few examples from the Italian resource:

```
08388207-n ita:lemma nobiltà
08388207-n ita:lemma aristocrazia
08388207-n ita:lemma patriziato
08388207-n ita:def_0
    l'insieme degli aristocratici
08388207-n ita:def_1
    l'insieme dei nobili
...

14842992-n ita:lemma terra
14842992-n ita:lemma terreno
14842992-n ita:lemma suolo
14842992-n ita:def_0 parte
    superficiale della crosta
    terrestre sulla quale si
    sta o si cammina
14842992-n ita:exe_0 si piegò
    con fatica per raccogliere da
    terra i sacchetti, pronta a
    salire sull'autobus
14842992-n ita:exe_1 l'uomo
    cominciò a rotolarsi per terra
    in preda a dolori lancinanti
```

⁶<https://www.w3.org/RDF/> for more details.

⁷OntoLex-Lemon is indeed representing an ontology of lexical elements.

⁸French, Spanish and Italian, with a focus on the latter. See <http://compling.hss.ntu.edu.sg/omw/> for downloading the resources. For more details see also (Bond and Paik, 2012).

As the reader can see in the two examples above, OMW resources deliver information on the synset number, together with the part-of-speech of the associated lemma. In some cases, definitions (marked with *ita: def*) are provided, as well as examples (marked with *ita: exe*).

This format is used for all languages of the OMW corpus. This eases its mapping to a formal representation supporting the interoperability and interlinking of language resources, such as the OntoLex-Lemon model (see Section 4).

3 MMorph

MMorph was originally developed by ISSCO at the University of Geneva in the past MULTEXT project⁸. For our purposes, we used the extended MMorph version developed at DFKI LT Lab (*MMorph3*). This version includes huge lexical resources for English, French, German, Italian and Spanish. Very generally, the tool relates a word to a morphosyntactic description (MSD) containing free-definable attribute and values. The MMorph lexicon which is used to realize such MSD consists of a set of lexical entries and structural rules.⁹ For example, the following rule creates in Italian a noun plural concatenating the noun stem and the gender-specific suffixes:

Listing 1: Rule for noun plural generation in Italian. Note how the rule ensures that the gender doesn't change in the plural form.

```
N.ms: "o" NSuffix[num=sing gen=masc
    type=oa]
N.mp: "i" NSuffix[num=plur gen=masc
    type=oa]
N.fs: "a" NSuffix[num=sing gen=fem
    type=oa]
N.fp: "e" NSuffix[num=plur gen=fem
    type=oa]

FlexN: Noun[gen=$1 num=$2 form=surf]
    <- Noun[gen=$1 num=sing
        form=stem type=$T]
    N_ASfix[gen=$1 num=$2
        type=$T]
```

This rule will apply only to the lexical entries (feminine and/or masculine nouns) matching the defined features, e.g.

```
Noun[gen=masc num=sing form=stem
    type=oa]
    "patriziat" = "patriziato"
    "suol" = "suolo"
```

⁸See <https://www.issco.unige.ch/en/research/projects/MULTEXT.html> for more details on the resulting MMorph2.3.4 version.

⁹See (Petitpierre and Russell, 1995)

The morphology is completed by a set of spelling rules to catch the orthographic peculiarities of a specific language (e.g. *fung + i = funghi* in Italian).

The MMorph lexica can be dumped to full form lists for the usage in further programs, as can be seen in the following examples:

```
"nobiltà" = "nobiltà"
  Noun[ gen=fem num=sing | plur ]
"suoli" = "suolo"
  Noun[ gen=masc num=plur ]
"suolo" = "suolo"
  Noun[ gen=masc num=sing ]
```

The entries above are completed by labelled features for gender and number, but the user can freely define further features, if needed (e.g. *clitics* for verbal entries or *rection* of prepositions). Multiple values of a feature are expressed by “|”.

Because of their well-structured form, the dumped Mmorph lexica are ideally suited for the mapping into the OntoLex-Lemon format.

4 OntoLex-Lemon

The OntoLex-Lemon model was originally developed with the aim to provide a rich linguistic grounding for ontologies, meaning that the natural language expressions used in the description of ontology elements are equipped with an extensive linguistic description.¹⁰ This rich linguistic grounding includes the representation of morphological and syntactical properties of a lexical entry as well as the syntax-semantics interface, i.e. the meaning of these lexical entries with respect to an ontology or to specialized vocabularies.

The main organizing unit for those linguistic descriptions is the lexical entry, which enables the representation of morphological patterns for each entry (a MWE, a word or an affix). The connection of a lexical entry to an ontological entity is marked mainly by the *denotes* property or is mediated by the *LexicalSense* or the *LexicalConcept* properties, as represented in Figure 1, which displays the core module of the model.

OntoLex-Lemon is based on and extends the *lemon* model (McCrae et al., 2012). A major difference is that OntoLex-Lemon includes an explicit way to encode conceptual hierarchies, using the SKOS standard.¹¹ As shown

¹⁰See (McCrae et al., 2012), (Cimiano et al., 2016) and also https://www.w3.org/community/ontolex/wiki/Final_Model_Specification.

¹¹SKOS stands for “Simple Knowledge Organization Sys-

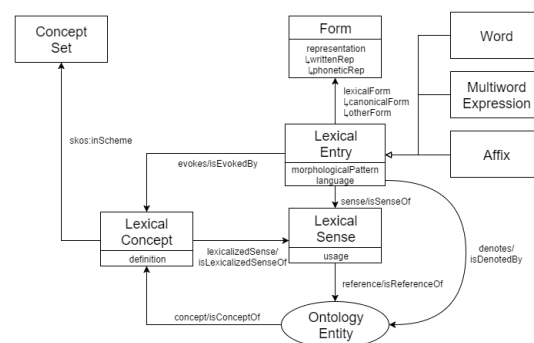


Figure 1: The core module of OntoLex-Lemon: Ontology Lexicon Interface. Graphic taken from <https://www.w3.org/2016/05/ontolex/>.

in Figure 1, lexical entries can be linked, via the *ontolex:evokes* property, to such SKOS concepts, which can represent WordNet synsets. This structure parallels the relation between lexical entries and ontological resources, which is implemented either directly by the *ontolex:reference* property or mediated by the instances of the *ontolex:LexicalSense* class.¹² The *ontolex:LexicalConcept* class seems to be most appropriate to model the “sets of cognitive synonyms (synsets)”¹³ described by Princeton WordNet (PWN), while the *ontolex:LexicalSense* class is meant to represent the bridge between lexical and ontological entities.

5 Mapping the OMW Resources to OntoLex-Lemon

As mentioned above, the format generated by the OMW initiative is very convenient to map dif-

ferent”. SKOS provides “a model for expressing the basic structure and content of concept schemes such as thesauri, classification schemes, subject heading lists, taxonomies, folksonomies, and other similar types of controlled vocabulary” (<https://www.w3.org/TR/skos-primer/>)

¹²Quoting from Section 3.6 “Lexical Concept” <https://www.w3.org/2016/05/ontolex/>: “We [...] capture the fact that a certain lexical entry can be used to denote a certain ontological predicate. We capture this by saying that the lexical entry denotes the class or ontology element in question. However, sometimes we would like to express the fact that a certain lexical entry evokes a certain mental concept rather than that it refers to a class with a formal interpretation in some model. Thus, in lemon we introduce the class *Lexical Concept* that represents a mental abstraction, concept or unit of thought that can be lexicalized by a given collection of senses. A lexical concept is thus a subclass of *skos:Concept*.”

¹³Quoted from <https://wordnet.princeton.edu/>.

ferent information onto more complex representation frameworks. To transform the OWN data onto the OntoLex-Lemon representation, a Python script was used. A design decision was to extract only the synset information and to encode the synsets as instances of the `LexicalConcept` class of OntoLex-Lemon. As some OWM lemmas are present in the MMorph resources, we just link the synsets to those lemmas, which are encoded as instances of the OntoLex-Lemon `LexicalEntry` class (see next section). We will need to create new instances of the OntoLex-Lemon `LexicalEntry` class for the OWM lemmas not present in the MMorph resources.

We have now 15553 such `LexicalConcept` instances for Italian. This is due to the fact that we consider only the subset of ItalWordNet that has been curated by OMW. We also noted that we have less instances of the `LexicalConcept` as lines for each synset in the original files, as the synset indices are represented by a unique URI in OntoLex-Lemon.

In Listing 2 we show examples of the OntoLex-Lemon encoding of two synsets for Spanish.¹⁴ The lemmas associated with these synsets are “cura”. In Section 7, we explain how the synsets are linked to the lemmas, which are differentiated in the OntoLex-Lemon representation, but not in the original OMW file.

Listing 2: The OntoLex-Lemon representation of two Spanish synsets

```
:synset_spawn-13491616-n
  rdf:type ontolex:LexicalConcept ;
  skos:inScheme :spawnet .

:synset_spawn-10470779-n
  rdf:type ontolex:LexicalConcept ;
  skos:inScheme :spawnet .
```

6 Mapping MMorph to OntoLex-Lemon

To transform the MMorph data into OntoLex-Lemon we used a Python script including the `rdflib` module¹⁵, which supports the generation of RDF-graphs in `rdf : xml`, `turtle`, or other relevant formats. In Listing 3, we show examples of the resulting data for the lemma “viola” in Italian.

¹⁴For the representation of OntoLex-Lemon data, we chose the `turtle` syntax serialization. More on the `turtle` syntax: <https://www.w3.org/TR/turtle/>.

¹⁵See <https://github.com/RDFLib/rdflib> for more details.

Listing 3: The OntoLex-Lemon entry for *viola*

```
:lex_viola_fem a ontolex:LexicalEntry ;
  lexinfo:partOfSpeech lexinfo:noun ;
  ontolex:canonicalForm :form_viola_f ;
  ontolex:otherForm :form_viola_f_pl .

:lex_viola_masc a ontolex:LexicalEntry ;
  lexinfo:partOfSpeech lexinfo:noun ;
  ontolex:canonicalForm :form_viola_m ;

:form_viola_f a ontolex:Form ;
  lexinfo:gender lexinfo:feminine ;
  lexinfo:number lexinfo:singular ;
  ontolex:writtenRep "viola"@it .

:form_viola_f_pl a ontolex:Form ;
  lexinfo:gender lexinfo:feminine ;
  lexinfo:number lexinfo:plural ;
  ontolex:writtenRep "violet"@it .

:form_viola_m a ontolex:Form ;
  lexinfo:gender lexinfo:masculine ;
  lexinfo:number lexinfo:plural ,
    lexinfo:singular ;
  ontolex:writtenRep "viola"@it .
```

As the reader can observe, we have two lexical entries for the entry “viola”, as this is requested by the OntoLex-Lemon guidelines, following which a word with different grammatical genders should have one lexical entry per gender. “Viola” in feminine is the music instrument, while in masculine it means “violet”. This is in fact an important feature for linking synsets to lemmas having distinct genders, as we will exemplify in Section 7.

The transformation of nominal entries from MMorph to the OntoLex-Lemon format resulted in 21085 instances of the class `LexicalEntry` for Italian. We still need to consider the lemmas of the OMW resources that are not in MMorph. This is concerning mostly multiword entries in OMW.

We will also investigate the use of other lexical resources, but the current use of the MMorph was motivated by the fact that we could have access to the different languages available in one and the same format, facilitating thus the uniform mapping into OntoLex-Lemon.

7 Linking the OMW Resources to the MMorph Resources

We see the use of OntoLex-Lemon for representing WordNets not only as a chance to port information from one format to another (including the possibility to publish WordNets in the Linguistic Linked Open Data cloud¹⁶), but also as an opportunity to extend the coverage of WordNet descrip-

¹⁶See <http://linguistic-lod.org/llod-cloud> and (Chiaros et al., 2012)

tions to more complex lexical phenomena, beyond lemma and PoS considerations. One case that has been studied in the recent past concerns the meaning that can be specifically associated to English plurals listed in PWN (Gromann and Declerck, 2019). We are interested in applying a similar approach to grammatical gender: we could link a Wordnet synset to a specific gender, as this information is normally not included in the Wordnets, which consider only the part-of-speech of the associated lemmas.

OntoLex-Lemon supports this linking in a straightforward manner. As can be seen in Figure 1, there is a property putting a `LexicalConcept` in relation to a `LexicalEntry`, i.e. the property `evokes` and its reverse `isEvokedBy`. Therefore we just need to add this property to both the OntoLex-Lemon representations of a synset and its corresponding entry. In Listing 4 we show such a case, taking again the word “cura” as an example.

Listing 4: Interlinking a synset and an entry for *cura*

```
:synset_spawn-13491616-n
  rdf:type ontolex:LexicalConcept ;
  skos:inScheme :spawnnet ;
  ontolex:evokes :lex_cura_1 .

:lex_cura_1 a ontolex:LexicalEntry ;
  lexinfo:gender lexinfo:fem ;
  lexinfo:partOfSpeech lexinfo:noun ;
  ontolex:canonicalForm :form_cura ;
  ontolex:otherForm :form_cura_plural ;
  ontolex:isEvokedBy
    :synset_spawn-1349161-n .

:synset_spawn-10470779-n
  rdf:type ontolex:LexicalConcept ;
  skos:inScheme :spawnnet ;
  ontolex:evokes :lex_cura_2 .

:lex_cura_2 a ontolex:LexicalEntry ;
  lexinfo:gender lexinfo:mas ;
  lexinfo:partOfSpeech lexinfo:noun ;
  ontolex:canonicalForm :form_cura ;
  ontolex:otherForm :form_cura_plural ;
  ontolex:isEvokedBy
    :synset_spawn-10470779-n .
```

Just adding the properties `evokes` and its reverse `isEvokedBy` to the corresponding elements in the generated OntoLex-Lemons data sets is providing for this morphological enrichment of the original Wordnets. Once the original (different types of) resources have been mapped onto the OntoLex-Lemon model, it is very easy to interlink or even to merge them into a richer representation. An extension of this work consists in describing restric-

tions on the usage of certain Wordnet concepts, as for example in the Italian case of the noun “bene” versus its plural form “beni”, or English “silk” versus the plural form “silks”, which are associated with different and sometimes not shareable meanings.¹⁷ We are making use for this of a strategy described in an extension to the core module of OntoLex-Lemon, called “Lexicog”,¹⁸ which foresees the description of instances of a class named `FormRestriction`, so that it is possible to state that a meaning is available only with the use of a specific form, like singular or plural.

8 Conclusion

We described our work on porting Open Multilingual Wordnet resources into the OntoLex-Lemon model, in order to establish an interlinking with corresponding morphological resources, such as the MMorph resource set. For this purpose, the morphological resources were also ported onto OntoLex-Lemon. As a result we noticed that this model can be easily used for bridging the WordNet type of lexical resources to a full description of lexical entries, which could possibly lead to an extension of the coverage of WordNets beyond the consideration of lemmas and PoS information.

We documented our interlinking work with the example of the full morphological representation of Italian words, putting them in relation with the corresponding OMW data sets. We also started to investigate the description of usage restrictions, which allows us to state formally that certain Wordnet concepts should be used only in the singular or in the plural form.

As a final goal of our work, we see the interlinked or merged resources in the Linguistic Linked Open Data (LLOD) cloud. We will investigate how our work can be combined with resources present in the LLOD, especially with the BabelNet framework, which is already integrating a huge number of lexical resources, including Princeton WordNet, and encyclopedic data sets (Ehrmann et al., 2014).

¹⁷The reader can see the different meanings associated to those plural words while querying for those in the user interface of PWN: <http://wordnetweb.princeton.edu/perl/webwn>.

¹⁸The current state of this “Lexicography” module is available at <https://www.w3.org/community/ontolex/wiki/Lexicography>.

Acknowledgments

The presented work has been supported in part by the H2020 project “Prêt-à-LLOD” with Grant Agreement number 825182. Contributions by Thierry Declerck have been supported additionally in part and by the H2020 project “ELEXIS” with Grant Agreement number 731015.

References

- Francis Bond and Ryan Foster. 2013. Linking and extending an open multilingual wordnet. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 1352–1362, Sofia.
- Francis Bond and Kyonghee Paik. 2012. A survey of wordnets and their licenses. *Small*, 8(4):5.
- Francis Bond, Piek Vossen, John P McCrae, and Christiane Fellbaum. 2016. Cili: the collaborative interlingual index. In *Proceedings of the Global WordNet Conference*, volume 2016.
- Christian Chiarcos, Sebastian Nordhoff, and Sebastian Hellmann, editors. 2012. *Linked Data in Linguistics - Representing and Connecting Language Data and Language Metadata*. Springer.
- Philipp Cimiano, John P. McCrae, and Paul Buitelaar. 2016. Lexicon Model for Ontologies: Community Report.
- Maud Ehrmann, Francesco Cecconi, Daniele Vannella, John Philip McCrae, Philipp Cimiano, and Roberto Navigli. 2014. Representing multilingual data as linked data: the case of BabelNet 2.0. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)*, pages 401–408, Reykjavik, Iceland, May. European Languages Resources Association (ELRA).
- Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA.
- Dagmar Gromann and Thierry Declerck. 2019. Towards the detection and formal representation of semantic shifts in inflectional morphology. In Maria Eskevich, Gerard de Melo, Christian Fth, John P. McCrae, Paul Buitelaar, Christian Chiarcos, Bettina Klimek, and Milan Dojchinovski, editors, *2nd Conference on Language, Data and Knowledge (LDK)*, volume 70 of *OpenAccess Series in Informatics (OA-SIcs)*, pages 21:1–21:15. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, 5.
- John McCrae, Guadalupe Aguado de Cea, Paul Buitelaar, Philipp Cimiano, Thierry Declerck, Asunción Gómez-Pérez, Jorge Gracia, Laura Hollink, Elena Montiel-Ponsoda, Dennis Spohr, and Tobias Wunner. 2012. Interchanging lexical resources on the Semantic Web. *Language Resources and Evaluation*, 46(6):701–709.
- John P. McCrae, Christiane Fellbaum, and Philipp Cimiano. 2014. Publishing and linking wordnet using lemon and rdf. In *Proceedings of the 3rd Workshop on Linked Data in Linguistics*.
- John P. McCrae, Paul Buitelaar, and Philipp Cimiano. 2017. The OntoLex-Lemon Model: Development and Applications. In Iztok Kosem, Jelena Kallas, Carole Tiberius, Simon Krek, Miloš Jakubíček, and Vít Baisa, editors, *Proceedings of eLex 2017*, pages 587–597. INT, Trojána and Lexical Computing, Lexical Computing CZ s.r.o., 9.
- Dominique Petitpierre and Graham. Russell. 1995. MMORPH: The Multext morphology program. Multext deliverable 2.3.1, ISSCO, University of Geneva.
- Emanuele Pianta, Luisa Bentivogli, and Christian Girardi. 2002. Multiwordnet: Developing an aligned multilingual database. In *Proceedings of the First International Conference on Global WordNet*, pages 293–302, Mysore, India.
- Antonio Toral, Stefania Bracale, Monica Monachini, and Claudia Soria. 2010. Rejuvenating the italian wordnet: upgrading, standardising, extending. In *Proceedings of the 5th International Conference of the Global WordNet Association (GWC-2010)*, Mumbai.

A Comparison of Representation Models in a Non-Conventional Semantic Similarity Scenario

Andrea Amelio Ravelli

University of Florence

andreaamelio.ravelli@unifi.it

Oier Lopez de Lacalle and Eneko Agirre

University of the Basque Country

e.agirre@ehu.eus

oier.lopezdelacalle@ehu.eus

Abstract

Representation models have shown very promising results in solving semantic similarity problems. Normally, their performances are benchmarked on well-tailored experimental settings, but what happens with unusual data? In this paper, we present a comparison between popular representation models tested in a non-conventional scenario: assessing action reference similarity between sentences from different domains. The action reference problem is not a trivial task, given that verbs are generally ambiguous and complex to treat in NLP. We set four variants of the same tests to check if different pre-processing may improve models performances. We also compared our results with those obtained in a common benchmark dataset for a similar task.¹

1 Introduction

Verbs are the standard linguistic tool that humans use to refer to actions, and action verbs are very frequent in spoken language (~50% of total verbs occurrences) (Moneglia and Panunzi, 2007). These verbs are generally ambiguous and complex to treat in NLP tasks, because the relation between verbs and action concepts is not one-to-one: e.g. (a) *pushing a button* is cognitively separated from (b) *pushing a table to the corner*; action (a) can also be predicated through *press*, while *move* can be used for (b) and not vice-versa (Moneglia, 2014). These represent two different *pragmatic actions*, despite of the verb used to describe it, and all the possible objects that can undergo the action. Another example could be the ambiguity behind a sentence like *John pushes the bottle*: is the

agent applying a continuous and controlled force to move the object from position A to position B, or is he carelessly shoving an object away from its location? These are just two of the possible interpretation of this sentence *as is*, without any other lexical information or pragmatic reference.

Given these premises, it is clear that the task of automatically classifying sentences referring to actions in a fine-grained way (e.g. *push/move* vs. *push/press*) is not trivial at all, and even humans may need extra information (e.g. images, videos) to precisely identify the exact action. One way could be to consider action reference similarity as a Semantic Textual Similarity (STS) problem (Agirre et al., 2012), assessing that lexical semantic information encodes, at a certain level, the action those words are referring to. The simplest way is to make use of pre-computed word embeddings, which are ready to use for computing similarity between words, sentences and documents. Various models have been presented in the past years that make use of well-known static word embeddings, like word2vec, GloVe and FastText (Mikolov et al., 2013; Pennington et al., 2014; Bojanowski et al., 2017). Recently, the best STS models rely on representations obtained from contextual embeddings, such as ELMO, BERT and XLNet (Peters et al., 2018; Devlin et al., 2018; Yang et al., 2019).

In this paper, we are testing the effectiveness of representation models in a non-conventional scenario, in which we do not have labeled data to train STS systems. Normally, STS is performed on sentence pairs that, on one hand, can have very close or distinct meaning, i.e. the assertion of similarity is easy to formulate; on the other hand, all sentences derive from the same domain, thus they share some syntactic regularities and vocabulary. In our scenario, we are computing STS between textual data from two different resources, IMAGACT and LSMDC16 (described respectively in

¹Copyright ©2019 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

5.1 and 5.2), in which the language used is highly different: from the first, synthetic and short captions; from the latter, audio descriptions. The objective is to benchmark word embedding models in the task of estimating the action concept expressed by a sentence.

2 Related Works

Word embeddings are abstract representations of words in the form of dense vectors, specifically tailored to encode semantic information. They represent an example of the so called transfer learning, as the vectors are built to minimize certain objective function (i.e., guessing the next word in a sentence), but successfully applied on different unrelated tasks, such as searching for words that are semantically related. In fact, embeddings are typically tested on semantic similarity/relatedness datasets, where a comparison of the vectors of two words is meant to mimic a human score that assesses the grade of semantic similarity between them.

The success of word embeddings on similarity tasks has motivated methods to learn representations of longer pieces of text such as sentences (Pagliardini et al., 2017), as representing their meaning is a fundamental step on any task requiring some level of text understanding. However, sentence representation is a challenging task that has to consider aspects such as compositionality, phrase similarity, negation, etc. The Semantic Textual Similarity (STS) task (Cer et al., 2017) aims at extending traditional semantic similarity/relatedness measures between pair of words in isolation to full sentences, and is a natural dataset to evaluate sentence representations. Through a set of campaigns, STS has distributed set of manually annotated datasets where annotators measure the similarity among sentences with a score that ranges between 0 (no similarity) to 5 (full equivalence).

In the recent years, evaluation campaigns that agglutinate many semantic tasks have been set up, with the objective to measure the performance of many natural language understanding systems. The most well-known benchmarks are SentEval² (Conneau and Kiela, 2018) and GLUE³ (Wang et al., 2019). They share many of existing

tasks and datasets, such as sentence similarity.

3 Problem Formulation

We cast the problem as a fine-grained action concept classification for verbs in LSMDC16 captions (e.g. *push* as *move* vs *push* as *press*, see Figure 1). Given a caption and the target verb from LSMDC16, our aim is to detect the most similar caption in IMAGACT that describe the action. The inputs to our model are the target caption and an inventory of captions that categorize the possible action concepts of the target verb. The model ranks the captions in the inventory according to the textual similarity with the target caption, and, similar to a kNN classifier, the model assigns the action label of k most similar captions.

4 Representation Models

In this section we describe the pretrained embeddings used to represent the contexts. Once we get the representation of each caption, the final similarity is computed based on cosine of the two representation vectors.

4.1 One-hot Encoding

This is the most basic textual representation, in which text is represented as binary vector indicating the words occurring in the context (Manning et al., 2008). This way of representing text creates long and sparse vectors, but it has been successfully used in many NLP tasks.

4.2 GloVe

The Global Vector model (GloVe)⁴ (Pennington et al., 2014) is a log-linear model trained to encode semantic relationships between words as vector offsets in the learned vector space, combining global matrix factorization and local context window methods.

Since GloVe is a word-level vector model, we compute the mean of the vectors of all words composing the sentence, in order to obtain the sentence-level representation. The pre-trained model from GloVe considered in this paper is the 6B-300d, counting a vocabulary of 400k words with 300 dimensions vectors and trained on a dataset of 6 billion tokens.

²<https://github.com/facebookresearch/SentEval>

³<https://gluebenchmark.com/>

⁴<https://nlp.stanford.edu/projects/glove/>

4.3 BERT

The Bidirectional Encoder Representations from Transformer (BERT)⁵ (Devlin et al., 2018) implements a novel methodology based on the so called *masked language model*, which randomly masks some of the tokens from the input, and predicts the original vocabulary id of the masked word based only on its context.

Similarly with GloVe, we extract the token embeddings of the last layer, and compute the mean vector to obtain the sentence-level representation. The BERT model used in our test is the BERT-Large Uncased (24-layer, 1024-hidden, 16-heads, 340M parameters).

4.4 USE

The Universal Sentence Encoder (USE) (Cer et al., 2018) is a model for encoding sentences into embedding vectors, specifically designed for transfer learning in NLP. Based on a deep averaging network encoder, the model is trained for a variety text length, such as sentences, phrases or short paragraphs, and in a variety of semantic task including the STS. The encoder returns the corresponding vector of the sentence, and we compute similarity using cosine formula.

5 Datasets

In this section, we briefly introduce the resources used to collect sentence pairs for our similarity test. Figure 1 shows some examples of data, aligned by action concepts.

5.1 IMAGACT

IMAGACT⁶ (Moneglia et al., 2014) is a multilingual and multimodal ontology of action that provides a video-based translation and disambiguation framework for action verbs. The resource is built on an ontology containing a fine-grained categorization of action concepts (*acs*), each represented by one or more visual prototypes in the form of recorded videos and 3D animations. IMAGACT currently contains 1,010 scenes, which encompass the actions most commonly referred to in everyday language usage.

Verbs from different languages are linked to *acs*, on the basis of competence-based annotation from mother tongue informants. All the verbs

that productively predicates the action depicted in an *ac* video are in *local equivalence* relation (Panunzi et al., 2018b), i.e. the property that different verbs (even with different meanings) can refer to the same action concept. Moreover, each *ac* is linked to a short *synthetic* caption (e.g. *John pushes the button*) for each locally equivalent verb in every language. These captions are formally defined, thus they only contain the minimum arguments needed to express an action.

We exploited IMAGACT conceptualization due to its *action-centric* approach. In fact, compared to other linguistic resources, e.g. WordNet (Fellbaum, 1998), BabelNet (Navigli and Ponzetto, 2012), VerbNet (Schuler, 2006), IMAGACT focuses on actions and represents them as visual concepts. Even if IMAGACT is a smaller resource, its action conceptualization is more fine-grained. Other resources have more broad scopes, and for this reason senses referred to actions are often vague and overlapping (Panunzi et al., 2018a), i.e. all possible actions can be gathered under one synset. For instance, if we look at the senses of *push* in Wordnet, we find that only 4 out of 10 synsets refer to concrete actions, and some of the glosses are not really exhaustive and can be applied to a wide set of different actions:

- push, force (move with force);
- push (press against forcefully without moving);
- push (move strenuously and with effort);
- press, push (make strenuous pushing movements during birth to expel the baby).

In such framework of categorization, all possible actions referred by *push* can be gathered under the first synset, except from those specifically described by the other three.

For the experiments proposed in this paper, only the English captions have been used, in order to test our method in a monolingual scenario.

5.2 LSMDC16

The Large Scale Movie Description Challenge Dataset⁷ (LSMDC16) (Rohrbach et al., 2017) consists in a parallel corpus of 128,118 sentences obtained from audio descriptions for visually impaired people and scripts, aligned to video clips

⁵<https://github.com/google-research/bert>

⁶<http://www.imagact.it>

⁷<https://sites.google.com/site/describingmovies/home>

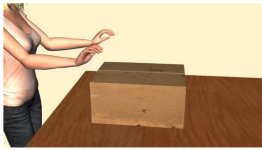


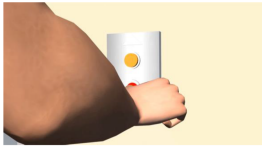




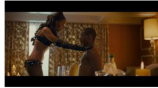
IMAGACT	LSMDC16
<p>ac_id: 40374041</p>  <p>PUSH: Mary pushes the box away SHOVE: Mary shoves the box away</p>	 <p>someone pushes him away</p>  <p>she pushes the plate away</p>
<p>ac_id: cbd1726a</p>  <p>PUSH: John pushes the button PRESS: John presses the button</p>	 <p>she presses a red button on the wall</p>  <p>the nazi officer pushes the snake's eye</p>
<p>ac_id: e017360a</p>  <p>PUSH: Mary pushes the basket under the table</p>	 <p>someone pushes the boxes out of his way</p>  <p>she pushes him onto the bed</p>

Figure 1: An example of aligned representation of action concepts in the two resources. On the left, action concepts with prototype videos and captions for all applicable verbs in IMAGACT; on the right, the video-caption pairs in LSMDC16, classified according to the depicted and described action.

from 200 movies. This dataset derives from the merging of two previously independent datasets, MPII-MD (Rohrbach et al., 2015) and M-VAD (Torabi et al., 2015). The language used in audio descriptions is particularly rich of references to physical action, with respect to reference corpora (e.g. BNC corpus) (Salway, 2007).

For this reason, LSMDC16 dataset could be considered a good source of video-caption pairs of action examples, comparable to data from IMAGACT resource.

6 Experiments

Given that the objective is not to discriminate distant actions (e.g. *opening a door* vs. *taking a cup*) but rather to distinguish actions referred to by the same verb or set of verbs, the experiments herein described have been conducted on a sub-set of the LSMDC16 dataset, that have been manually annotated with the corresponding *acs* from IMAGACT. The annotation has been carried on by one expert annotator, trained on IMAGACT conceptualization framework, and revised by a supervisor. In this way, we created a Gold Standard for the evaluation of the compared systems.

6.1 Gold Standard

The Gold Standard test set (GS) has been created by selecting one starting verb: *push*. This verb has been chosen according to the fact that, as a general action verb, it is highly frequent in the use, it applies to a high number of *acs* in the IMAGACT Ontology (25 *acs*) and it has a high occurrence both in IMAGACT and LSMDC16.

From the IMAGACT Ontology, all the verbs in relation of local equivalence with *push* in each of its *acs* have been queried⁸, i.e all the verbs that predicate at least one of the *acs* linked to *push*. Then, all the captions in LSMDC16 containing one of those verbs have been manually annotated with the corresponding *ac*'s id. In total, 377 video-caption pairs have been correctly annotated⁹ with 18 *acs*, and they have been paired with 38 captions for the verbs linked to the same *acs* in IMAGACT, consisting in a total of 14,440 similarity

⁸The verbs collected for this experiment are: *push*, *insert*, *press*, *ram*, *nudge*, *compress*, *squeeze*, *wheel*, *throw*, *shove*, *flatten*, *put*, *move*. *Move* and *put* have been excluded from this list, due to the fact that this verbs are too general and apply to a wide set of *acs*, with the risk of introducing more noise in the computation of the similarity; *flatten* is connected to an *ac* that found no examples in LSMDC16, so it has been excluded too.

⁹Pairs with no action in the video, or pairs with a novel or difficult to assign *ac* have been excluded from the test.

judgements.

It is important to highlight that the manual annotation took into account the visual information conveyed with the captions (i.e. videos from both resources), that made possible to precisely assign the most applicable *ac* to the LSMDC16 captions.

6.2 Pre-processing of the data

As stated in the introduction, STS methods are normally tested on data within the same domain. In attempt to leverage some differences between IMAGACT and LSMDC16, basic pre-processing have been applied.

Length of caption in the two resources vary: captions in IMAGACT are *artificial*, and they only contain minimum syntactic/semantic elements to describe the *ac*; captions in LSMDC16 are transcription of more natural spoken language, and usually convey information on more than one action at the same time. For this reason, LSMDC16 captions have been splitted in shorter and simpler sentences. To do that, we parsed the original caption with StanforNLP (Qi et al., 2018), and rewrote simplified sentences by collecting all the words in a dependency relation with the targeted verbs. Table 1 shows an example of the splitting process.

FULL	<i>As he crashes onto the platform, someone hauls him to his feet and pushes him back towards someone.</i>	✓
SPLIT	<i>he crashes onto the platform and</i>	✗
	<i>As someone hauls him to his feet</i>	✗
	<i>pushes him back towards someone</i>	✓

Table 1: Example of the split text after processing the output of the dependency parser. From the original caption (FULL) we obtain three sub-captions (SPLIT). Only the one with the target verb is used (✓), and the rest is ignored (✗).

LSMDC16 dataset is anonymised, i.e. the pronoun *someone* is used in place of all proper names; on the contrary, captions in IMAGACT always have a proper name (e.g. John, Mary). We automatically substituted IMAGACT proper names with *someone*, to match with LSMDC16.

Finally, we also removed stop-words, which are often the first lexical elements to be pruned out from texts, prior of any computation, because

they do not convey semantic information, and they sometimes introduce noise in the process. Stop-words removal has been executed in the moment of calculating the similarity between caption pairs, i.e. tokens corresponding to stop-words have been used for the representation by contextual models, but then discharged when computing sentence representation.

With these pre-processing operations, we obtained 4 variants of testing data:

- plain (LSMDC16 splitting only);
- anonIM (anonymisation of IMAGACT captions by substitution of proper names with *someone*);
- noSW (stop-words removing from both resources);
- anonIM+noSW (combination of the two previous ones).

7 Results

To benchmark the performances of the four models, we also defined a baseline that, following a binomial distribution, randomly assigns an *ac* of the GS test set (actually, baseline is calculated analytically without simulations). Parameters of the binomial are calculated from the GS test set. Table 2 shows the results at different recall@*k* (i.e. ratio of examples containing the correct label in the top *k* answers) of the three models tested.

All models show slightly better results compared to the baseline, but they are not much higher. Regarding the pre-processing, any strategy (noSW, anonIM, anonIM+noSW) seems not to make difference. We were expecting low results, given the difficulty of the task: without taking into account visual information, also for a human annotator most of those caption pairs are ambiguous.

Surprisingly, GloVe model, the only one with static pre-trained embeddings based on statistical distribution, outperforms the baseline and other contextual models by ~ 0.2 in recall@10. It is not an exciting result, but it shows that STS with pre-trained word embedding might be effective to speed up manual annotation tasks, without any computational cost. Probably, one reason to explain the lower trend in results obtained by contextual models (BERT, USE) could be that these systems have been penalized by the splitting process of LSMDC16 captions. Example in Table

Model	Pre-processing	recall@1	recall@3	recall@5	recall@10
ONE-HOT ENCODING	plain	0.195	0.379	0.484	0.655
	noSW	0.139	0.271	0.411	0.687
	anonIM	0.197	0.4	0.482	0.624
	anonIM+noSW	0.155	0.329	0.453	0.65
GLOVE	plain	0.213	0.392	0.553	0.818
	noSW	0.182	0.408	0.505	0.755
	anonIM	0.218	0.453	0.568	0.774
	anonIM+noSW	0.279	0.453	0.553	0.761
BERT	plain	0.245	0.439	0.539	0.632
	noSW	0.247	0.484	0.558	0.679
	anonIM	0.239	0.434	0.529	0.645
	anonIM+noSW	0.2	0.384	0.526	0.668
USE	plain	0.213	0.403	0.492	0.616
	noSW	0.171	0.376	0.461	0.563
	anonIM	0.239	0.471	0.561	0.666
	anonIM+noSW	0.179	0.426	0.518	0.637
Random baseline		0.120	0.309	0.447	0.658

Table 2: STS results for the models tested on IMAGACT-LSMDC scenario.

1 shows a good splitting result, while processing some other captions leads to less-natural sentence splitting, and this might influence the global result.

Model	Pre-processing	Pearson
GLOVE	plain	0.336
BERT	plain	0.47
USE	plain	0.702

Table 3: Results on STS-benchmark.

We run similar experiments on the publicly available STS-benchmark dataset¹⁰ (Cer et al., 2017), in order to see if the models show similar behaviour when benchmarked on a more conventional scenario. The task is similar to the one presented herein: it consists in the assessment of pairs of sentences according to their degree of semantic similarity. In this task, models are evaluated by the Pearson correlation of machine scores with human judgments. Table 3 shows the expected results: Contextual models outperform GloVe based model in a consistent way, and USE outperform the rest by large margin (about 20-30 points better overall). It confirms that model performances are task-dependent, and that results obtained in *non-conventional* scenarios can be counter-intuitive if compared to results obtained in conventional ones.

¹⁰<http://ixa2.si.ehu.es/stswiki/index.php/STSBenchmark>

8 Conclusions and Future Work

In this paper we presented a comparison of four popular representation models (one-hot encoding, GloVe, BERT, USE) in the task of semantic textual similarity on a non-conventional scenario: action reference similarity between sentences from different domains.

In the future, we would like to extend our Gold Standard dataset, not only in terms of dimension (i.e. more LSMDC16 video-caption pairs annotated with *acs* from IMAGACT), but also in terms of annotators. It would be interesting to observe to what extent the visual stimuli offered by video prototypes can be interpreted clearly by more than one annotator, and thus calculate the inter-annotator agreement. Moreover, we plan to extend the evaluation to other representation models as well as state-of-the-art supervised models, and see if their performances in canonical tests are confirmed on our scenario. We would also try to augment data used for this test, by exploiting dense video captioning models, i.e. videoBERT (Sun et al., 2019).

Acknowledgements

This research was partially supported by the Spanish MINECO (DeepReading RTI2018-096846-B-C21 (MCIU/AEI/FEDER, UE)), ERA-Net CHISTERA LIHLITH Project funded by Agencia Estatal de Investigación (AEI, Spain) projects PCIN-2017-118/AEI and PCIN-2017-085/AEI, the Basque Government (excellence research group, IT1343-19), and the NVIDIA GPU grant program.

References

- Eneko Agirre, Daniel Cer, Mona Diab, and Aitor Gonzalez-Agirre. 2012. SemEval-2012 Task 6: A pilot on semantic textual similarity. In **SEM 2012 - 1st Joint Conference on Lexical and Computational Semantics*, pages 385–393. Universidad del Pais Vasco, Leioa, Spain, January.
- Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics (TACL)*, 5(1):135–146.
- Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. 2017. SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 1–14, Vancouver, Canada, August. Association for Computational Linguistics.
- Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. 2018. Universal sentence encoder. *CoRR*.
- Alexis Conneau and Douwe Kiela. 2018. Senteval: An evaluation toolkit for universal sentence representations. *arXiv preprint arXiv:1803.05449*.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT - Pre-training of Deep Bidirectional Transformers for Language Understanding. *CoRR*, 1810:arXiv:1810.04805.
- Christiane Fellbaum. 1998. *WordNet: an electronic lexical database*. MIT Press.
- Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to Information Retrieval*. Cambridge University Press, New York, NY, USA.
- Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed Representations of Words and Phrases and their Compositionality. In C J C Burges, L Bottou, M Welling, Z Ghahramani, and K Q Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 3111–3119. Curran Associates, Inc.
- Massimo Moneglia and Alessandro Panunzi. 2007. Action Predicates and the Ontology of Action across Spoken Language Corpora. In M Alcántara Plá and Th Declerck, editors, *Proceedings of the International Workshop on the Semantic Representation of Spoken Language (SRS� 2007)*, pages 51–58, Salamanca.
- Massimo Moneglia, Susan Brown, Francesca Frontini, Gloria Gagliardi, Fahad Khan, Monica Monachini, and Alessandro Panunzi. 2014. The IMAGACT Visual Ontology. An Extendable Multilingual Infrastructure for the representation of lexical encoding of Action. *LREC*, pages 3425–3432.
- Massimo Moneglia. 2014. The variation of Action verbs in multilingual spontaneous speech corpora. *Spoken Corpora and Linguistic Studies*, 61:152.
- Roberto Navigli and Simone Paolo Ponzetto. 2012. Babelnet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193(0):217 – 250.
- Matteo Pagliardini, Prakhar Gupta, and Martin Jaggi. 2017. Unsupervised learning of sentence embeddings using compositional n-gram features. *CoRR*, abs/1703.02507.
- Alessandro Panunzi, Lorenzo Gregori, and Andrea Amelio Ravelli. 2018a. One event, many representations. mapping action concepts through visual features. In James Pustejovsky and Ielka van der Sluis, editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).
- Alessandro Panunzi, Massimo Moneglia, and Lorenzo Gregori. 2018b. Action identification and local equivalence of action verbs: the annotation framework of the imagact ontology. In James Pustejovsky and Ielka van der Sluis, editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).
- Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. GloVe: Global vectors for word representation. In *Conference on Empirical Methods in Natural Language Processing*, pages 1532–1543. Stanford University, Palo Alto, United States, January.
- Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of NAACL-HLT*, pages 2227–2237.
- Peng Qi, Timothy Dozat, Yuhao Zhang, and Christopher D Manning. 2018. Universal Dependency Parsing from Scratch. *CoNLL Shared Task*.
- Anna Rohrbach, Marcus Rohrbach, Niket Tandon, and Bernt Schiele. 2015. A dataset for Movie Description. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Anna Rohrbach, Atousa Torabi, Marcus Rohrbach, Niket Tandon, Christopher Pal, Hugo Larochelle, Aaron Courville, and Bernt Schiele. 2017. Movie Description. *International Journal of Computer Vision*, 123(1):94–120, January.
- Andrew Salway. 2007. A corpus-based analysis of audio description. In Jorge Díaz Cintas, Pilar Orero, and Aline Remael, editors, *Media for All*, pages 151–174. Leiden.
- Karin Kipper Schuler. 2006. *VerbNet: A Broad-Coverage, Comprehensive Verb Lexicon*. Ph.D. thesis, University of Pennsylvania.
- Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. 2019. Videobert: A joint model for video and language representation learning. *CoRR*, abs/1904.01766.
- Atousa Torabi, Christopher J Pal, Hugo Larochelle, and Aaron C Courville. 2015. Using Descriptive Video Services to Create a Large Data Source for Video Annotation Research. *cs.CV:arXiv:1503.01070*.
- Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *International Conference on Learning Representations*.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime G. Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *CoRR*.

How Much Competence Is There in Performance?

Assessing the Distributional Hypothesis in Word Bigrams

Johann Seltmann

University of Potsdam[◇]

jseltmann@uni-potsdam.de

Luca Ducceschi

University of Trento[♣]

luca.ducceschi@unitn.it

Aur lie Herbelot

University of Trento[♣]

aurelie.herbelot@unitn.it

[◇]Department of Linguistics, [♣]Dept. of Psychology and Cognitive Science, [♣]Center for Mind/Brain Sciences,
Dept. of Information Engineering and Computer Science

Abstract

The field of Distributional Semantics (DS) is built on the ‘distributional hypothesis’, which states that meaning can be recovered from statistical information in observable language. It is however notable that the computations necessary to obtain ‘good’ DS representations are often very involved, implying that if meaning is derivable from linguistic data, it is not directly encoded in it. This prompts questions related to fundamental questions about language acquisition: if we regard text data as linguistic *performance*, what kind of ‘innate’ mechanisms must operate over that data to reach *competence*? In other words, how much of semantic acquisition is truly data-driven, and what must be hard-encoded in a system’s architecture? In this paper, we introduce a new methodology to pull those questions apart. We use state-of-the-art computational models to investigate the amount and nature of transformations required to perform particular semantic tasks. We apply that methodology to one of the simplest structures in language: the word bigram, giving insights into the specific contribution of that linguistic component.¹

1 Introduction

The traditional notions of *performance* and *competence* come from Chomsky’s work on syntax (Chomsky, 1965), where much emphasis is put on the mental processes underpinning language acquisition. Chomsky posits the existence of a Universal Grammar, *innate* in the human species, which gets specialised to the particular language of a speaker. By exposure to the imperfect utterances of their community (referred to as *performance* data), an individual configures their UG to

reach some ideal knowledge of that community’s language, thereby reaching *competence*.

The present paper borrows the notions of ‘performance’, ‘competence’ and ‘innateness’ to critically analyse the semantic ‘acquisition’ processes simulated by Distributional Semantics models (DSMs). Our goal is to tease apart how much of their observed competence is due to the performance data they are exposed to, and how much is contributed by ‘innate’ properties of those systems, i.e. by their specific architectures.

DSMs come in many shapes. Traditional unsupervised architectures rely on counting co-occurrences of words with other words or documents (Turney and Pantel, 2010; Erk, 2012; Clark, 2012). Their neural counterparts, usually referred to as ‘predictive models’ (Baroni et al., 2014) learn from a language modelling task over raw linguistic data (e.g. Word2Vec, Mikolov et al., 2013, GloVe Pennington et al., 2014). The most recent language embedding models (Vaswani et al., 2017; Radford et al., 2018), ELMo (Peters et al., 2018), or BERT (Devlin et al., 2018) compute *contextualised* word representations and sentence representations, yielding state-of-the-art results on sentence-related tasks, including translation. In spite of their differences, all models claim to rely on the *Distributional Hypothesis* (Harris, 1954; Firth, 1957), that is, the idea that distributional patterns of occurrences in language correlate with specific aspects of meaning.

The Distributional Hypothesis, as stated in the DSM literature, makes semantic acquisition sound like an extremely data-driven procedure. But we should ask to what extent meaning indeed is to be found in statistical patterns. The question is motivated by the observation that the success of the latest DSMs relies on complex mechanisms being applied to the underlying linguistic data or the task at hand (e.g. attention, self-attention, negative

¹Copyright  2019 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

sampling, particular objective functions). Such mechanisms have been shown to apply very significant transformations to the original input data: for instance, the Word2Vec objective function introduces parallelisms in the space that make it perform particularly well on analogy tasks (Gittens et al., 2017). Models such as BERT apply extensive processing to the input through stacks of encoders. So while meaning can be *derived* from training regimes involving raw data, it is not directly encoded in it.

Interestingly, Harris himself (Harris, 1954) points out that a) distributional structure is in no simple relation to the structure of meaning; b) different distributions in language encode different phenomena with various levels of complexity. We take both points as highlighting the complex relation between *linguistic* structure and the *cognitive* mechanisms that are necessary to apply to the raw input to retrieve semantic information. The point of our paper is to understand better what is encoded in observable linguistic structures (at the level of raw performance data), and how much distortion of the input needs to be done to acquire meaning (i.e. what cognitive mechanisms are involved in learning semantic competence).

In the spirit of Harris, we think it is worth investigating the behaviour of specific components of language and understand which aspects of meaning they encode, and to what extent. The present work illustrates our claim by presenting an exploratory analysis of one of the simplest recoverable structure in corpora: the word bigram. Our methodology is simple: we test the raw distributional behaviour of the constituent over different tasks, comparing it to a state-of-the-art model. We posit that each task embodies a specific aspect of competence. By inspecting the difference in performance between the simplest and more complex models, we get some insight into the way a particular structure (here, the bigram) contributes to the acquisition of specific linguistic faculties. The failures of raw linguistic data to encode a particular competence points at some necessary, ‘innate’ constraint of the acquisition process, which might be encoded in a model’s architecture as well as the specific task that it is required to solve.

In what follows, we propose to investigate the behaviour of the bigram with respect to three different levels of semantic competence, corresponding to specific tasks from the DS literature: a)

word relatedness; b) sentence relatedness; c) sentence autoencoding (Turney, 2014; Bowman et al., 2016). The first two tasks test to which extent the linguistic structure under consideration encodes topicality: if it does, it should prove able to cluster together similar lexical items, both in isolation and as the constituents of sentences. The third task evaluates the ability of a system to build a sentence representation and from that representation alone, recover the original utterance. That is, it tests *distinguishability* of representations. Importantly, distinguishability is at odds with the relatedness tasks which favour *clusterability*. The type of space learned from the raw data will necessarily favour one or the other. Our choice of tasks thus allows us to understand which type of space can be learned from the bigram: we will expand on this in our discussion (§6).²

2 Related work

The Distributional Hypothesis is naturally encoded in *count-based* models of Distributional Semantics (DS), which build lexical representations by gathering statistics over word co-occurrences. Over the years, however, these simple models have been superseded by so-called *predictive* models such as Word2Vec (Mikolov et al., 2013) or FastText (Bojanowski et al., 2017), which operate via language modeling tasks. These neural models involve sets of more or less complex procedures, from subsampling to negative sampling and subword chunking, which give them a clear advantage over methods that stick more closely to distributions in corpora. At the level of higher constituents, the assumption is that a) additional composition functions must be learned over the word representations to generate meaning ‘bottom-up’ (Clark, 2012; Erk, 2012); b) the semantics of a sentence influences the meaning of its parts ‘top-down’, leading to a notion of contextualised word semantics, retrievable by yet another class of distributional models (Erk and Padó, 2008; Erk et al., 2010; Thater et al., 2011; Peters et al., 2018). Bypassing the word level, some research investigates the meaning of sentences directly. Following from classic work on seq2seq architectures and attention, various models have been proposed to generate sentence embeddings through highly param-

²Our code for this investigation can be found under <https://github.com/sejo95/DSGeneration.git>.

eterised stacks of encoders (Vaswani et al., 2017; Radford et al., 2018; Devlin et al., 2018).

This very brief overview of work in DS shows the variety of models that have been proposed to encode meaning at different levels of constituency, building on more and more complex mechanisms. Aside from those efforts, much research has also focused on finding ideal hyperparameters for the developed architectures (Bullinaria and Levy, 2007; Baroni et al., 2014), ranging from the amount of context taken into account by the model to the type of task it should be trained on. Overall, it is fair to say that if meaning can be retrieved from raw language data, the process requires knowing the right transformations to apply to that data, and the right parametrisation for those transformations, including the type of linguistic structure the model should focus on. One important question remains for the linguist to answer: how much semantics was actually contained in corpus statistics, and where? We attempt to set up a methodology to answer this question, and use two different types of tasks (relatedness and autoencoding) to support our investigation.

While good progress has been made in the DS community on modelling relatedness, distinguishability has received less attention. Some approaches to autoencoding suggest using syntactic elements (such as syntax trees) for decomposition of an embedding vector into a sentence (Dinu and Baroni, 2014; Iyyer et al., 2014). However, some research suggests that this may not be necessary and that continuous bag-of-words representations and n-gram models contain enough word order information to reconstruct sentences (Schmaltz et al., 2016; Adi et al., 2017). Our own methodology is inspired by White et al. (2016b), who decode a sentence vector into a bag of words using a greedy search over the vocabulary. In order to also recover word order, those authors expand their original system in White et al. (2016a) by combining it with a traditional trigram model, which they use to reconstruct the original sentence from the bag of words.

3 Methodology

3.1 A bigram model of Distributional Semantics

We construct a count-based DS model by taking bigrams as our context windows. Specifically, for a word w_i , we construct an embedding vec-

tor \vec{v}_i which has one entry for each word w_j in the model. The entry \vec{v}_{ij} then contains the bigram probability $p(w_j|w_i)$.

We talked in our introduction of ‘raw’ linguistic structure without specifying at which level it is to be found. Following Church and Hanks (1990), we consider the joint probability of two events, relative to their probability of occurring independently, to be a good correlate of the fundamental psycholinguistic notion of *association*. As per previous work, we thus assume that a PMI-weighted DS space gives the most basic representation of the information contained in the structure of interest. For our bigram model, the numerator and denominator of the PMI calculation exactly correspond to elements in our bigram matrix B weighted by elements of our unigram vector U :

$$pmi(w_i, w_j) \equiv \log \frac{p(w_j|w_i)}{p(w_j)} \quad (1)$$

In practice, we use PPMI weighting and map every negative PMI value to 0.

Word relatedness: following standard practice, we compute relatedness scores as the cosine similarity of two PPMI-weighted word vectors, $\cos(\vec{w}_i, \vec{w}_j)$. For evaluation, we use the MEN test collection (Bruni et al., 2014), which contains 3000 word pairs annotated for relatedness; we compute the spearman ρ correlation between system and human scores.

Sentence relatedness: we follow the proof given by Paperno and Baroni (2016), indicating that the meaning of a phrase ab in a count-based model with PMI weighting is roughly equivalent to the addition of the PMI-weighted vectors of a and b (shifted by some usually minor correction). Thus, we can compute the similarity of two sentences $S1$ and $S2$ as:

$$\cos\left(\sum_{w_i \in S_1} \vec{w}_i, \sum_{w_j \in S_2} \vec{w}_j\right) \quad (2)$$

We report sentence relatedness scores on the SICK dataset (Marelli et al., 2014), which contains 10,000 utterance pairs annotated for relatedness. We calculate the relatedness for each pair in the dataset and order the pairs according to the results. We then report the spearman correlation between the results of the model and the ordering of the dataset.

Autoencoding of sentences: White et al. (2016b) encode a sentence as the sum of the word

embedding vectors of the words of that sentence. They *decode* that vector (the *target*) back into a bag of words in two steps. The first step, *greedy addition* begins with an empty bag of words. In each step a word is selected, such that the sum of the word vectors in the bag and the vector of the candidate item is closest to the target (using Euclidian distance as similarity measure). This is repeated until no new word could bring the sum closer to the target than it already is. The second step, *n-Substitution* begins with the bag of n words found in the greedy addition. For each subbag of size $m \leq n$ it considers replacing it with another possible subbag of size $\leq m$. The replacement that brings the sum closest to the target vector is chosen. We follow the same procedure, except that we only consider subbags of size 1, i.e. substitution of single words, for computational efficiency. In addition, the bigram component of our model B lets us turn the bags of words back into an ordered sequence.³ We use a beam search to perform this step, following [Schmaltz et al. \(2016\)](#).

We evaluate sentence autoencoding in two ways. First, we test the bag-of-words reconstruction on its own, by feeding the system the encoded sentence embedding and evaluating whether it can retrieve *all* single words contained in the original utterance. We report the proportion of perfectly reconstructed bags-of-words across all test instances. Second, we test the entire autoencoding process, including word re-ordering. We use two different metrics: a) **the BLEU score:** ([Papineni et al., 2002](#)), which computes how many n-grams of a decoded sentence are shared with several reference sentences, giving a precision score; b) **the CIDEr-D score:** ([Vedantam et al., 2015](#)) which accounts for both precision and recall and is computed using the average cosine similarity between the vector of a candidate sentence and a set of reference vectors. For this evaluation, we use the PASCAL-50S dataset (included in CIDEr-D), a caption generation dataset, that contains 1000 images with 50 reference captions each. We encode and decode the first reference caption for each image and use the remaining 49 as reference for the CIDEr and BLEU calculations.

For the actual implementation of the model, we

³Note that although a bigram language model would normally perform rather poorly on sentence generation, having a constrained bag-of-words to reorder makes the task considerably simpler.

build B and U from 90% of the BNC (≈ 5.4 million sentences), retaining 10% for development purposes. We limit our vocabulary to the 50000 most common words in the corpus, therefore the matrix is of the size 50002×50002 , including tokens for sentence beginning and end.

3.2 Comparison

In what follows, we compare our model to two Word2Vec models, which provide an upper bound for what a DS model may be to achieve. One model, W2V-BNC, is trained from scratch on our BNC background corpus, using gensim ([Řehůřek and Sojka, 2010](#)) with 300 dimensions, window size ± 5 , and ignoring words that occur less than five times in the corpus. The other model, W2V-LARGE, is given by out-of-the-box vectors released by [Baroni et al. \(2014\)](#): that model is trained on 2.5B words, giving an idea of the system’s performance on larger data. In all cases, we limit the vocabulary to the same 50,000 words included in the bigram model.

Note that given space restrictions, we do not disentangle the contribution of the models themselves and the particular type of linguistic structure they are trained on. Our results should thus be taken as indication of the amount of information encoded in a raw bigram model compared to what can be obtained by a state-of-the-art model using the best linguistic structure at its disposal (here, a window of ± 5 words around the target).

4 Results

Word relatedness: the bigram model obtains an acceptable $\rho = 0.48$ on the MEN dataset. W2V-BNC and W2V-LARGE perform very well, reaching $\rho = 0.72$ and $\rho = 0.80$. Note that whilst the bigram model lags well behind W2V, it achieves its score with what is in essence a unidirectional model with window of size 1 – that is, with as minimal input as it can get, seeing 10 times less co-occurrences than W2V-BNC.

Sentence relatedness: the bigram model obtains $\rho = 0.40$ on the sentence relatedness task. Interestingly, that score increases by 10 points, to $\rho = 0.50$, when filtering away frequent words with probability over 0.005. W2V-BNC and W2V-LARGE give respectively $\rho = 0.59$ and $\rho = 0.61$.

Sentence autoencoding: we evaluate sentence autoencoding on sentences from the Brown corpus ([Kučera and Francis, 1967](#)), using seven bins

sent. length	original sents.		in matrix	
	W2V	CB	W2V	CB
3-5	0.556	0.792	0.686	0.988
6-8	0.380	0.62	0.646	0.988
9-11	0.279	0.586	0.548	1.0
12-14	0.210	0.578	0.402	1.0
15-17	0.178	0.338	0.366	0.978
18-20	0.366	0.404	0.984	0.974
21-23	0.306	0.392	0.982	0.968

Table 1: Fraction of exact matches in bag-of-word reconstruction (W2V refers to W2V-LARGE)

for different sentence lengths (from 3-5 words to 21-23 words). Each bin contains 500 sentences. In some cases, the sentences contained words that aren't present in the matrix and which are therefore skipped for encoding. We thus look at two different values: a) in how many cases the reconstruction returns exactly the words in the sentence; b) in how many cases the reconstruction returns the words in the sentence which are contained in the matrix (results in Table 1).

The bigram model shines in this task: ignoring words not contained in the matrix leads to almost perfect reconstruction. In comparison, the W2V model has extremely erratic performance (Table 1), with scores decreasing as a function of sentence length (from 0.686 for length 3-5 to 0.366 for length 15-17), but increasing again for lengths over 18.

One interesting aspect of the bigram model is that it also affords a semantic competence that W2V does not naturally have: encoding a sequence and decoding it back into an ordered sequence. We inspect how well the model does at that task, compared to a random reordering baseline. Results are listed in Table 2. The bigram model clearly beats the baseline for all sentence lengths. But it is expectedly limited by the small n-gram size provided by the model. Table 3 contains examples of sentences from the brown corpus and their reconstructions. We see that local ordering is reasonably modeled, but the entire sentence structure fails to be captured.

5 Discussion

On the back of our results, we can start commenting on the particular contribution of bigrams to the semantic competences tested here. First, bigrams are moderately efficient at capturing relat-

	all	2-10	11-23
CIDEr-D bigram	1.940	1.875	2.047
BLEU bigram	0.193	0.209	0.176
CIDEr-D random	1.113	1.1	1.134
BLEU random	0.053	0.059	0.045

Table 2: CIDEr-D and BLEU scores on reordering of bags-of-words using our bigram matrix and random reordering. Results are given for all sentences as well as sentences of lengths 2-10 and 11-23.

Original sentence	Reconstruction
They have to be.	they have to be .
Six of these were proposed by religious groups.	by these were six of religious groups proposed .
His reply, he said, was that he agreed to the need for unity in the country now.	the need for the country , in his reply , he said that he was now agreed to unity .

Table 3: Examples of decoded and reordered sentences. All words in the original sentences were retrieved by the model, but the ordering is only perfectly recovered in the first case.

edness: in spite of encoding extremely minimal co-occurrence information, they manage to make for two thirds of W2V's performance, trained on the same data with a much larger window and a complex algorithm (see $\rho = 0.48$ for the bigram model vs $\rho = 0.72$ for W2V-BNC). So relatedness, the flagship task of DS, seems to be present in the most basic structures of language use, although in moderate amount.

The result of the bigram model on sentence relatedness is consistent with its performance at the word level. The improved result obtained by filtering out frequent words, though, reminds us that logical terms are perhaps not so amenable to the distributional hypothesis, despite indications to the contrary (Abrusán et al., 2018).

As for sentence autoencoding, the excellent results of the bigram model might at first be considered trivial and due to the dimensionality of the space, much larger for the bigram model than for W2V. Indeed, at the bag-of-words level, sentence reconstruction can in principle be perfectly achieved by having a space of the dimensionality of the vocabulary, with each word symbolically expressed as a one-hot vector.⁴ However,

⁴To make this clear, if we have a vocabulary $V =$

as noted in §2, the ability to encode relatedness is at odds with the ability to distinguish between meanings. There is a trade-off between having a high-dimensionality space (which allows for more discrimination between vectors and thus easier reconstruction – see White et al., 2016b) and capturing latent features between concepts (which is typically better achieved with lower dimensionality). Interestingly, bigrams seem to be biased towards more symbolic representations, generating representations that distinguish very well between word meanings, but they do also encapsulate a reasonable amount of lexical information. This makes them somewhat of a hybrid constituent, between proper symbols and continuous vectors.

6 Conclusion

So what can be said about bigrams as distributional structure? They encode a very high level of lexical discrimination while accounting for some basic semantic similarity. They of course also encode minimal sequential information which can be used to retrieve local sentence ordering. Essentially, they result in representations that are perhaps more ‘symbolic’ than continuous. It is important to note that the reasonable correlations obtained on relatedness tasks were achieved *after* application of PMI weighting, implying that the raw structure requires some minimal preprocessing to generate lexical information.

On the back of our results, we can draw a few conclusions with respect to the relation of performance and competence at the level of bigrams. Performance data alone produces very distinct word representations without any further processing. Some traces of lexical semantics are present, but require some hard-encoded preprocessing step in the shape of the PMI function. We conclude from this that as a constituent involved in acquisition, the bigram is mostly a marker of the uniqueness of word meaning. Interestingly, we note that the notion of contrast (words that differ in form differ in meaning) is an early feature of children’s language acquisition (Clark, 1988). The fact that it is encoded in one of the most simple structures in language is perhaps no coincidence.

In future work, we plan a more encompassing study of other linguistic components. Crucially,

$\{cat, dog, run\}$ and we define $cat = [100]$, $dog = [010]$ and $run = [001]$, then, trivially, $[011]$ corresponds to the bag-of-words $\{dog, run\}$.

we will also investigate which aspects of state-of-the-art models such as W2V contribute to score improvement on lexical aspects of semantics. We hope to thus gain insights into the specific cognitive processes required to bridge the gap between raw distributional structure as it is found in corpora, and actual speaker competence.

References

- Márta Abrusán, Nicholas Asher, and Tim Van de Cruys. 2018. Content vs. function words: The view from distributional semantics. In *Proceedings of Sinn und Bedeutung 22*.
- Yossi Adi, Einat Kermany, Yonatan Belinkov, Ofer Lavi, and Yoav Goldberg. 2017. Fine-grained analysis of sentence embeddings using auxiliary prediction tasks. International Conference on Learning Representations (ICLR), Toulon, France.
- Marco Baroni, Georgiana Dinu, and Germán Kruszewski. 2014. Don’t count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors. In *ACL (1)*, pages 238–247.
- Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.
- Samuel R. Bowman, Luke Vilnis, Oriol Vinyals, Andrew Dai, Rafal Jozefowicz, and Samy Bengio. 2016. [Generating sentences from a continuous space](#). In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 10–21. Association for Computational Linguistics.
- E. Bruni, N. K. Tran, and M. Baroni. 2014. [Multi-modal distributional semantics](#). *Journal of Artificial Intelligence Research*, 49:1–47.
- John A Bullinaria and Joseph P Levy. 2007. [Extracting semantic representations from word co-occurrence statistics: A computational study: A computational study](#). *Behavior Research Methods*, 39(3):510–526.
- Noam Chomsky. 1965. *Aspects of the theory of syntax*. MIT Press.
- Kenneth Ward Church and Patrick Hanks. 1990. Word association norms, mutual information, and lexicography. *Computational linguistics*, 16(1):22–29.
- Eve V Clark. 1988. On the logic of contrast. *Journal of Child Language*, 15(2):317–335.
- Stephen Clark. 2012. Vector space models of lexical meaning. In Shalom Lappin and Chris Fox, editors, *Handbook of Contemporary Semantics – second edition*. Wiley-Blackwell.

- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
- Georgiana Dinu and Marco Baroni. 2014. How to make words with vectors: Phrase generation in distributional semantics. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 624–633.
- Katrin Erk. 2012. Vector space models of word meaning and phrase meaning: a survey. *Language and Linguistics Compass*, 6:635–653.
- Katrin Erk and Sebastian Padó. 2008. A structured vector space model for word meaning in context. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing (EMNLP2008)*, pages 897–906, Honolulu, HI.
- Katrin Erk, Sebastian Padó, and Ulrike Padó. 2010. A flexible, corpus-driven model of regular and inverse selectional preferences. *Computational Linguistics*, 36(4):723–763.
- John Rupert Firth. 1957. *A synopsis of linguistic theory, 1930–1955*. Philological Society, Oxford.
- Alex Gittens, Dimitris Achlioptas, and Michael W Mahoney. 2017. Skip-gram- zipf+ uniform= vector additivity. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 69–76.
- Zelig Harris. 1954. Distributional structure. *Word*, 10(2-3):146–162.
- Mohit Iyyer, Jordan Boyd-Graber, and Hal Daumé III. 2014. Generating sentences from semantic vector space representations. In *NIPS Workshop on Learning Semantics*.
- Henry Kučera and Winthrop Nelson Francis. 1967. *Computational analysis of present-day American English*. Dartmouth Publishing Group.
- Marco Marelli, Stefano Menini, Marco Baroni, Luisa Bentivogli, Raffaella Bernardi, and Roberto Zamparelli. 2014. A sick cure for the evaluation of compositional distributional semantic models. In *LREC*, pages 216–223.
- Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013. Linguistic regularities in continuous space word representations. In *HLT-NAACL*, pages 746–751.
- Denis Paperno and Marco Baroni. 2016. When the whole is less than the sum of its parts: How composition affects pmi values in distributional semantic vectors. *Computational Linguistics*, 42(2):345–350.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. *Bleu: A method for automatic evaluation of machine translation*. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 311–318, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.
- Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237.
- Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training.
- Radim Řehůřek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta. ELRA. <http://is.muni.cz/publication/884893/en>.
- Allen Schmalz, Alexander M. Rush, and Stuart Shieber. 2016. *Word ordering without syntax*. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2319–2324, Austin, Texas. Association for Computational Linguistics.
- S. Thater, H. Fürstenu, and M. Pinkal. 2011. Word meaning in context: A simple and effective vector model. In *Proceedings of the 5th International Joint Conference on Natural Language Processing*, Chiang Mai, Thailand.
- Peter D. Turney. 2014. *Semantic composition and decomposition: From recognition to generation*. *CoRR*, abs/1405.7908.
- Peter D. Turney and Patrick Pantel. 2010. From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*, 37:141–188.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.
- R. Vedantam, C. L. Zitnick, and D. Parikh. 2015. *Cider: Consensus-based image description evaluation*. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4566–4575.

L. White, R. Togneri, W. Liu, and M. Bennamoun. 2016a. [Modelling sentence generation from sum of word embedding vectors as a mixed integer programming problem](#). In *2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW)*, pages 770–777.

Lyndon White, Roberto Togneri, Wei Liu, and Mohammed Bennamoun. 2016b. Generating bags of words from the sums of their word embeddings. In *17th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing)*.

Jointly Learning to See, Ask, *Decide when to Stop*, and then GuessWhat

Ravi Shekhar[†], Alberto Testoni[†], Raquel Fernández^{*} and Raffaella Bernardi[†]

[†]University of Trento, ^{*}University of Amsterdam

ravi.shekhar@unitn.it alberto.testoni@unitn.it

raquel.fernandez@uva.nl raffaella.bernardi@unitn.it

Abstract

We augment a task-oriented visual dialogue model with a decision-making module that decides which action needs to be performed next given the current dialogue state, i.e. whether to ask a follow-up question or stop the dialogue. We show that, on the *GuessWhat?! game*, the new module enables the agent to succeed at the game with shorter and hence less error-prone dialogues, despite a slightly decrease in task accuracy. We argue that both dialogue quality and task accuracy are essential features to evaluate dialogue systems.¹

1 Introduction

The development of conversational agents that ground language in visual information is a challenging problem that requires the integration of dialogue management skills with multimodal understanding. A common test-bed to make progress in this area are guessing tasks where two dialogue participants interact with the goal of letting one of them guess a visual target (Das et al., 2017a; de Vries et al., 2017; Das et al., 2017b). We focus on the *GuessWhat?! game*, which consists in guessing a target object within an image which is visible to both participants. One participant (the Questioner) is tasked with identifying the target object by asking yes-no questions to the other participant (the Oracle), who is the only one who knows the target. Participants are free to go on with the task for as many turns as required.

Most models of the Questioner agent in the *GuessWhat?! game* consist of two disconnected modules, a Question Generator and a Guesser, which are trained independently with Supervised

Learning or Reinforcement Learning (de Vries et al., 2017; Strub et al., 2017). In contrast, Shekhar et al. (2019) model these two modules jointly. They show that thanks to its joint architecture, their Questioner model leads to dialogues with higher linguistic quality in terms of richness of the vocabulary and variability of the questions, while reaching a performance similar to the state of the art with Reinforcement Learning. They argue that achieving high task success is not the only criterion by which a visually-grounded conversational agent should be judged. Crucially, the dialogue should be coherent, with no unnatural repetitions nor irrelevant questions. We claim that to achieve this, a conversational agent needs to learn a strategy to decide how to respond at each dialogue turn, based on the dialogue history and the current context. In particular, the Questioner model has to learn when it has gathered enough information and it is therefore ready to guess the target.

In this work, we extend the joint Questioner architecture proposed by Shekhar et al. (2019) with a decision-making component that decides whether to ask a follow-up question to identify the target referent, or to stop the conversation to make a guess. Shekhar et al. (2018) had added a similar module to the baseline architecture by de Vries et al. (2017). Here we show that the novel joint architecture by Shekhar et al. (2019) can also be augmented with a decision-making component and that this addition leads to further improvements in the quality of the dialogues. Our extended Questioner agent reaches a task success comparable to Shekhar et al. (2019), but it asks fewer questions, thus significantly reducing the number of games with repetitions.

¹Copyright © 2019 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

2 Task and Models

2.1 Task

The *GuessWhat?!* dataset² was collected via Amazon Mechanical Turk by de Vries et al. (2017). The task involves two human participants who see a real-world image, taken from the MS-COCO dataset (Lin et al., 2014). One of the participants (the Oracle) is assigned a target object in the image and the other participant (the Questioner) has to guess it by asking Yes/No questions to the Oracle. There are no time constraints to play the game. Once the Questioner is ready to make a guess, the list of candidate objects is provided and the game is considered successful if the Questioner picks the target object. The dataset consists of around 155k English dialogues about approximately 66k different images. Dialogues contain on average 5.2 questions-answer pairs.

We use the same train (70%), validation (15%), and test (15%) splits as de Vries et al. (2017). The test set contains new images not seen during training. Following Shekhar et al. (2019), we use two experimental setups for the number of questions to be asked by the Questioner, motivated by prior work: 5 questions (5Q) as de Vries et al. (2017), and 8 questions (8Q) as Strub et al. (2017).

2.2 Models

We focus on developing a Questioner agent able to decide when it has asked enough information to identify the target object. We first describe the baseline model proposed by de Vries et al. (2017). Then we describe the model proposed by Shekhar et al. (2019) and extend it with a decision making module.

Baseline de Vries et al. (2017) model the Questioner agent of the *GuessWhat?!* game as two disjoint models a Question Generator (QGen) and a Guesser trained independently. After a fixed number of questions by QGen, the Guesser selects a candidate object.

QGen is implemented as a Recurrent Neural Network (RNN) with a transition function handled with Long-Short-Term Memory (LSTM), on which a probabilistic sequence model is built with a Softmax classifier. Given the overall image (encoded by extracting its VGG features) and the current dialogue history (i.e., the previous sequence

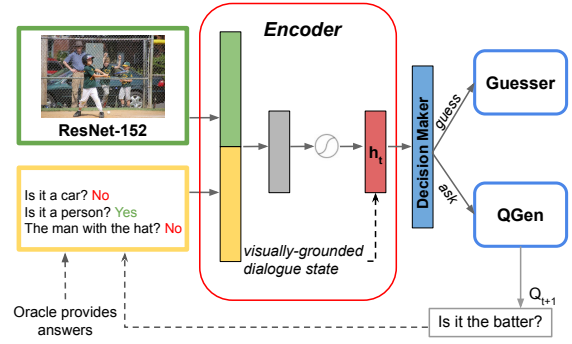


Figure 1: Proposed visually-grounded dialogue state encoder with a decision-making component.

of questions and answers), QGen produces a representation of the visually grounded dialogue (the RNN’s hidden state QH_{t-1} at time $t - 1$ in the dialogue) that encodes information useful to generate the next question q_t . The best performing model of the Guesser by de Vries et al. (2017) represents candidate objects by their object category and spatial coordinates. These features are passed through a Multi-Layer Perceptron (MLP) to get an embedding for each object. The Guesser also takes as input the dialogue history processed by an LSTM, whose hidden state GH_{t-1} is of the same size as the MLP output. A dot product between both returns a score for each candidate object in the image.

Shekhar et al. (2018) extend the baseline architecture of de Vries et al. (2017) with a third model, a decision-making component that determines, after each question/answer pair, whether the QGen model should ask another question or whether the Guesser model should guess the target object.

Grounded Dialogue State Encoder (GDSE)

Shekhar et al. (2019) address one of the fundamental weakness of the Questioner model by de Vries et al. (2017), i.e., having two disconnected QGen and Guesser modules. They tackle this issue with a multi-task approach, where a common visually-grounded dialogue state encoder (GDSE) is used to generate questions and guess the target object. Two learning paradigms are explored: supervised learning (SL) and co-operative learning (CL). In SL, the Questioner model is trained using human data. While in CL, the Questioner model is trained on both generated and human data. First, the Guesser is trained on the generated questions and answers and then the QGen is “readapted” using the human data. Their results show that

²Available at <https://guesswhat.ai/download>.

training these two modules jointly improves the performance of the Questioner model, reaching a task success comparable to RL-based approaches (Strub et al., 2017).

Adding a Decision Making module (GDSE-DM) We extend the GDSE model of Shekhar et al. (2019) with a decision-making component (DM). The DM determines whether QGen should ask a follow-up question or the Guesser should guess the target object, based on the image and dialogue history. As shown in Figure 1, the DM component is modelled as a binary classifier that uses the visually-grounded dialogue state h_t to decide whether to *ask* or *guess*. It is implemented by a Multi Layer Perceptron (MLP_d) trained together with the encoder with negative log-likelihood loss:

$$\mathcal{L}_D = -\log p(\text{dec}_{label}) \quad (1)$$

where dec_{label} is the decision label, i.e., ‘ask’ or ‘guess’. The MLP_d consists of three hidden layers whose dimensions are 256, 64, and 16, respectively; after each hidden layer a ReLU non-linearity is applied.

To train the DM, we need decision labels. For the SL setting, we follow the label generation procedure introduced by Shekhar et al. (2018): decision labels are generated by annotating all the last question-answer pairs in the games with *guess* and all other question-answer pairs as *ask*. For the CL setting, we label the question/answer pairs based on whether the Guesser module is able to correctly predict the target object given the current dialogue. If the Guesser module is able to make a correct prediction after a given question/answer pair, we label that dialogue state with *guess* and otherwise with *ask*. This process results in an unbalanced dataset for the DM where the guess label makes up for only 20% of states. We address this class imbalance by adding a weighing factor, α , to the loss. The balanced loss is given by

$$\mathcal{L}_D = \alpha_{label} \cdot (-\log p(\text{dec}_{label})) \quad (2)$$

where $\alpha_{guess} = 0.8$ and $\alpha_{ask} = 0.2$.

The DM, for both SL and CL, is trained with Cross Entropy loss in a supervised manner using decision labels after each question/answer pair. During inference, the model continues to ask questions unless the DM chooses to end the conversation or the maximum number of questions has been reached. The GDSE-DM model trained with

Model	5Q	8Q
Baseline	41.2	40.7
GDSE-SL	47.8	49.7
GDSE-CL	53.7 (± 8.3)	58.4 (± 1.2)
GDSE-SL-DM	46.78	49.12
GDSE-CL-DM	49.77(± 1.16)	53.89($\pm .24$)

Table 1: Test set accuracy for each model (for setups with 5 and 8 questions).

SL and CL will be referred to as SL-DM and CL-DM, respectively. It has to be highlighted that the tasks of generating a question and guessing the target object are not equally challenging: while the Guesser has to learn the probability distribution of the set of possible objects in the image, QGen needs to fit the distribution of natural language words, which is a much harder task. As in Shekhar et al. (2019), we address this issue by making the learning schedule task-dependent using a *modulo-n* training setup. In the SL setting, n indicates after how many epochs of QGen training the Guesser is updated together with QGen; for CL, QGen is updated at every n^{th} epoch, while the Guesser is updated at all other epochs. We found the optimal value of n to be equal to 5 for both the SL and the CL setting. The models are trained for 100 epochs with Adam optimizer and a learning rate of 0.0001 and we select the Questioner module with the best performance on the validation set.

3 Results

In this section, we report the task success accuracy of our GDSE-DM model, which extends the joint GDSE architecture with a decision-making component. Following Shekhar et al. (2019), to neutralize the effect of random sampling in CL training, we use 3 runs and report mean and standard deviation.

Table 1 gives an overview of the accuracy results obtained by the models. Our main goal is to show the effect of adding a DM module to the joint GDSE architecture. We therefore do not compare to other approaches that use RL.³ As we can see, adding a DM to the GDSE model decreases its accuracy by 0.5-1% in the supervised learning setting and by 4-5% in the cooperative learning set-

³For completeness, the RL model by Strub et al. (2017) has accuracy 56.2(± 24) and 56.3($\pm .05$) for the 5Q and 8Q settings, respectively.

Model	5Q	8Q
GDSE-SL-DM	3.83	5.49
GDSE-CL-DM	4.02(± 0.10)	5.46(± 0.10)

Table 2: Average number of questions asked by the GDSE-DM models when the maximum number of questions is set to 5 or 8.

ting. We believe that the higher drop in accuracy of the CL-DM model can be attributed to the decision labels used by this model. In the SL-DM setting, the model is trained on human data, which leads to a more reliable decision label. In contrast, in the CL-DM setting, the model is trained on automatically generated data, which includes possible errors by both the QGen and the Oracle. Overall, this results in more noisy dialogues. We think that, due to the accumulation of these errors, the decision labels of the generated dialogue deviate significantly from the human data and thus the DM fails to capture them.

Despite the drop in task success, the DM agent seems to be more efficient. Table 2 shows that the average number of questions asked by the DM-based models is lower: the GDSE model without a DM always asks the maximum number of questions allowed (either 5 or 8 questions); while, on average, the GDSE-DM agent asks around 3.8 to 5.5 questions, even when it is allowed to ask up to 8. As we shall see in the next section, this leads to dialogues that are more natural and less repetitive.

4 Analysis

In this section, we look into the advantage brought about by the DM in terms of the quality of the dialogues produced by the model.

Following Shekhar et al. (2019), we report statistics about the dialogue produced by the models with respect to lexical diversity (measured as type/token ratio over all games), question diversity (measured as the percentage of unique questions over all games), and percentage of games with at least one repeated question (see Table 3). The main drawback of the models asking a fixed number of questions is that they repeat questions within the same dialogue. While the introduction of the joint GDSE architecture by Shekhar et al. (2019) substantially reduced the percentage of games with repeated questions with respect to the baseline model (from 93.5% to 52.16%), more


	Lexical diversity	Question diversity	% Games with repeated Q's
Baseline	0.030	1.60	93.50
GDSE-SL	0.101	13.61	55.80
GDSE-CL	0.115 (± 0.02)	14.15 (± 3.0)	52.19 (± 4.7)
GDSE-SL-DM	0.047	1.62	42.47
GDSE-CL-DM	0.135(± 0.02)	10.25(± 2.46)	32.51(± 6.45)
Humans	0.731	47.89	—

Table 3: Statistics of the linguistic output of all models with the 8Q setting compared to human dialogues in all test games.

than 50% of dialogues included repetitions, which make them unnatural. We can see how adding a DM component to GDSE addresses this important problem: with the CL-DM setting, the percentage of games with repeated questions goes down to 32.51% (-19.68%, from 52.19 to 32.51). The reduction is also substantial for the SL-DM model (-13.33%, from 55.80 to 42.47) albeit less impressive.

Given that the number of questions asked by the DM-based models is lower (as shown in Table 2), it is to be expected that the lexical and question diversity of the dialogues produced by these models will also be somewhat lower. Indeed, we observe a rather significant drop in diversity for the SL-DM setting. The CL-DM model, on the other hand, is rather robust to diversity loss: in fact, lexical diversity increases slightly with respect to GDSE-CL (0.135 vs. 0.115 on average), while question diversity decreases by a couple of points only, remaining much higher than that of the baseline model.

Following Shekhar et al. (2019), we also looked into the distribution of the types of questions asked by the models. Questions are divided into two broad categories; ENTITY (about the target object category, e.g., ‘*is it a animal?*’) and ATTRIBUTE (about the target object property, e.g., ‘*is it green one?*’). ENTITY questions are further sub-divided into ‘object category’ and ‘super-category’. ATTRIBUTE questions are divided into ‘color’, ‘shape’, ‘size’, ‘texture’, ‘location’, and ‘action’ questions. Table 4 provides distribution of questions by different models. Compared with their counterparts, the DM-based models ask more object questions. The SL-DM also lowers significantly the number of location questions (from 37.09 to 21.70), which are the type of question most commonly repeated by the various models, as shown by Shekhar et al. (2019). We also computed the Kullback-Leibler

	Baseline [failure]		GDSE-SL [success]		GDSE-CL [success]	
	1. is it a person? no		1. is it a person? no		1. is it a person? no	
	2. is it a tree? no		2. is it a ski? yes		2. is it a tree? no	
	3. is it a tree? no		3. is it on the person in the front? yes		3. is it a ski? yes	
	4. is it a tree? no		DM stops asking to guess		4. is it both skis of the person in front? yes	
	5. is it a tree? no		4. is it the left most person? yes		DM stops asking to guess	
	6. is it a tree? no		5. both skis? yes		5. is it on the left? yes	
	7. is it a tree? no		6. both skis? yes		6. is it on the left? yes	
	8. is it a tree? no		7. the whole skis? yes		7. it is second to the right? no	
			8. the whole skis? yes		8. is it the one on the left? yes	

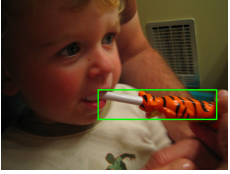
	Baseline [success]		GDSE-SL [success]		GDSE-CL [success]	
	1. is it the baby? no		1. is it a person? no		1. is it a person? no	
	2. is it the toothbrush? yes		2. is it a toothbrush? yes		2. is it a toothbrush? yes	
	3. is it the toothbrush? yes		3. is it the whole thing? yes		DM stops asking to guess	
	4. is it the toothbrush? yes		DM stops asking to guess		3. is it the one kid holding? yes	
	5. is it the toothbrush? yes		4. the one in the hand? yes		4. is it the one left of the baby? no	
	6. is it the toothbrush? yes		5. the one in the front? yes		5. is it next to the one with the red handle? yes	
	7. is it the toothbrush? yes		6. the whole thing? yes		6. is it next to the baby? no	
	8. is it the entire toothbrush? yes		7. the whole of it? yes		7. is it one next to the baby? no	
			8. all of it? yes		8. next to the whole object? yes	

Figure 2: Game examples where most models succeed at guessing the target object (framed). In red, the point in the dialogue where the DM component decides to stop asking questions and guess the target. Many of the questions asked after the decision point by the GDSE model without a DM are repeated, and thus do not add any extra information.

(KL) divergence to measure how the output of each model differs from the human distribution of fine-grained question classes. We can see that GDSE-DM models have comparatively higher degree of divergence than GDSE, in particular the SL-DM model, which asks a substantially larger proportion of ENTITY questions.

The sample dialogues in Figure 2 provide a qualitative illustration of the output of our models, showing how the DM-based Questioner stops asking questions when it has enough information to guess the target object.

Question type	BL	SL	CL	SL-DM	CL-DM	H
ENTITY	49.00	48.07	46.51	71.03	51.36	38.11
SUPER-CAT	19.6	12.38	12.58	15.35	15.40	14.51
OBJECT	29.4	35.70	33.92	55.68	35.97	23.61
ATTRIBUTE	49.88	46.64	47.60	27.27	45.21	53.29
COLOR	2.75	13.00	12.51	10.57	8.41	15.50
SHAPE	0.00	0.01	0.02	0.0	0.07	0.30
SIZE	0.02	0.33	0.39	0.01	0.67	1.38
TEXTURE	0.00	0.13	0.15	0.01	0.25	0.89
LOCATION	47.25	37.09	38.54	21.70	39.92	40.00
ACTION	1.34	7.97	7.60	3.96	8.01	7.59
Not classified	1.12	5.28	5.90	1.70	3.43	8.60
KL wrt Human	0.953	0.042	0.038	1.48	0.055	—

Table 4: Percentage of questions per question type in all the test set games played by humans (H) and the models with the 8Q setting, and KL divergence from human distribution of fine-grained question types.

5 Conclusion

We have enriched the Questioner agent in the goal-oriented dialogue game *GuessWhat?!* with a Decision Making (DM) component. Based on the visually grounded dialogue state, our Questioner model learns whether to ask a follow-up question or to stop the conversation to guess the target object. We show that the dialogue produced by our model has less repetitions and less unnecessary questions, thus potentially leading to more efficient and less unnatural interactions – a well known limitation of current visual dialogue systems. As in Shekhar et al. (2018), where a simple baseline model was extended with a DM component, task accuracy slightly decreases while the quality of the dialogues increases.

A first attempt to partially tackle the issue within the *GuessWhat?!* game was made by Strub et al. (2017), who added a <stop> token to the vocabulary of the question generator module to learn when to stop asking questions using Reinforcement Learning. This is a problematic approach as it requires the QGen to generate probabilities over a non-linguistic token; further, the decision to ask more questions or guess is a binary decision and thus it is not desirable to incorporate it within the large softmax output of the QGen.

Jiaping et al. (2018) propose a hierarchical RL-based Questioner model for the *GuessWhich* image-guessing game introduced by Chattopad-

hyay et al. (2017) using the *VisDial* dataset (Das et al., 2017a). The first RL layer is a module that learns to decide when to stop asking questions. We believe that a decision making component for the *GuessWhich* game is an ill-posed problem. In this game, the Questioner does not see the pool of candidate images while carrying out the dialogue. Hence, it will never know when it has gathered enough information to distinguish the target image from the distractors. In any case, our work shows that a simple approach can be used to augment visually-grounded dialogue systems with a DM without having to use the high complexity of RL paradigms.

Task accuracy and dialogue quality are equally important aspects of visually-grounded dialogue systems. It remains to be seen how such systems can reach higher task accuracy while profiting from the better quality that DM-based models produce.

References

- Prithvijit Chattopadhyay, Deshraj Yadav, Viraj Prabhu, Arjun Chandrasekaran, Abhishek Das, Stefan Lee, Dhruv Batra, and Devi Parikh. 2017. Evaluating visual conversational agents via cooperative human-ai games. In *Proceedings of the Fifth AAAI Conference on Human Computation and Crowdsourcing (HCOMP)*.
- Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José M.F. Moura, Devi Parikh, and Dhruv Batra. 2017a. Visual Dialog. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.
- Abhishek Das, Satwik Kottur, José M.F. Moura, Stefan Lee, and Dhruv Batra. 2017b. Learning cooperative visual dialog agents with deep reinforcement learning. In *International Conference on Computer Vision (ICCV)*.
- Zhang Jiaping, Zhao Tiancheng, and Yu Zhou. 2018. Multimodal hierarchical reinforcement learning policy for task-oriented visual dialog. In *Proceeding of the SigDial Conference*, pages 140–150. Association for Computational Linguistics.
- Sang-Woo Lee, Yu-Jung Heo, and Byoung-Tak Zhang. 2017. Answerer in questioner’s mind for goal-oriented visual dialogue. In *NIPS Workshop on Visually-Grounded Interaction and Language (ViGIL)*. ArXiv:1802.03881. Last version Feb. 2018.
- T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, Dollar, P., and C. L. Zitnick. 2014. Microsoft COCO: Common objects in context. In *Proceedings of ECCV (European Conference on Computer Vision)*.
- Ravi Shekhar, Tim Baumgärtner, Aashish Venkatesh, Elia Bruni, Raffaella Bernardi, and Raquel Fernández. 2018. Ask no more: Deciding when to guess in referential visual dialogue. In *Proceedings of the 27th International Conference on Computational Linguistics (COLING)*, pages 1218–1233.
- Ravi Shekhar, Aashish Venkatesh, Tim Baumgärtner, Elia Bruni, Barbara Plank, Raffaella Bernardi, and Raquel Fernández. 2019. Beyond task success: A closer look at jointly learning to see, ask, and guesswhat. In *NAACL*.
- Florian Strub, Harm de Vries, Jeremie Mary, Bilal Piot, Aaron Courville, and Olivier Pietquin. 2017. End-to-end optimization of goal-driven and visually grounded dialogue systems. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*.
- Harm de Vries, Florian Strub, Sarath Chandar, Olivier Pietquin, Hugo Larochelle, and Aaron C. Courville. 2017. Guesswhat?! Visual object discovery through multi-modal dialogue. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Junjie Zhang, Qi Wu, Chunhua Shen, Jian Zhang, Jianfeng Lu, and Anton van den Hengel. 2018. Goal-oriented visual question generation via intermediate rewards. In *Proceedings of the European Conference of Computer Vision (ECCV)*.
- Yan Zhu, Shaoting Zhang, and Dimitris Metaxas. 2017. Interactive reinforcement learning for object grounding via self-talking. In *NIPS Workshop on Visually-Grounded Interaction and Language (ViGIL)*.

Creating a Multilingual Terminological Resource using Linked Data: the case of Archaeological Domain in the Italian language

Giulia Speranza, Carola Carlino

UNIOR NLP Research Group
University of Naples “L’Orientale”
Naples, Italy

{gsperanza, ccarlino}@unior.it

Sina Ahmadi

Insight Centre for Data Analytics
National University of Ireland
Ireland, Galway

sina.ahmadi@insight-centre.org

Abstract

English. The lack of multilingual terminological resources in specialized domains constitutes an obstacle to the access and reuse of information. In the technical domain of cultural heritage and, in particular, archaeology, such an obstacle still exists for Italian language. This paper presents an effort to fill this gap by collecting linguistic data using existing Collaboratively-Constructed Resources and those on the Web of linked data. The collected data are then used to linguistically enrich the ICCD Archaeological Finds Thesaurus— a monolingual Italian thesaurus. Our terminological resource contains 446 terms with translations in four languages and is publicly available in the Resource Description Framework (RDF) in the Ontolex-Lemon model.

1 Introduction

Multilingual domain-specific linguistic resources, such as thematic dictionaries and terminological resources (*terminologies* further in the text), are knowledge repositories providing information about terms and their semantic relationships in a specific domain and across languages. Currently, most European languages, including Italian, lack terminologies in the field of cultural heritage (Dong, 2017). With cultural heritage one defines the tangible and intangible objects that constitute the culture of each society such as monuments but also songs, traditions and history (Dorr, 2009).

Given the expanding amount of cultural data on the Semantic Web and a plethora of publicly-available resources in various languages as Linked Open Data (LOD), the Web provides solutions for enhancing multilingualism in terminologies (Brugman et al., 2008). Nowadays, many Collaboratively-Constructed Resources (CCRs), or Collaborative Knowledge Bases (CKBs), such as Wiktionary¹ and Wikipedia², are created by decentralized communities of volunteers in different domains.

CCRs differ from Linguistic Knowledge Bases (LKBs), such as WordNet (Miller, 1995) and FrameNet (Baker et al., 1998), which are instead created by experts in specific fields with higher quality control. Some scholars, such as Müller and Gurevych (2008) and Hovy et al. (2013), pointed out several weaknesses of LKBs such as the low coverage of domain-specific vocabulary, restriction to common vocabulary and the difficulty in continuous maintenance resulting out-dated data.

Moreover, despite the application of CCRs in various natural language processing (NLP) tasks (Zesch et al., 2008; Nakayama et al., 2008; Meyer and Gurevych, 2012), processing heterogeneous and often unstructured data linguistically requires syntactic, lexical and ontological information (Bouayad-Agha et al., 2012; Davies, 2009). This can be efficiently addressed thanks to the current advances in applying computational techniques to the disciplines of the humanities, known as digital humanities (DH), and accessibility of linguistic resources on the Web with movements such as the Linguistic Linked Open Data (LLOD) (Chiarcos et al., 2013).

Regarding the field of cultural heritage, multilingualism is still a challenge due to the tendency of experts to store terminologies monolingually (Vavliakis et al., 2012). We investigated some on-

Copyright ©2019 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

¹<https://www.wiktionary.org/>

²<https://www.wikipedia.org/>

line multilingual terminologies such as the Getty Vocabularies³ (Baca and Gill, 2015) which contains thesauri in art, architecture and cultural objects, iDAI.vocab—the German Archaeological Institute archaeological vocabulary⁴, the *UNESCO Thesaurus*⁵, the European Heritage Network thesauri⁶ and the Loterre Controlled Vocabulary in art and archaeology⁷. Among these resources, only the Art & Architecture Thesaurus (AAT) by Getty and the iDAI.vocab are exploitable due to a partial domain-specific similarity with our dataset; nevertheless, none of them provide lexicographic descriptions of the terms.

In this paper, we propose an approach for semi-automatically creating a multilingual terminology in the technical domain of archaeology and cultural heritage by enriching an existing Italian ontology with linguistic information. Our approach can be applied to any domain and language. Our case study is the archaeological thesaurus provided by the Central Institute for Catalogue and Documentation (ICCD) for describing archaeological finds in Italian (Felicetti et al., 2013). The enriched information are evaluated by annotators, and then converted into the Ontolex-Lemon model in the Resource Description Framework (RDF). Our resource provides linguistic information of 446 Italian terms with translations in four languages.

2 Related Work

Leveraging resources on the Web for extracting and processing information is a common practice in NLP tasks (Lin and Katz, 2003; Cucerzan and Brill, 2004). Previous studies focusing on extracting data from CCRs showed that this is a valuable resource for collecting lexicographic data and promoting multilingualism (Kilgarriff and Grefenstette, 2001; Lin and Krizhanovsky, 2011).

Bourgonje et al. (2016) develop a platform for digital curation technologies using a Semantic Web layer which provides linguistic analysis and discourse information. This platform allows knowledge experts to create digital content and ex-

plore a collection of documents related to a specific domain. Project FREME (Dojchinovski et al., 2016) is a framework for multilingual and semantic enrichment of digital content where linguistic linked open data workflows are used along with linguistic and NLP ontologies. The EuroTermBank project (Vasiljevs et al., 2008) aims at improving the terminology infrastructure of the European languages by creating a centralized online terminology bank and collecting terminologies from various European institutions to facilitate the production, use and distribution of digital content and promote cultural diversity.

Dannélls et al. (2013) also focus on the domain of cultural heritage and use Wikipedia to retrieve translations for the task of text generation. Dong (2017) uses three multilingual semantic resources, GeoNames, DBpedia and Wiktionary, to enrich English information for Chinese Genealogical Linked Data in the field of cultural heritage. Declerck et al. (2012) use Wiktionary to expand a taxonomy of folk catalogue in English with multilingual translations.

Providing terminologies in Linked Data has been also addressed by previous researchers. Cimiano et al. (2015) present an approach for publishing and linking terminological resources using linked data principles. They provide a service for transforming term bases in TBX–TermBase eXchange, an open XML-based standard format for terminological data, to RDF using *lemon* model. Similarly, McCrae et al. (2011) show the conversion of WordNet and Wiktionary data into Lemon model. Sérasset et al. (2015) focused on creating a RDF Lemon-based multilingual resource with data extracted from Wiktionary.

3 Case Study

The dataset used in this study is the Italian ICCD “*RA Thesaurus per la descrizione dei reperti archeologici*” (en. RA Thesaurus for the description of archaeological finds) published by the ICCD (Istituto Centrale per il Catalogo e la Documentazione) in collaboration with the Italian Ministry of Cultural Heritage and Activities (MiBAC). The ICCD Thesaurus (Mancinelli, 2014) is an open monolingual Italian vocabulary (last updated in 2014), which was created with the final aim of regulating the terminology to be used to identify archaeological finds in Italy. In the ICCD Thesaurus different levels for the representation of the

³<https://www.getty.edu/research/tools/vocabularies/>

⁴<https://archwort.dainst.org>

⁵<http://vocabularies.unesco.org/browser/thesaurus/en/>

⁶<https://www.coe.int/en/web/culture-and-heritage/herein-heritage-network>

⁷<https://www.loterre.fr/skosmos/27X/>


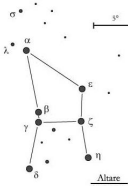
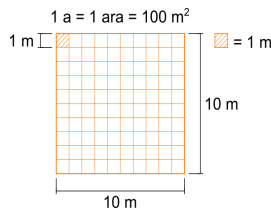

ara (n)			
			
Ornithology (Q44703)	Astronomy (Q333)	Metrology (Q394)	Archaeology (Q23498)

Figure 1: An example of the Italian word *ara* (n) which can appear in various terminological domains.

terms are provided: the first level indicates the object itself, e.g. *colonna* (en. column); other levels refer to the morphology which indicates the type and shape of the object, e.g. *colonna dorica*, (en. doric column), and part which specifies the part of the object, e.g. *base*, *capitello* (en. base, capital). Furthermore, it is enriched with a short description and sometimes images of the object described. The ICCD Thesaurus is published as LOD on a designed platform⁸ and can be accessed through various formats.

Regarding archaeological finds, the Italian terminology in this field is composed of both technical terms and common vocabulary from everyday language. Technical terms may be perceived as more or less technical on a continuum: there are technical terms which might be so frequent, also in the common vocabulary, that their meaning is generally understood by the majority of literate people, e.g. *capitello* (en. capital), *altare* (en. altar), and less frequent terms used and known only by experts in the field, e.g. *acroterio* (en. acroterion), *archivolto* (en. archivolt). On the other hand, many common words are used to describe archaeological finds, e.g. *bottiglia* (en. bottle), *collana* (en. necklace), which, of course, sound more comprehensible also to non-experts.

A jargon, such as the language of archaeology, often reuse already-existing words instead of creating ad hoc new terms, assigning them a different meaning (Gotti, 1991; Scarpa, 2008; Gualdo and Telve, 2011). In fact, several examples of semantic redeterminations were registered in the ICCD Thesaurus such as the word *ghianda* which comes from a common vocabulary, where it has the general meaning of acorn, but, in the specialized domain, is used to identify a particular kind of pro-

jectile weapon, thus acquiring a totally different new meaning. Despite being precise and unique in their terminology, it is not rare to find homographs and polysemous words also in specialized jargons. For example the Italian word *ara* can be found at least in four different domains (ornithology, astronomy, metrology and archaeology) with different meanings but the same written form, as shown in Figure 1.

Furthermore, for the specialized domain of archaeology, many analogies with the anatomical parts of the human body are observed, e.g. column foot and neck-amphora. In linguistics and rhetoric, this phenomenon is a figure of speech called *catachresis*, which is based on mixed metaphoric and metonymic expressions which allow an economic reuse of a previous lexicon.

In order to further specify the morphology or the function of a cultural object, many multi-word expressions (MWEs), mostly composed of Noun+Preposition+Noun, are also used in the Italian terminology, e.g. *altare a mensa*. There are also many compounds such as *semicolonna* and *monoansata* (respectively, half-column and one-handled in English). In addition, a conspicuous part of domain-specific terminology comes both from Greek and Latin words (e.g. *rhyton*, *cingulum*) or presents Greek or Latin prefixoids which contribute to make this specialized lexicon even more difficult to understand and highly technical. Finally, there are also some loan-words such as *menhir* and *applique* which come from Breton and French.

4 Methodology

Given a list of terms in the source dataset, we first retrieve those concepts to which the term is associated on Wikidata, i.e. concepts with `rdfs:label` as a predicate and the term as an

⁸<http://dati.beniculturali.it/>

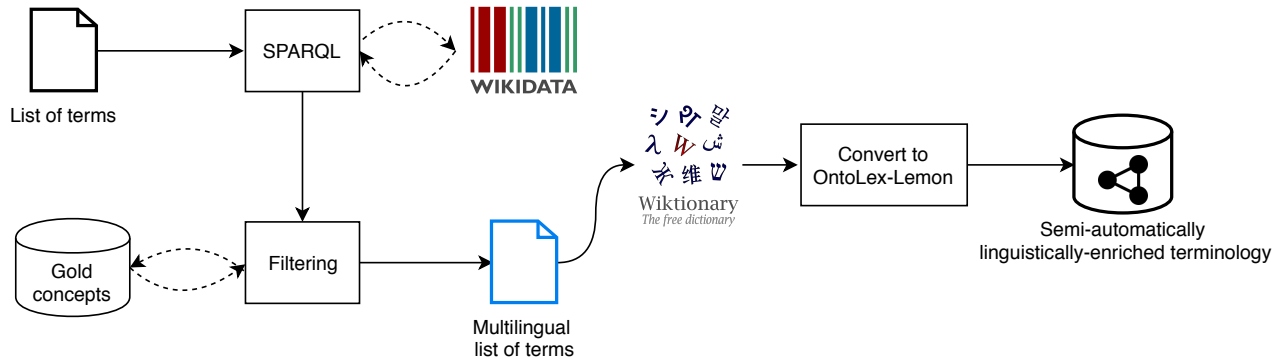


Figure 2: Terminological enrichment process

object as follows:

```
SELECT ?ConceptID {
  ?ConceptID rdfs:label "T"@it.
}
```

where the ID of the concepts associated with the term *T* are returned.

Since a word can be used in various domains with different senses, it is possible to retrieve more than one concept for a term. Therefore, the relevance of the retrieved concepts to our terminological field is examined based on the semantic relationships, such as subclass-of, part-of and instance-of, between the retrieved concepts and those to which we assume that the terms are associated. Such concepts, henceforth referred to as *gold concepts*, are collected based on the knowledge of the experts in the domain and manual collection from Wikidata. The SPARQL query for this verification can be described as follows:

```
ASK {
  wd:ConceptID (wdt:P361|wdt:P279|
    wdt:P31)+ wd:GoldConceptID.
}
```

where `wd:ConceptID` and `wd:GoldConceptID` refer to the ID of the retrieved concepts and the gold concepts, respectively. `P279`, `P361` and `P31` are the Wikipedia properties for subclass-of, part-of and instance-of properties on Wikidata. A list of the gold concepts in the field of archaeology is provided in Appendix A.

Filtering retrieved data from Wikidata enables us to disambiguate the terms based on the concepts. For instance, the Italian word *calice* appears as a label for several concepts such as wine glass, calyx and chalice, to which only the latter is

relevant to our terminological field, therefore selected in this step. Following the collection of the candidate concepts, we retrieve the labels of the concepts in our target languages, namely, English, French, German and Italian. The choice of the languages was dependent on our evaluation means. The retrieved terms are then enriched by linguistic information from Wiktionary. This process is illustrated in Figure 2.

4.1 Conversion to OntoLex-Lemon

In the recent years, there have been efforts to create specific data models providing support for representing linguistic data on the Semantic Web. The OntoLex-Lemon (McCrae et al., 2017) is a model based on the Lexicon Model for Ontologies (lemon) which provides rich linguistic grounding for ontologies, such as representation of morphological and syntactic properties of lexical entries. This model draws heavily on previous lexical data models, particularly LexInfo (Cimiano et al., 2011), LIR (Montiel-Ponsoda et al., 2008) and LMF (Francopoulo et al., 2006), with improvements such as being RDF-native, descriptive and modular justifying its promising adaptability in linguistic resource management.

The previous step yields a tabular format of the lexicographic information, making it possible to convert the data semi-automatically into RDF triples in OntoLex-Lemon. Figure 3 illustrates the equivalent of the Italian entry *ascia* in the output terminology in RDF Turtle in OntoLex-Lemon. In addition to the linguistic information, each entry is linked to the original concept in the source dataset, i.e. ICCD, using the `skos:concept` property. Similarly, the Wikipedia page describing the term is provided using `ontolex:denotes` property.

In addition to OntoLex-Lemon core model, we

```

:lexicon a lime:Lexicon;
lime:entry :ascia ;
lime:language
<http://www.lexvo.org/page/iso639-1/it>.

:ascia a ontolex:LexicalEntry,
ontolex:Word ;
ontolex:canonicalForm :form_ascia ;
rdfs:label "ascia"@it ;
lexinfo:partOfSpeech lexinfo:noun ;
lexinfo:gender lexinfo:feminine .

:form_ascia a ontolex:Form ;
dct:language
<www.lexvo.org/page/iso639-1/it>;
ontolex:writtenRep "ascia"@it ;
lexinfo:number lexinfo:singular ;
ontolex:sense :ascia_n_sense ;
ontolex:denotes wd:Q2517447;
<https://it.wikipedia.org/wiki/Ascia>;
dct:subject wd:Q382995 ;
owl:sameAs dati:009000000004 .

:trans a vartrans:Translation ;
vartrans:source :ascia_n_sense ;
vartrans:target
frl:fr_herminette_sense .

```

Figure 3: The description of the term *ascia* in Ontolex-Lemon

used the following modules:

- Linguistic Metadata (*lime*) to describe metadata at the level of the lexicon-ontology interface with information such as lexical entries and language.
- Syntax and Semantics (*synsem*) enables us to describes syntactic behaviour. We use syntactic frames to relate a lexical entry to one of its various syntactic roles, such as the canonical form of the word *ascia*.
- Lexinfo (*lexinfo*) (Cimiano et al., 2011) for describing relevant linguistic categories and properties, particularly part-of-speech (POS), gender and number.
- Variation and Translation (*vartrans*) is used to describe relations between lexical entries, particularly translations.

Among the 4000 terms provided in the source dataset, i.e. the ICCD Thesaurus, only 446 terms could be retrieved from Wikipedia. This can be due to the technicality of the source dataset which is confined to Italian archaeological finds, therefore describes cultural objects which might not be

present outside Italy. On the hand, Wikidata is constantly being enriched and may had incomplete data when the queries were run. With respect to Wiktionary, among the retrieved terms, 26 terms were available without linguistic descriptions such as part-of-speech (PoS) tags and gender. We observed that the majority of missing terms were of Latin or Greek etymology. As Wiktionary is a Collaboratively-Constructed Resource, a manual verification and completion of the retrieved data was carried out. Some of the erroneous data were due to homographs such as *ancora* and polysemous terms which may belong to more than one grammatical category, such as *piatto* meaning “plate” as a noun while “flat” as an adjective.

5 Conclusion

In this paper, we demonstrated the usage of LOD and CCR in enriching terminological ontologies. As a case study, we used an ontology in Italian in the field of cultural heritage and archaeology to create multilingual terminologies. The results of the manual evaluation and implementation process show that leveraging such resources is a valid option for enriching ontologies linguistically. Nonetheless, since CCRs are created by a community effort, a manual verification was carried out for creating gold-standard datasets.

Finally, the effort of this study can be framed within the more general context of contributing to the implementation and advancement of the multilingual Web of Data and the LLOD movement. The multilingual resource that we are proposing can be used in several professional figures among which lexicographers, translators, museum and exhibition experts, archaeologists and researchers.

Further experiments will concern retrieving MWEs as we have not included them in the current study due to the scarce availability on Wikidata and Wiktionary. MWEs are a topic increasingly handled in NLP, and their processing is fundamental for NLP tasks ranging from POS tagging to Machine Translation to obtain better and more reliable results (Monti et al., 2018). We are also interested in creating gold concepts more efficiently, particularly using topic modelling techniques, and integrating more resources, particularly ConceptNet (Liu and Singh, 2004) which contains many resources such as WordNets and DBpedia.

This project is openly available at <https://github.com/sinaahmadi/sparql4respop>.

Acknowledgments

We want to thank SmartApps for providing useful material and information for the realization of this project. This project has been partially supported by the PON Ricerca e Innovazione 2014/20 and the POR Campania FSE 2014/2020 funds. Sina Ahmadi is also supported by the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015.

References

- Murtha Baca and Melissa Gill. 2015. Encoding multilingual knowledge systems in the digital age: the getty vocabularies. *NASKO*, 42(4):232–243.
- Collin F Baker, Charles J Fillmore, and John B Lowe. 1998. The berkeley framenet project. In *Proceedings of the 17th international conference on Computational linguistics-Volume 1*, pages 86–90. Association for Computational Linguistics.
- Nadjet Bouayad-Agha, Gerard Casamayor, Simon Mille, Marco Rospocher, Horacio Saggion, Luciano Serafini, and Leo Wanner. 2012. From ontology to NL: Generation of multilingual user-oriented environmental reports. In *International Conference on Application of Natural Language to Information Systems*, pages 216–221. Springer.
- Peter Bourgonje, Julian Moreno-Schneider, Jan Nehring, Georg Rehm, Felix Sasaki, and Ankit Srivastava. 2016. Towards a platform for curation technologies: enriching text collections with a Semantic Web layer. In *European Semantic Web Conference*, pages 65–68. Springer.
- Hennie Brugman, Véronique Malaisé, and Laura Hollink. 2008. A common multimedia annotation framework for cross linking cultural heritage digital collections. In *6th International Conference on Language Resources and Evaluation (LREC 2008)*.
- Christian Chiarcos, Philipp Cimiano, Thierry Declerck, and John P McCrae. 2013. Linguistic linked open data (llo). introduction and overview. In *Proceedings of the 2nd Workshop on Linked Data in Linguistics (LDL-2013): Representing and linking lexicons, terminologies and other language data*, pages i–xi.
- Philipp Cimiano, Paul Buitelaar, John McCrae, and Michael Sintek. 2011. Lexinfo: A declarative model for the lexicon-ontology interface. *Web Semantics: Science, Services and Agents on the World Wide Web*, 9(1):29–51.
- Philipp Cimiano, John P McCrae, Víctor Rodríguez-Doncel, Tatiana Gornostay, Asunción Gómez-Pérez, Benjamin Siemoneit, and Andis Lagzdins. 2015. Linked terminologies: applying linked data principles to terminological resources. In *Proceedings of the eLex 2015 Conference*, pages 504–517.
- Silviu Cucerzan and Eric Brill. 2004. Spelling correction as an iterative process that exploits the collective knowledge of web users. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 293–300.
- Dana Dannélls, Aarne Ranta, Ramona Enache, Mariana Damova, and Maria Mateva. 2013. Multilingual access to cultural heritage content on the Semantic Web. In *Proceedings of the 7th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pages 107–115.
- Rob Davies. 2009. Europeanalocal—its role in improving access to Europe's cultural heritage through the European digital library. In *Proceedings of IACH workshop at ECDL2009 (European Conference on Digital Libraries)*, Aarhus, September.
- Thierry Declerck, Karlheinz Mörth, and Piroska Lendvai. 2012. Accessing and standardizing wiktionary lexical entries for supporting the translation of labels in taxonomies for digital humanities. In *Proceedings of LREC*.
- Martin Doerr. 2009. Ontologies for cultural heritage. In *Handbook on ontologies*, pages 463–486. Springer.
- Milan Dojchinovski, Felix Sasaki, Tatjana Gornostaja, Sebastian Hellmann, Erik Mannens, Frank Salliau, Michele Osella, Phil Ritchie, Giannis Stoitsis, Kevin Koidl, Markus Ackermann, and Nilesch Chakraborty. 2016. FREME: Multilingual semantic enrichment with linked data and language technologies. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 4180–4183, Portorož, Slovenia, May. European Language Resources Association (ELRA).
- Hang Dong. 2017. Enrichment of cross-lingual information on Chinese genealogical Linked Data. *iConference 2017 Proceedings Vol. 2*.
- Achille Felicetti, Tiziana Scarselli, Maria Letizia Mancinelli, and Franco Niccolucci. 2013. Mapping ICCD archaeological data to CIDOC-CRM: the RA schema. *A Mapping of CIDOC CRM Events to German Wordnet for Event Detection in Texts*, 11.
- Gil Francopoulo, Monte George, Nicoletta Calzolari, Monica Monachini, Nuria Bel, Mandy Pet, and Claudia Soria. 2006. Lexical markup framework (LMF). In *International Conference on Language Resources and Evaluation-LREC 2006*, page 5.
- Maurizio Gotti. 1991. *I linguaggi specialistici: caratteristiche linguistiche e criteri pragmatici*. La Nuova Italia.
- Riccardo Gualdo and Stefano Telve. 2011. *Linguaggi specialistici dell'italiano*. Carocci.
- Eduard Hovy, Roberto Navigli, and Simone Paolo Ponzetto. 2013. Collaboratively built semi-structured content and artificial intelligence: The story so far. *Artificial Intelligence*, 194:2–27.
- Adam Kilgarriff and Gregory Grefenstette. 2001. Web as corpus. In *Proceedings of Corpus Linguistics 2001*, pages 342–344. Corpus Linguistics. Readings in a Widening Discipline.
- Jimmy Lin and Boris Katz. 2003. Question answering from the web using knowledge annotation and knowledge mining techniques. In *Proceedings of the twelfth international conference on Information and knowledge management*, pages 116–123. ACM.
- Feiyu Lin and Andrew Krizhanovsky. 2011. Multilingual ontology matching based on Wiktionary data accessible via SPARQL endpoint. *arXiv preprint arXiv:1109.0732*.
- Hugo Liu and Push Singh. 2004. Conceptnet practical commonsense reasoning tool-kit. *BT technology journal*, 22(4):211–226.

- Maria Letizia Mancinelli. 2014. Strumenti terminologici. Scheda RA. reperti archeologici. thesaurus per la definizione del bene. introduzione e indicazioni per l'uso. ICCD - Servizio beni archeologici.
- John McCrae, Dennis Spohr, and Philipp Cimiano. 2011. Linking lexical resources and ontologies on the semantic web with Lemon. In *Extended Semantic Web Conference*, pages 245–259. Springer.
- John P McCrae, Julia Bosque-Gil, Jorge Gracia, Paul Buitelaar, and Philipp Cimiano. 2017. The Ontolex-Lemon model: development and applications. In *Proceedings of eLex 2017 conference*, pages 19–21.
- Christian M Meyer and Iryna Gurevych. 2012. *Wiktionary: A new rival for expert-built lexicons? Exploring the possibilities of collaborative lexicography*. na.
- George A Miller. 1995. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.
- Johanna Monti, Ruslan Mitkov, Violeta Seretan, and Gloria Corpas Pastor. 2018. Multiword units in machine translation and translation technology. In Ruslan Mitkov, Johanna Monti, Violeta Seretan, and Gloria Corpas Pastor, editors, *Multiword units in machine translation and translation technology*, pages 1–38. John Benjamins Publishing Company.
- Elena Montiel-Ponsoda, Guadalupe Aguado De Cea, Asunción Gómez-Pérez, and Wim Peters. 2008. Modelling multilinguality in ontologies. *Coling 2008: Companion volume: Posters*, pages 67–70.
- Christof Müller and Iryna Gurevych. 2008. Using wikipedia and wiktionary in domain-specific information retrieval. In *Workshop of the Cross-Language Evaluation Forum for European Languages*, pages 219–226. Springer.
- Kotaro Nakayama, Minghua Pei, Maike Erdmann, Masahiro Ito, Masumi Shirakawa, Takahiro Hara, and Shojiro Nishio. 2008. Wikipedia mining wikipedia as a corpus for knowledge extraction.
- Federica Scarpa. 2008. *La traduzione specializzata. Un approccio didattico professionale*. Milano: Hoepli, 2nd edition.
- Gilles Sérasset. 2015. Dbnary: Wiktionary as a lemon-based multilingual lexical resource in RDF. *Semantic Web*, 6(4):355–361.
- Andrejs Vasiljevs, Signe Rirdance, and Andris Liedskalnins. 2008. Eurotermbank: Towards greater interoperability of dispersed multilingual terminology data. In *Proceedings of the First International Conference on Global Interoperability for Language Resources ICGL*, pages 213–220.
- Konstantinos N Vavliakis, Georgios Th Karagiannis, and Pericles A Mitkas. 2012. Semantic Web in cultural heritage after 2020. In *Proceedings of the 11th International Semantic Web Conference (ISWC), Boston, MA, USA*, pages 11–15.
- Torsten Zesch, Christof Müller, and Iryna Gurevych. 2008. Extracting lexical semantic knowledge from Wikipedia and Wiktionary. In *LREC*, volume 8, pages 1646–1652.

Appendix A

architecture	Q12271
archaeology	Q10855079
artificial physical object	Q8205328
art	Q735
archaeological artifact	Q220659
architectural element	Q391414
architectural order	Q217175
container	Q987767
vase	Q191851
clothing in ancient Greece	Q522648
clothing in ancient Rome	Q2457980
tool	Q39546
roof tile	Q268547
religious object	Q21029893
visual artwork	Q4502142
costume accessory	Q1065579
sculpture	Q860861
religious object	Q21029893
accessory	Q362200
building component	Q19603939
bijou	Q3575260

Table 1: Concepts used for disambiguation of Wikidata concepts (*gold concepts*)

Vir is to Moderatus as Mulier is to Intemperans Lemma Embeddings for Latin

Rachele Sprugnoli, Marco Passarotti, Giovanni Moretti
CIRCSE Research Centre, Università Cattolica del Sacro Cuore
Largo Agostino Gemelli 1, 20123 Milano

{rachele.sprugnoli, marco.passarotti, giovanni.moretti}@unicatt.it

Abstract

English. This paper presents a new set of lemma embeddings for the Latin language. Embeddings are trained on a manually annotated corpus of texts belonging to the Classical era: different models, architectures and dimensions are tested and evaluated using a novel benchmark for the synonym selection task. A qualitative evaluation is also performed on the embeddings of rare lemmas. In addition, we release vectors pre-trained on the “Opera Maiora” by Thomas Aquinas, thus providing a resource to analyze Latin in a diachronic perspective.¹

1 Introduction

Any study of the ancient world is inextricably bound to empirical sources, be those archaeological relics, artifacts or texts. Most ancient texts are written in dead languages, one of the distinguishing features of which is that both their lexicon and their textual evidence are essentially closed, without any new substantial addition. This finite nature of dead languages, together with the need of empirical data to their study, makes the preservation and the careful analysis of their legacy a core task of the (scientific) community. Although computational and corpus linguistics have mainly focused on building tools and resources for modern languages, there has always been large interest in providing scholars with collections of texts written in dead or historical languages (Berti, 2019). Not by chance, one of the first electronic corpora ever produced is the “Index Thomisticus” (Busa, 1974 1980), the opera omnia of Thomas Aquinas written in Latin in the 13th century. Owing to its

wide diachronic span covering more than two millennia, as well as its diatopic distribution across Europe and the Mediterranean, Latin is the most resourced historical language with respect to the availability of textual corpora. Large collections of Latin texts, e.g. the *Perseus Digital Library*² and the corpus of Medieval Italian Latinity *ALIM*³, can now be processed with state-of-the-art computational tools and methods to provide linguistic resources that enable scholars to exploit the empirical evidence provided by such datasets to the fullest. This is particularly promising given that the quality of many textual resources for Latin, carefully built over decades, is high.

Recent years have seen the rise of language modeling and feature learning techniques applied to linguistic data, resulting in so-called “word embeddings”, i.e. empirically trained vectors of lexical items in which words occurring in similar linguistic contexts are assigned close vectorial space. The semantic meaningfulness and motivation of word embeddings stems from the basic assumption of distributional semantics, according to which the distributional properties of words mirror their semantic similarities and/or differences, so that words sharing similar contexts tend to have similar meanings.

In this paper, we present and evaluate a number of embeddings for Latin built from a manually lemmatized dataset containing texts from the Classical era.⁴ In addition, we release embeddings trained on a manually lemmatized corpus of medieval texts to facilitate diachronic analyses. This research is performed in the context of the *LiLa: Linking Latin* project, which seeks to build a Knowledge Base of linguistic resources for Latin connected via a common vocabulary of knowledge

¹Copyright ©2019 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

²<http://www.perseus.tufts.edu/hopper/>

³<http://www.alim.dfil.univr.it/>

⁴Word embeddings built on tokens of the same dataset are also available online.

description following the principles of the Linked Data framework.⁵ Our contribution provides the community with new resources to be connected in the LiLa Knowledge Base aimed at supporting data-driven socio-cultural studies of the Latin world. The added value of our lemma embeddings for Latin results from the interdisciplinary blending of state-of-the-art methods in computational linguistics with the long tradition of Latin corpora creation: on the one hand the embeddings are evaluated with techniques hitherto applied to modern languages data only, on the other they are built from high quality datasets heavily used by scholars working on Latin.

2 Related Work

Word embeddings are crucial to many Natural Language Processing (NLP) tasks (Collobert et al., 2011; Lample et al., 2016; Yu et al., 2017). Numerous pre-trained word vectors generated with different algorithms have been released, typically generated from huge amounts of contemporary texts written in modern languages. The interest towards this type of distributional approach has emerged also in the Digital Humanities, as evidenced by publications on the use of word embeddings trained on literary texts or historical documents (Hamilton et al., 2016; Leavy et al., 2018; Sprugnoli and Tonelli, 2019). Although to a lesser extent, the literature also reports works on word embeddings for dead languages, including Latin.

Both Facebook and the organizers of the CoNLL shared tasks on multilingual parsing have pre-computed and released word embeddings trained on Latin texts crawled from the web: the former using the fastText model on Common Crawl and Wikipedia dumps (Grave et al., 2018a), the latter applying word2vec to Common Crawl only (Zeman et al., 2018). Both resources were developed by relying on automatic language detection engines: they are very big in terms of vocabulary size⁶ but highly noisy due to the presence of languages other than Latin. In addition, they include terms related to modern times, such as movie stars, TV series, companies (e.g., *Cumberbatch*, *Simpson*, *Google*), making them unsuitable for the study of language use in ancient texts. The automatic detection of language has

also been employed by Bamman (2012) to collect a corpus of Latin books available from Internet Archive. The corpus spans from 200 BCE to the 20th century and contains 1.38 billion tokens: embeddings trained on this corpus⁷ were used to investigate the relationship between concepts and historical characters in the work of Casiodorus (Bjerva and Praet, 2015). However, these word vectors are affected by OCR errors present in the training corpus: 25% of the embedding vocabulary contains non-alphanumeric characters, e.g. *-**-, iftud^*. The quality of the corpus used to train the Latin word embeddings available through the SemioGraph interface⁸, on the other hand, is high: these embeddings are based on the “Computational Historical Semantics” database, a manually curated collection of 4,000 Latin texts written between the 2nd and the 15th century AD (Jussen and Rohmann, 2015). In SemioGraph, more than one hundred word vectors can be visually explored searching by Part-of-Speech (PoS) labels and text genres: however, these vectors cannot be downloaded for further analysis and were generated with one model only, i.e. word2vec.

With respect to the works cited above, in this paper we rely on manually lemmatized texts free of OCR errors, we focus on a period not covered by the “Computational Historical Semantics” database and we test two models to learn lemma representations. It is worth noting that none of the previously mentioned studies have carried out an evaluation of the trained Latin embeddings; we, on the contrary, provide both quantitative and qualitative evaluations of our vectors.

3 Dataset Description

Our lemma vectors were trained on the “Opera Latina” corpus (Denooz, 2004). This textual resource has been collected and manually annotated since 1961 by the Laboratoire d’Analyse Statistique des Langues Anciennes (LASLA) at the University of Liège⁹. It includes 158 texts from 20 different Classical authors covering various genres, such as treatises (e.g. “Annales” by Tacitus), letters (e.g. “Epistulae” by Pliny the Younger), epic poems (e.g. “Aeneis” by Virgil), elegies

⁵<https://lila-erc.eu/>

⁶For example, the size of the CoNLL embeddings vocabulary is 1,082,365 words.

⁷<http://www.cs.cmu.edu/~dbamman/latin.html>

⁸<http://semigraph.texttechnologylab.org/>

⁹<http://web.philo.ulg.ac.be/lasla/textes-latins-traites/>

TARGET WORDS	SYNONYMS	DECOY WORDS		
<i>decretum</i> /decree	<i>edictum</i> /proclamation	<i>flagitium</i> /shameful act	<i>adolesco</i> /to grow up	<i>stipendiarius</i> /tributary
<i>saepe</i> /often	<i>crebro</i> /frequently	<i>conquiro</i> /to seek for	<i>ululatus</i> /howling	<i>frugifer</i> /fertile
<i>rogo</i> /to ask	<i>oro</i> /to ask for	<i>columna</i> /column	<i>retorqueo</i> /to twist back	<i>errabundus</i> /vagrant
<i>exilis</i> /thin	<i>macer</i> /emaciated	<i>moles</i> /pile	<i>mortalitas</i> /mortality	<i>audens</i> /daring

Table 1: Examples taken from the Latin benchmark for the synonym selection task.

(e.g. “Elegiae” by Propertius), plays (both comedies and tragedies e.g. “Aulularia” by Plautus and “Oedipus” by Seneca), and public speeches (e.g. “Philippicae” by Cicero)¹⁰.

The corpus contains several layers of linguistic annotation, such as lemmatization, PoS tagging and tagging of inflectional features, organized in space-separated files. “Opera Latina” contains approximately 1,700,000 words (punctuation is not present in the corpus), corresponding to 133,886 unique tokens and 24,339 unique lemmas.

4 Experimental Setup

We tested two different vector representations, namely word2vec (Mikolov et al., 2013a) and fastText (Bojanowski et al., 2017): the former is based on linear bag-of-words contexts generating a distinct vector for each word, whereas the latter is based on a bag of character n-grams, that is, the vector for a word (or a lemma) is the sum of its character n-gram vectors. Lemma vectors were pre-computed using two dimensionalities (100, 300) and two models: skip-gram and Continuous Bag-of-Words (CBOW). In this way, we had the possibility of evaluating both modest and high dimensional vectors and two architectures: skip-gram is designed to predict the context given a target word, whereas CBOW predicts the target word based on the context. The window size was 10 lemmas for skip-gram and 5 for CBOW. The other training options were the same for the two models:

- number of negatives sampled: 25;
- number of threads: 20;
- number of iterations over the corpus: 15;
- minimal number of word occurrences: 5.

Embeddings were trained on the lemmatized “Opera Latina” in order to reduce the data sparsity due to the high inflectional nature of Latin. Moreover, we lower-cased the text and converted *v* into *u* (so that *vir* ‘man’ becomes *uir*) to fit the lexicographic conventions of some Latin dictionaries

¹⁰The corpus can be queried through an online interface after requesting credentials: <http://cip193.philo.ulg.ac.be/OperaLatina/>

	word2vec		fastText	
	cbow	skip-gram	cbow	skip-gram
100	81.14%	79.83%	80.57%	86.91%
300	80.86%	79.48%	79.43%	86.40%

Table 2: Results of the synonym selection task calculated on the whole benchmark.

	word2vec		fastText	
	cbow	skip-gram	cbow	skip-gram
100	81.48%	85.18%	77.77%	87.03%
300	76.63%	85.18%	75.92%	90.74%

Table 3: Results of the synonym selection task calculated on a subset of the benchmark containing only questions with lemmas sharing the same PoS.

(Glare, 1982) and corpora. With the minimal number of lemma occurrences set to 5, we obtained a vocabulary size of 11,327 lemmas.

5 Evaluation

Word embeddings resulting from the experiments described in the previous Section were tested performing both an intrinsic and a qualitative evaluation (Schnabel et al., 2015). To the best of our knowledge, these methods, although well documented in the literature, have never been applied to the evaluation of Latin embeddings.

5.1 Synonym Selection Task

In the synonym selection task, the goal is to select the correct synonym of a target lemma out of a set of possible answers (Baroni et al., 2014). The most commonly used benchmark for this task is the Test of English as a Foreign Language (TOEFL), consisting of multiple-choice questions each involving five terms: the target words and another four, one of which is a synonym of the target word and the remaining three decoys (Landauer and Dumais, 1997). The original TOEFL dataset is made of only 80 questions but extensions have been proposed to widen the set of multiple-choice questions using external resources such as WordNet (Ehlert, 2003; Freitag et al., 2005).

In order to create a TOEFL-like benchmark for Latin, we relied on four digitized dictionaries

	contrudo /to thrust	frugaliter /thrifty	auspicatus /consecrated by auspices
fastText-skip	<i>protrudo</i> */to thrust forward <i>extrudo</i> */to thrust out	<i>frugalis</i> */thrifty <i>frugalitas</i> */economy	<i>auspicato</i> */after taking the auspices <i>auspicium</i> */auspices
fastText-cbow	<i>contego</i> */to cover <i>contraho</i> /to collect	<i>aliter</i> /differently <i>negligenter</i> /neglectfully	<i>auguratus</i> */the office of augur <i>pontificatus</i> /the office of pontifex
word2vec-skip	<i>infodio</i> /to bury <i>tabeo</i> /to melt away	<i>frugi</i> */frugal <i>quaerito</i> /to seek earnestly	<i>erycinus</i> /Erycinian <i>parilia</i> /the feast of Pales
word2vec-cbow	<i>refundo</i> /to pour back <i>infodio</i> /to bury	<i>lautus</i> /neat <i>frugi</i> */frugal	<i>erycinus</i> /Erycinian <i>parilia</i> /the feast of Pales

Table 4: Examples of the nearest neighbors of rare lemmas

of Latin synonyms (Hill, 1804; Dumesnil, 1819; Von Doederlein and Taylor, 1875; Skřivan, 1890) available online in XML Dictionary eXchange format¹¹. Starting from the digital versions of the dictionaries, we proceeded as follows:

- we downloaded and parsed the XML files so as to extract only the information useful for our purposes, that is, the dictionary entry and the synonyms;
- we merged the content of all dictionaries to obtain the largest possible list of lemmas with their corresponding synonyms. Unlike “Opera Latina” and the other synonym dictionaries, Dumesnil (1819) often lemmatizes verbs under the infinite form; therefore, for the sake of uniformity, we used LEM-LAT v3¹² to obtain the first person, singular, present, active (or passive, in case of deponent verbs), indicative form of all verbs registered in that dictionary in their present infinite form (e.g. *accingere* ‘to gird on’ → *accingo*) (Passarotti et al., 2017). At the end of this phase, we obtained a new resource containing 2,759 unique entries and covering all types of PoS, together with their synonyms;
- multiple-choice questions were created by taking each entry as a target lemma, then adding its first synonym and another three lemmas randomly chosen from the “Opera Latina” corpus;
- a Latin language expert manually checked samples of multiple-choice questions so as to be sure that the three randomly chosen lemmas were in fact decoy lemmas.

Table 1 provides some examples of the multiple-choice questions generated using the procedure described above .

¹¹<https://github.com/nikita-moor/latin-dictionary>

¹²<https://github.com/CIRCSE/LEMLAT3>

We computed the performance of the embeddings by calculating the cosine similarity between the vector of the target lemma and that of the other lemmas, picking the candidate with the largest cosine. Questions containing lemmas not included in the vocabulary, and thus vectorless, are automatically filtered out; results are given in terms of accuracy. As shown in Table 2, fastText proved to be the best lemma representation for the synonym selection task with the skip-gram architecture achieving an accuracy above 86%. This result can be explained by the fact that fastText is able to model morphology by taking into consideration sub-word units (i.e. character n-grams) and joining lemmas from the same derivational families. In addition, the skip-gram architecture works well with small amounts of training data like ours. It is also worth noting that, for both architectures and models, vectors with a modest dimensionality achieved a slightly higher accuracy with respect to embeddings with 300 dimensions.

The error analysis revealed specific types of linguistic and semantic relations, other than synonymy, holding between the target lemma and the decoy lemma that resulted having the largest cosine: for example, meronymy (e.g., target word: *annalis* ‘chronicles’ - synonym: *historia* ‘narrative of past events’ - answer: *charta* ‘paper’) and morphological derivation (e.g. target word: *consors* ‘having a common lot’ - synonym: *particeps* ‘sharer’ - answer: *sors* ‘lot’).

As an additional analysis, we repeated our evaluation on a subset of the benchmark containing 85 questions made of lemmas sharing the same PoS, e.g. *auxilior* ‘to assist’, *adiuuvo* ‘to help’, *censeo* ‘to assess’, *reuerto* ‘to turn back’, *humo* ‘to bury’. Results reported in Table 3 confirm that the skip-gram architecture provides the best accuracy for this task achieving a score above 90% for fastText embeddings with 300 dimensions. We also note an improvement of the accuracy for word2vec (+5%). The reasons behind these results need further in-

vestigations.

5.2 Qualitative Evaluation on Rare Lemma Embeddings

One of the main differences between word2vec and fastText is that the latter is supposed to be able to generate better embeddings for words that occur rarely in the training data. This is due to the fact that rare words in word2vec have few neighbor context words from which to learn the vector representation, whereas in fastText even rare words share their character n-grams with other words, making it possible to represent them reliably. To validate this hypothesis, we performed a qualitative evaluation of the nearest neighbors of a small set of randomly selected lemmas appearing between 5 and 10 times only in the “Opera Latina” corpus. Two Latin language experts manually checked the two most similar lemmas (in terms of cosine similarity) induced by the different 100-dimension embeddings we trained. Table 4 presents a sample of the selected rare lemmas and their neighbors: an asterisk marks neighbors that two experts judged as most semantically-related to the target lemma. This manual inspection, even if based on a small set of data, shows that the embeddings trained using the fastText model with the skip-gram architecture can find more similar lemmas than those trained with other models and architectures.

6 A Diachronic Perspective

Diachronic analyses are particularly relevant for Latin given that its use spans more than two millennia. To support this type of study we release, together with the embeddings presented in the previous Sections, lemma vectors trained on the “Opera Maiora”, written by Thomas Aquinas in the 13th century. “Opera Maiora” is a set of philosophical and religious works comprising some 4.5 million words (Passarotti, 2015): all texts are manually lemmatized and tagged at the morphological level (Passarotti, 2010) and are part of the “Index Thomisticus” (IT) corpus.

Before training the embeddings, we pre-processed the texts following the conventions adopted in “Opera Latina”: we lower-cased, removed punctuation, and converted *v* and *j* into *u* and *i*, respectively. Embeddings were trained with the configuration that reported the best results in the evaluation described in Section 5 (i.e. fastText

with the skip-gram architecture and 100 dimensions). For a comparative analysis with the embeddings of “Opera Latina”, we aligned the embeddings of “Opera Maiora” to the same coordinate axes using the unsupervised alignment algorithm provided with the fastText code (Grave et al., 2018b). Thanks to this alignment, we can inspect the nearest neighbors (nn) of lemmas in the two embeddings. For example, the lemma *ordo* shifts from social class or military rank (among the top 10 nn in the “Opera Latina” embeddings we find, in this order, *equester* ‘cavalry’, *legionarius* ‘legionary’, *turmatim* ‘by squadrons’) to referring to the concept of order and intellectual structure in Thomas Aquinas (nn in “Opera Maiora”: *ordinatio* ‘setting in order’, *coordinatio* ‘arranging together’, *ordino* ‘set in order’) (Busa, 1977). Another interesting case is *spiritus*: in the Classical era it refers to ‘breath’ (nn in “Opera Latina”: *spiro* ‘to blow’, *exspiro* ‘to exhale’, *spiramentum* ‘draught’), while in Aquinas’ Christian writings it associated with the Holy Ghost (nn: *sanctio* ‘to make sacred’, *donum* ‘gift’, *paracletus* ‘protector’) (Busa, 1983).

7 Conclusion and Future Work

In this paper we presented a new set of Latin embeddings based on high quality lemmatized corpora and a new benchmark for the synonym selection task. The aligned embeddings can be visually explored through a web interface and all the resources are freely available online: <https://embeddings.lila-erc.eu>.

Several future works are envisaged. For example, we plan to develop new benchmarks, like the analogy test (Mikolov et al., 2013b) or the rare words dataset (Luong et al., 2013), for the intrinsic quantitative evaluation of Latin embeddings. Moreover, embeddings could be used to improve the linking of datasets in the LiLa Knowledge Base. We would also like to extend the diachronic analysis to the embeddings trained on the “Computational Historical Semantics” database as soon as these become available.

This work represents the first step towards the development of a new set of resources for the analysis of Latin. This effort is laying the foundations of the first campaign devoted to the evaluation of NLP tools for Latin, EvaLatin.

Acknowledgments

This work is supported by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme via the "LiLa: Linking Latin" project - Grant Agreement No. 769994. The authors also wish to thank Andrea Peverelli for his expert support on Latin and Chris Culy for providing his code for the embeddings visualization.

References

- David Bamman and David Smith. 2012. Extracting two thousand years of Latin from a million book library. *Journal on Computing and Cultural Heritage (JOCCH)*, 5(1):1–13.
- Marco Baroni, Georgiana Dinu, and Germán Kruszewski. 2014. Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 238–247.
- Monica Berti. 2019. *Digital Classical Philology: Ancient Greek and Latin in the Digital Revolution*, volume 10. Walter de Gruyter GmbH & Co KG.
- Johannes Bjerva and Raf Praet. 2015. Word embeddings pointing the way for late antiquity. In *Proceedings of the 9th SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH)*, pages 53–57.
- Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.
- Roberto Busa. 1974–1980. *Index Thomisticus: sancti Thomae Aquinatis operum omnium indices et concordantiae, in quibus verborum omnium et singulorum formae et lemmata cum suis frequentis et contextibus variis modis referuntur quaeque / consociata plurimum opera atque electronico IBM automato usus digessit Robertus Busa SJ*. Frommann - Holzboog.
- Roberto Busa. 1977. Ordo dans les oeuvres de st. thomas d'aquin. *II Coll. Intern. del Lessico Intellettuale Europeo*, pages 59–184.
- Roberto Busa. 1983. De voce spiritus in operibus s. thomae aquinatis. *IV Coll. Intern. del Lessico Intellettuale Europeo*, pages 191–222.
- Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural language processing (almost) from scratch. *Journal of machine learning research*, 12(Aug):2493–2537.
- Joseph Denooz. 2004. Opera latina: une base de données sur internet. *Euphrosyne*, 32:79–88.
- Jean Baptiste Gardin Dumesnil. 1819. *Latin Synonyms: With Their Different Significations: and Examples Taken from the Best Latin Authors*. GB Whittaker.
- Bret R Ehlert. 2003. *Making accurate lexical semantic similarity judgments using word-context co-occurrence statistics*. University of California, San Diego.
- Dayne Freitag, Matthias Blume, John Byrnes, Edmond Chow, Sadik Kapadia, Richard Rohwer, and Zhiqiang Wang. 2005. New experiments in distributional representations of synonymy. In *Proceedings of the Ninth Conference on Computational Natural Language Learning*, pages 25–32. Association for Computational Linguistics.
- Peter G.W. Glare. 1982. *Oxford latin dictionary*. Oxford univ. press.
- Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. 2018a. Learning Word Vectors for 157 Languages. In Nicoletta Calzolari (Conference chair), Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Koiti Hasida, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hlne Mazo, Asuncion Moreno, Jan Odijk, Stelios Piperidis, and Takenobu Tokunaga, editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, pages 3843–3847, Miyazaki, Japan, May 7–12, 2018. European Language Resources Association (ELRA).
- Edouard Grave, Armand Joulin, and Quentin Berthet. 2018b. Unsupervised Alignment of Embeddings with Wasserstein Procrustes. pages 1880–1890.
- William L. Hamilton, Jure Leskovec, and Dan Jurafsky. 2016. Diachronic word embeddings reveal statistical laws of semantic change. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1489–1501, Berlin, Germany, August. Association for Computational Linguistics.
- John Hill. 1804. *The Synonymes in the Latin Language, Alphabetically Arranged; with Critical Dissertations Upon the Force of Its Prepositions, Both in a Simple and Compounded State: By John Hill, LL. D. Professor of Humanity in the University, and Fellow of the Royal Society, of Edinburgh*. James Ballantyne, for Longman and Rees, London.
- Bernhard Jussen and Gregor Rohmann. 2015. Historical Semantics in Medieval Studies: New Means and Approaches. *Contributions to the History of Concepts*, 10(2):1–6.
- Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer.

2016. Neural architectures for named entity recognition. In *Proceedings of NAACL-HLT*, pages 260–270.
- Thomas K. Landauer and Susan T. Dumais. 1997. A solution to Plato’s problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review*, 104(2):211–240.
- Susan Leavy, Karen Wade, Gerardine Meaney, and Derek Greene. 2018. Navigating literary text with word embeddings and semantic lexicons. In *Workshop on Computational Methods in the Humanities 2018 (COMHUM 2018)*, Luasanne, Switzerland, 4–5 June 2018.
- Thang Luong, Richard Socher, and Christopher Manning. 2013. Better word representations with recursive neural networks for morphology. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, pages 104–113.
- Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. *Proceedings of Workshop at ICLR*.
- Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013b. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.
- Marco Passarotti, Marco Budassi, Eleonora Litta, and Paolo Ruffolo. 2017. The Lemlat 3.0 Package for Morphological Analysis of Latin. In *Proceedings of the NoDaLiDa 2017 Workshop on Processing Historical Language*, number 133, pages 24–31. Linköping University Electronic Press.
- Marco Passarotti. 2010. Leaving behind the less-resourced status. The case of Latin through the experience of the Index Thomisticus Treebank. In *7th SaLTMiL Workshop on Creation and use of basic lexical resources for less-resourced languages LREC 2010, Valetta, Malta, 23 May 2010 Workshop programme*, pages 27–32.
- Marco Passarotti. 2015. What you can do with linguistically annotated data. from the index thomisticus to the index thomisticus treebank. In *Reading Sacred Scripture with Thomas Aquinas: Hermeneutical Tools, Theological Questions and New Perspectives*, pages 3–44.
- Tobias Schnabel, Igor Labutov, David Mimno, and Thorsten Joachims. 2015. Evaluation methods for unsupervised word embeddings. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 298–307.
- Arnošt Skřivan. 1890. *Latinská synonymika pro školu i dum*. V CHRUDIMI.
- Rachele Sprugnoli and Sara Tonelli. 2019. Novel event detection and classification for historical texts. *Computational Linguistics*, 45(2):229–265.
- Ludwig Von Doederlein and Samuel Harvey Taylor. 1875. *Döderlein’s Hand-book of Latin Synonymes*. WF Draper.
- Liang-Chih Yu, Jin Wang, K Robert Lai, and Xuejie Zhang. 2017. Refining word embeddings for sentiment analysis. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 534–539.
- Daniel Zeman, Jan Hajič, Martin Popel, Martin Potthast, Milan Straka, Filip Ginter, Joakim Nivre, and Slav Petrov. 2018. CoNLL 2018 shared task: Multilingual parsing from raw text to universal dependencies. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 1–21.

Valutazione umana di Google Traduttore e DeepL per le traduzioni di testi giornalistici dall'inglese verso l'italiano

Mirko Tavosanis

Dipartimento di Filologia, letteratura e linguistica

Università di Pisa

Via Santa Maria 36 – 56126 Pisa PI

mirko.tavosanis@unipi.it

Riassunto¹

Il contributo presenta una valutazione delle prestazioni di Google Traduttore e di DeepL attraverso le interfacce web disponibili al pubblico. Per la valutazione è stato usato un campione di 100 frasi tratto da testi giornalistici in lingua inglese tradotti in italiano. Le traduzioni prodotte sono state valutate da esseri umani e i risultati della valutazione sono stati confrontati con il calcolo del punteggio BLEU. La valutazione umana dei sistemi automatici ha mostrato livelli di qualità vicini a quelli della traduzione umana, mentre il punteggio BLEU non ha mostrato una stretta corrispondenza con la valutazione umana.

Abstract

The paper describes an assessment of the performance of Google Translator and DeepL when the systems are used through their public web interfaces. The assessment was carried on a sample of 100 sentences from English-language journalistic texts translated into Italian. The translation outputs were evaluated by humans and the results of the evaluation were compared with the calculation of the BLEU score. Human evaluation of machine translation has shown quality levels very close to those of human translation,

while the BLEU score has not shown a strict connection with human evaluation.

1 Introduzione

I sistemi di traduzione automatica stanno assumendo un ruolo sempre più importante nella vita quotidiana, da soli o integrati in altre pratiche (Bersani Berselli 2011). La loro diffusione potrebbe anche produrre innovazioni strutturali e trasformare in profondità alcuni settori lavorativi, a cominciare dall'insegnamento delle lingue straniere (Ostler 2010; Tavosanis 2018).

Tuttavia, la valutazione delle effettive prestazioni di questi sistemi rimane un problema complesso sia dal punto di vista teorico sia dal punto di vista pratico. Inoltre, la difficoltà di valutazione è considerata da tempo uno dei vincoli principali anche per lo sviluppo dei sistemi di traduzione (Pieraccini 2012, p. 275; Hajič 2008, p. 85).

Per la valutazione sono state sviluppate numerose metriche di tipo automatico o semiautomatico; la più usata in tempi recenti è stata BLEU (Papinieni e altri, 2002). Il lavoro sulle metriche è costante e, in particolare, alla valutazione delle metriche è dedicato uno degli *shared tasks* delle conferenze WMT (i risultati della più recente sono presentati in Fourth Conference on Machine Translation 2019, pp. 494-525).

Tuttavia, nel corso degli ultimi anni è diventato evidente che le metriche più usate non sono in realtà in grado di descrivere adeguatamente le differenze e i miglioramenti di prestazioni dei sistemi più recenti di traduzione automatica, e in particolare di quelli basati su reti neurali. Il pro-

¹ Copyright © 2019 for this paper by its author.
Use permitted under Creative Commons License
Attribution 4.0 International (CC BY 4.0).

blema può essere descritto in generale come problema di scarsa correlazione tra le metriche e il giudizio umano. Usare come punto di riferimento il giudizio umano sembra d'altra parte del tutto corretto dal punto di vista metodologico: l'obiettivo dei sistemi di traduzione è principalmente quello di fornire traduzioni che gli esseri umani considerino di buon livello.

In particolare, la non perfetta correlazione tra BLEU e il giudizio umano è stata notata da tempo (per esempio: Callison-Burch, Osborne e Koehn 2006) e diversi valutatori hanno ribadito la necessità di considerare la valutazione umana come primaria (Callison-Burch e altri 2008, p. 72). La situazione è stata probabilmente resa meno evidente anche dall'abitudine frequente di valutare sistemi diversi confrontando le prestazioni tra di loro e non su una scala assoluta; tuttavia, questa prassi non è mai stata l'unica e i sistemi presentati nella principale campagna di valutazione sulla traduzione automatica, i task WMT, sono valutati solo con giudizi assoluti, non con giudizi relativi.

Il problema si è mostrato con particolare evidenza negli ultimi anni, in seguito alla rapida introduzione dei sistemi di traduzione basati su reti neurali. BLEU, come i sistemi di traduzione statistica (PB-SMT), basa il proprio funzionamento sugli n -grammi. Si ritiene però che questo meccanismo mostri un "inherent bias" contro i sistemi che adottano meccanismi di traduzione non basati su n -grammi, quali appunto i sistemi basati su reti neurali (Way 2018, p. 170).

Diverse verifiche hanno mostrato che in pratica BLEU sottovaluta fortemente i risultati dei sistemi di traduzione a reti neurali (Bentivogli e altri 2018a; Shterionov e altri 2018). Naturalmente, la validità di queste verifiche può essere relativizzata alle caratteristiche di specifici campioni. Tuttavia, i dati oggi disponibili giustificano l'idea che BLEU non possa essere usato come indicatore generale di qualità di questi sistemi.

In questo contesto non mancano dichiarazioni in cui si rivendica il raggiungimento della "parità" tra traduzione automatica e traduzione umana per alcuni sistemi commerciali (Hassan e altri 2018). Le verifiche indipendenti non hanno però al momento confermato questi risultati; al contrario, hanno evidenziato differenze significative (Toral e altri 2018).

Dichiarazioni del genere mostrano comunque l'utilità di una valutazione esterna delle prestazioni dei sistemi più usati. Anche il presente contributo concorre a questa attività, documentando lo stato delle cose per prodotti di ampia diffusione e in un contesto d'uso reale per una lingua su cui

le valutazioni sono state finora piuttosto ridotte. Alcuni testi generati con traduzione automatica sono stati quindi sottoposti a valutazione umana, assieme a prodotti di traduttori umani, con l'obiettivo di:

1. fornire una valutazione umana delle prestazioni (assolute e non relative) di due diversi sistemi
2. confrontare i risultati della valutazione umana con quelli della valutazione ottenuta attraverso BLEU

2 Il contesto della traduzione

Le verifiche descritte di seguito sono state compiute usando due sistemi liberamente accessibili al pubblico e spesso indicati come i migliori nel loro genere: Google Traduttore e DeepL.

I due sistemi non sono forse i più utilizzati su scala mondiale. Si può pensare che Google Traduttore sia il sistema più comunemente usato, ma in assenza di indicazioni ufficiali è possibile che questo primato vada in realtà assegnato al sistema di traduzione automatica di Facebook.

DeepL non solo è sicuramente meno noto di Google Traduttore, ma è probabilmente meno usato anche di un quarto sistema di traduzione, Microsoft Translator. Tuttavia, DeepL è frequentemente segnalato come uno dei migliori prodotti della sua categoria e nelle valutazioni con BLEU ha ottenuto negli ultimi anni punteggi spesso superiori a quelli di Google Traduttore (Heiss e Soffritti 2018).

3 Google Traduttore

Le origini di Google Traduttore risalgono al 2003, quando il servizio venne lanciato con il nome di Google Translate. In seguito il servizio è stato rinominato, per l'italiano, come Google Traduttore.

Alle origini, il sistema si basava su prodotti SYSTRAN. Già nel 2006 Google iniziò comunque a usare un sistema di traduzione statistica sviluppato in proprio, GSMT (Google Statistical Machine Translation). Caratteristica di questo sistema è l'uso dell'inglese come lingua ponte, per cui le traduzioni tra lingue diverse dall'inglese vengono fatte passando comunque da una versione in lingua inglese e poi ritradotte – con un peggioramento significativo della qualità rispetto alle traduzioni dirette da e verso l'inglese (una sintesi delle fasi di sviluppo è presentata in Tavosanis

2018, pp. 95-96). Le lingue coperte sono aumentate rapidamente e, anche se nell'ultimo anno non ne sono state aggiunte di nuove, nel luglio del 2019 risultavano in tutto 103 (la lista completa è disponibile sul sito <https://translate.google.com/>), traducibili reciprocamente per un totale di poco più di 10.000 diverse combinazioni.

Nel frattempo, Google ha sviluppato il prodotto inserendovi caratteristiche di intelligenza artificiale basate sull'apprendimento automatico e sulle reti neurali. Il 15 novembre 2016 è stato quindi annunciato il passaggio di una parte dei servizi di Google Traduttore dal sistema GSMT a quello GNMT (Google Neural Machine Translation). Rispetto al precedente, GNMT ha il vantaggio di tradurre, secondo gli sviluppatori, frase intera e non spezzoni di frase, curando in particolare la coesione grammaticale, che nei sistemi precedenti non sempre veniva rispettata (Turovsky 2016). Nel marzo 2017, il sistema GNMT era già disponibile per traduzioni tra otto lingue: inglese, cinese, francese, tedesco, giapponese, coreano, portoghese, spagnolo e turco. Nell'aprile dello stesso anno è stato esteso ad altre lingue europee, tra cui l'italiano (Google 2017).

4 DeepL

Realizzato dall'azienda tedesca DeepL GmbH, il sistema di traduzione DeepL è stato reso disponibile al pubblico nell'agosto del 2017 (sito: <https://www.deepl.com/>). Rispetto a Google, copre un numero relativamente ridotto di lingue, tutte di origine indoeuropea: italiano, inglese, tedesco, francese, spagnolo, portoghese, olandese, polacco e russo. Dal punto di vista tecnico, l'azienda ha dichiarato che il sistema di traduzione si basa su reti neurali, ma non ha fornito altre informazioni.

5 Procedura di valutazione

Per la valutazione del lavoro è stato usato un corpus di articoli di quotidiani e periodici. Tale scelta è stata fatta in base a diversi fattori:

- Importanza, in quanto l'italiano giornalistico è centrale nell'architettura dell'italiano contemporaneo (Bonomi 2002, Berruto 2012)
- Verosimiglianza, in quanto la traduzione di articoli di questo tipo è un impiego realistico dei sistemi descritti, nella loro versione rivolta all'utente generico e resa disponibile attraverso un'interfaccia web

- Disponibilità, in quanto è facile ottenere ragionevoli quantitativi di articoli in doppia versione, originali e tradotti
- Praticità, in quanto le traduzioni degli articoli spesso hanno una corrispondenza 1:1 tra le frasi del testo originale e quelle del testo tradotto.

Il lavoro è stato condotto su un campione di 100 frasi, valutate separatamente (da valutatori diversi) per l'adeguatezza (*adequacy*) e per la fluenza (*fluency*). Anche se i risultati delle verifiche WMT hanno confermato la maggior rilevanza dell'adeguatezza (Bentivogli e altri 2018b: 62), le due valutazioni diverse sono state conservate per verificare l'esistenza di differenze nei prodotti commerciali. Va comunque notato che dal punto di vista dell'adeguatezza, nonostante sia teoricamente possibile che una frase tradotta con sistemi a reti neurali non abbia nulla a che fare contenutisticamente con il testo di partenza, nella pratica non si è prodotto nessun caso di questo genere.

Le scale utilizzate sono state:

Adeguatezza

1. Il contenuto informativo dell'originale è stato completamente alterato
2. È stata trasmessa una parte del contenuto informativo, ma non la più importante
3. Circa metà del contenuto informativo è stata trasmessa
4. La parte più importante del contenuto informativo originale è stata trasmessa
5. Il contenuto informativo è stato tradotto completamente

Fluenza

1. Impossibile da ricondurre alla norma
2. Con più di due errori morfosintattici
3. Con non più di due errori morfosintattici e/o molti usi insoliti di collocazioni
4. Con non più di un errore morfosintattico e/o un uso insolito di collocazioni
5. Del tutto corretta

All'interno del campione sono state inserite casualmente frasi provenienti da un corpus di 15 articoli di quotidiani e periodici, scelti casualmente sulla base della disponibilità online sia del testo originale sia di una traduzione in lingua italiana. In alcuni casi, le traduzioni umane prese in esame sono opera di volontari ma sono comunque di buon livello qualitativo. I testi originali in inglese

sono stati ripuliti e sottoposti alle interfacce web di Google Traduttore e DeepL. Poiché queste interfacce, nella versione liberamente accessibile, accettano testi di una lunghezza massima di 5000 caratteri, i testi più lunghi sono stati scomposti in blocchi di lunghezza inferiore, rispettando i confini di frase (e spesso di capoverso). I blocchi stessi sono stati poi sottoposti individualmente ai sistemi.

Al termine della procedura, per ogni articolo erano quindi disponibili:

1. Il testo originale in lingua inglese
2. La traduzione umana
3. La traduzione prodotta da Google
4. La traduzione prodotta da DeepL

Le frasi da esaminare sono state selezionate in modo casuale. Sono poi state sottoposte ai valutatori in ordine casuale e senza indicazioni sulla loro origine: i valutatori non avevano quindi elementi esterni per decidere se l'origine di una singola frase era un traduttore umano, DeepL o Google. Nella valutazione per adeguatezza le frasi erano accompagnate dal testo originale in lingua inglese, secondo l'orientamento *DA-src* (Bentivogli e altri 2018b: 62), mentre nella valutazione per fluenza era disponibile solo il testo italiano. La valutazione è stata eseguita su carta, in condizioni controllate, per un tempo medio di un'ora per ogni campione.

I valutatori sono stati complessivamente 14: 6 hanno valutato l'adeguatezza, 8 la fluenza. La valutazione della fluenza è stata condotta su un campione più esteso di 147 frasi, per rendere la lunghezza dell'attività paragonabile a quella della valutazione dell'adeguatezza. Ai fini della valutazione sono state tuttavia usate solo le 100 frasi coincidenti con frasi valutate per adeguatezza.

Il gruppo dei valutatori era interamente formato da studenti del corso di laurea magistrale in Informatica umanistica dell'Università di Pisa. Tutti i valutatori avevano l'italiano come lingua madre e disponevano di una conoscenza della lingua inglese di livello B2 o superiore. Nessuno di loro aveva esperienza di attività redazionale o di revisione di traduzioni e nessuno è stato coinvolto nella fase di scelta e preparazione degli articoli.

Per migliorare l'omogeneità del risultato, una settimana prima della valutazione vera e propria è stata fatta una sessione di prova con i valutatori interessati. In questa sessione sono state valutate frasi diverse da quelle esaminate in seguito. I punteggi assegnati sono stati discussi sulla base dei

testi, cercando di arrivare quanto più possibile alla condivisione di parametri per il lavoro effettivo.

6 Esito della valutazione

Nel giudizio finale la varianza dei giudizi è stata piuttosto ridotta. Le medie della varianza calcolata su ogni singola frase sono state infatti:

	Adeguatezza	Fluenza
Google	0,3982	0,4631
DeepL	0,4312	0,4375
Umano	0,4320	0,3432
Totale	0,4192	0,4243

Tabella 1: Varianza media nei giudizi per frasi.

Per quanto riguarda la fluenza, la varianza massima (0,1728) si è avuta nei giudizi per questa traduzione, con sei punteggi 4 e due punteggi 3:

Originale: As Rahme served a frugal dish of rice in vine leaves, her son unspooled a familiar Palestinian narrative.

Traduzione DeepL: Mentre Rahme serviva un frugale piatto di riso in foglie di vite, suo figlio ha sboccato un racconto familiare palestinese.

Più consistente è stata la varianza massima per l'adeguatezza, con due frasi che hanno ottenuto il livello di 1,9592:

Originale: And though people can be induced to use social media addictively, while ordering Deliveroo night after night, pausing only to take an Uber to the pub, wedding addiction remains a rarity.

Traduzione Google: E anche se le persone possono essere indotte a usare i social media in modo assopito, mentre ordinano Deliveroo notte dopo notte, facendo una pausa solo per portare un Uber al pub, la dipendenza da matrimonio rimane una rarità.

Originale: And now the Trump administration, having failed to repeal the ACA when Republicans controlled Congress, is suing to have the whole thing declared unconstitutional in court – because what could be a better way to start off the 2020 campaign than taking insurance away from 20 million Americans?

Traduzione umana: E ora l'amministrazione Trump, non essendo riuscita ad abrogare l'ACA quando i repubblicani controllavano il Congresso, sta facendo causa per far dichiarare l'intera cosa

incostituzionale in tribunale - perché quale modo migliore di togliere l'assicurazione a 20 milioni di americani per iniziare la campagna del 2020?

Le frasi oggetto di valutazione sono state poi riassemblate in tre diversi documenti, a seconda dell'origine, ed è stato calcolato il punteggio BLEU per i prodotti della traduzione automatica, confrontati con la traduzione umana. La valutazione risultante è stata:

Traduttore	N. frasi	Media adeguatezza	Media fluenza	BLEU
Google	37	4,15	3,90	0,2538
DeepL	39	4,30	3,94	0,3254
Umano	24	4,60	4,46	n. a.

Tabella 2: Valutazione complessiva delle traduzioni.

Per la fluenza, va notato che il punteggio 5 è stato assegnato all'unanimità solo a pochissime frasi. Tuttavia, alcune frasi sia di Google sia di DeepL hanno ottenuto questo punteggio massimo, cosa che viceversa non è successa per le traduzioni umane. Questo giudizio è stato assegnato soprattutto a frasi brevi, ma non solo a esse. Per esempio, sono state valutate 5 queste traduzioni:

Originale: Which is weird, because the truth is that everyone's judging everyone else's relationships all the time.

Traduzione DeepL: Il che è strano, perché la verità è che tutti giudicano sempre le relazioni altrui.

Originale: In an attempt to avert this awful fate, the American Medical Association launched what it called Operation Coffee Cup, a pioneering attempt at viral marketing.

Traduzione Google: Nel tentativo di scongiurare questo terribile destino, l'American Medical Association lanciò quella che chiamò Operation Coffee Cup, un tentativo pionieristico di marketing virale.

7 Esame dei risultati

In risposta alle domande presentate nel paragrafo 1 è innanzitutto notevole l'alto livello raggiunto da entrambi i sistemi. Nessuno dei due può essere considerato all'altezza della traduzione umana, e

non mancano i casi di frasi tradotte in modo molto insoddisfacente, come questa (valutazione media 1,43):

Originale: If you are used to the boil-them-whole, admire, tug-leaf-by-leaf, scape-with-bottom-teeth school of artichoke preparation and eating, it comes as a shock when you first see Romans deal, in typically direct style, with their favourite vegetable.

Traduzione Google: Se sei abituato a bollire tutto, ammira, rimorchia la foglia per pianta, scolpisci i denti di fondo con la preparazione e il consumo di carciofo, diventa un vero shock quando vedi per la prima volta i romani, in genere stile diretto, con il loro vegetale preferito.

Tuttavia, nel complesso, colpisce che per esempio per l'adeguatezza la distanza relativa tra la traduzione umana e DeepL sia pari solo al 6,5%. Il dislivello per quanto riguarda la fluenza è maggiore, ma rimane comunque molto contenuto.

I dati confermano inoltre la superiorità delle prestazioni di DeepL già segnalata da diverse fonti, anche se la differenza con Google è molto contenuta. Il margine relativo di vantaggio di DeepL è infatti solo del 3,5% per l'adeguatezza e dell'1% per la fluenza.

Va notato che la differenza nella composizione del campione potrebbe spiegare parte dei risultati; all'interno di eventuali prove future sarebbe sicuramente opportuno sottoporre alla valutazione campioni omogenei. Tuttavia, per esempio, la lunghezza media delle frasi, che influenza in negativo la qualità della traduzione automatica, non solo è molto simile nei due campioni, ma è superiore nel caso del sistema che ha ottenuto la valutazione più alta. Il campione usato per DeepL ha infatti una lunghezza media di 25,79 token per frase, mentre in quello usato per Google il valore equivalente è di 25,03.

Per quanto riguarda BLEU, la correlazione con la valutazione umana risulta davvero debole. Il ridotto scarto tra Google e DeepL nella valutazione umana diventa infatti una differenza relativa del 22% con BLEU.

Soprattutto, però, è notevole la differenza rispetto ai punteggi BLEU per la traduzione umana spesso indicati in bibliografia (Papinien e altri 2002), che si aggirano attorno a 0,6. Per DeepL questo corrisponderebbe a una differenza relativa del 45,8%, difficile da considerare rappresentativa della differenza tra i risultati su una scala di giudizio assoluta.

Va inoltre notato che negli ultimi anni i punteggi BLEU di sistemi come Google o Microsoft Translator si sono spesso collocati tra 0,2 e 0,4 (Tavosanis 2018). In questo contesto, se il punteggio di DeepL è piuttosto elevato, quello di Google si avvicina alla media.

8 Conclusioni e sviluppi futuri

Il lavoro descritto qui rappresenta una delle prime concretizzazioni di un progetto più ampio, dedicato a studiare le possibilità di inserimento strutturale dei traduttori automatici nella pratica didattica delle lingue partendo dall'analisi delle prestazioni e della possibilità di integrare facilmente i prodotti nel percorso di un traduttore in formazione. Nel giro di pochi mesi dovrebbero essere quindi disponibili valutazioni più estese. Per la traduzione italiana, queste valutazioni potrebbero essere di particolare interesse, considerando non solo la rapidità dei miglioramenti recenti ma anche il fatto che l'italiano è stato relativamente poco rappresentato nelle analisi condotte finora.

Per gli sviluppi futuri, l'aver preso in esame un unico genere testuale, per quanto variato, è un limite evidente dell'analisi (Burchardt e altri 2017: 159-160): l'estensione della valutazione a tipologie diverse rispetto all'articolo di quotidiano o periodico potrebbe facilmente portare a risultati molto diversi da quelli descritti qui. L'inclusione di altri generi testuali rappresenta quindi senz'altro il requisito più importante nella progettazione di un lavoro di valutazione su scala più estesa. In quest'ottica, sembra particolarmente interessante l'estensione del lavoro a testi specialistici.

Bibliografia

- Bentivogli, Luisa, e altri (2018a). *Neural versus phrase-based mt quality: An in-depth analysis on english-german and english-french*. In *Computer Speech & Language*, 49, pp. 52-70.
- Bentivogli, Luisa, e altri (2018b). *Machine Translation Human Evaluation: an investigation of evaluation based on Post-Editing and its relation with Direct Assessment*. In *Proceedings of the 15th International Workshop on Spoken Language Translation, Iwslt*, pp. 62-69.
- Berruto, Gaetano (2012). *Sociolinguistica dell'italiano contemporaneo. Nuova edizione*. Roma: Carocci.
- Bersani Berselli, Gabriele (a cura di, 2011), *Usare la traduzione automatica*. Bologna: CLUEB.
- Bonomi, Ilaria (2002). *L'italiano giornalistico. Dall'inizio del '900 ai quotidiani online*. Firenze: Cesati.
- Burchardt, Aljoscha, e altri (2017). *A Linguistic Evaluation of Rule-Based, Phrase-Based, and Neural MT Engines*. In *The Prague Bulletin of Mathematical Linguistics*, 108, pp. 159-70.
- Callison-Burch, Chris, e altri (2008). *Further meta-evaluation of machine translation*. In *Proceedings of the third workshop on statistical machine translation*, Association for Computational Linguistics, pp. 70-106.
- Callison-Burch, Chris, Miles Osborne e Philipp Koehn (2006). *Re-evaluation the role of BLEU in machine translation research*. In *11th Conference of the European Chapter of the Association for Computational Linguistics*, Association for Computational Linguistics, pp. 249-256.
- Fourth Conference on Machine Translation (2019), *Proceedings of the Conference, Volume 2: Shared Task Papers, Day 1*, Stroudsburg, ACL <<http://www.statmt.org/wmt19/pdf/53/WMT-2019-2.pdf>>.
- Google (2017). *Translation API Language Support*. Sito Google <<https://cloud.google.com/translate/docs/languages>>.
- Hajič, Jan (2008). *Linguistics Meets the Exact Sciences*. In *A companion to digital humanities*, a cura di Susan Schreibman, Ray Siemens e John Unsworth, Hoboken, John Wiley & Sons, pp. 79-87.
- Hassan, Hany, e altri (2018). *Achieving human parity on automatic Chinese to English news translation*. arXiv preprint arXiv:1803.05567 (2018).
- Heiss, Christine e Marcello Soffritti (2018). *DeepL Traduttore e didattica della traduzione dall'italiano in tedesco-alcune valutazioni preliminari*. In *Translation and Interpreting for Language Learners (TAIL). Lessons in honour of Guy Aston, Anna Ciliberti, Daniela Zorzi*, a cura di Laurie Anderson, Laura Gavioli e Federico Zanettin, Milano, AItLA, pp. 241-258.
- Ostler, Nicholas (2010). *The Last Lingua Franca. English until the Return of Babel*. Londra: Allen Lane.
- Papinieni, Kishore, e altri (2002). *BLEU: A Method for Automatic Evaluation of Machine Translation*. In *ACL-2002: 40th Annual Meeting of the Association for Computational Linguistics*, ACL, Stroudsburg, pp. 311-8.
- Pieraccini, Roberto (2012). *The Voice in the Machine. Building Computers that Understand Speech*. Boston: MIT Press.
- Shterionov, Dimitar, e altri (2018). *Human versus automatic quality evaluation of NMT and PBSMT*. In *Machine Translation*, 32, 3, pp. 217-235.
- Tavosanis, Mirko (2018). *Lingue e intelligenza artificiale*. Roma: Carocci.
- Toral, Antonio, e altri (2018). *Attaining the unattainable? Reassessing claims of human parity in neural machine translation*. arXiv preprint arXiv:1808.10432.
- Turovsky, Barak (2016). *Found in translation: More accurate, fluent sentences in Google Translate*. Google Blog <<https://www.blog.google/products/translate/found-translation-more-accurate-fluent-sentences-google-translate/>>.
- Way, Andy (2018). *Quality expectations of machine translation*. In *Translation Quality Assessment*, Springer, Cham, pp. 159-178.

Prendo la Parola in Questo Consesso Mondiale: A Multi-Genre 20th Century Corpus in the Political Domain

Sara Tonelli[†], Rachele Sprugnoli[‡], Giovanni Moretti^{†‡}

[†]Fondazione Bruno Kessler, Trento, Italy

[‡]Università Cattolica, Milano, Italy

{satonelli, moretti}@fbk.eu

rachele.sprugnoli@unicatt.it

Abstract

English. In this paper we present a multi-genre corpus spanning 50 years of European history. It contains a comprehensive collection of Alcide De Gasperi's public documents, 2,762 in total, written or transcribed between 1901 and 1954. The corpus comprises different types of texts, including newspaper articles, propaganda documents, official letters and parliamentary speeches. The corpus is freely available and includes several annotation layers, i.e. key-concepts, lemmas, PoS tags, person names and geo-referenced places, representing a high-quality 'silver' annotation. We believe that this resource can foster research in historical corpus analysis, stylometry and computational social science, among others.¹

1 Introduction

In recent years, political scientists and history scholars have started to exploit the availability of digital material to enrich their research, taking advantage of freely accessible online archives and easy-to-use tools for text processing and data extraction. Active communities have been created around topics such as the study of Parliamentary corpora (see the ParlaCLARIN² and ParlaFormat workshops³), the analysis of political manifestos⁴ and of Presidential speeches.⁵ Despite the importance of this research field, copyright and availability in machine-readable format still represent

major issues, especially in those countries where no or only limited public initiatives have been undertaken to support the distribution of this kind of documents. For example, while in the US the Federal Digital System grants access to public Presidential documents through APIs and bulk-data repositories, in Italy an effort along this line has started only recently with the support of the Archive of the President of the Republic⁶, but has not delivered substantial results so far.

This work represents a first attempt to deal with this lack of data, since we present and make available a large corpus of Italian public documents in the political domain. In particular, we release a comprehensive collection of Alcide De Gasperi's public documents issued between 1901 and 1954, which had been previously published in four volumes by Il Mulino (De Gasperi, 2006; De Gasperi, 2008a; De Gasperi, 2008b; De Gasperi, 2009) but were not machine-readable. Our repository contains all documents in three formats: txt, XML and tab-separated. Raw text files contain only the body of the documents, and may be straightforwardly used to extract embeddings or topics. XML files include metadata that cover not only the title, the date and the place of publication, but also key-concepts automatically extracted from each text and genre labels manually assigned by domain experts. Furthermore, the release includes silver annotation for lemma, part of speech, person names and place names with associated coordinates in a CoNLL-like format. All files and the corresponding descriptions can be downloaded at <https://dh.fbk.eu/technologies/corpus-de-gasperi> (with CC BY-NC-SA license). The corpus can also be navigated using the ALCIDE platform (Moretti et al., 2016) at this link: <http://alcidedigitale.fbk.eu/>.

¹Copyright ©2019 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

²<https://www.clarin.eu/ParlaCLARIN>

³<https://www.clarin.eu/event/2019/parlaformat-workshop>

⁴<https://manifesto-project.wzb.eu/>

⁵<https://www.presidency.ucsb.edu/documents>

⁶<https://archivio.quirinale.it/aspr/>

2 Related Work

The political domain has been studied in computational linguistics from various perspectives. Annotated corpora have been created to analyse rhetoric and metaphors in political communication (Cardie and Wilkerson, 2008; Ahrens et al., 2018), study the impact of speeches on the audience (Guerini et al., 2013; Thomas et al., 2006) and understand the relationship between ideology and linguistic complexity (Schoonvelde et al., 2019). Resources have also been developed to train and test automatic systems for several types of NLP tasks, such as persuasiveness prediction (Strapparava et al., 2010), sentiment and emotion analysis (Young and Soroka, 2012; Rheault et al., 2016), text classification (Yu et al., 2008), topic-based agreement detection (Menini et al., 2017) and recognition of ideological positions (Hirst et al., 2010).

Many research activities have recently dealt with the digitisation and release of corpora containing historical political texts. For example, the corpus of speeches given in the British Parliament from 1803 to 2005 (i.e. the Hansard Corpus) has been automatically tagged using the Historical Thesaurus Semantic Tagger (Piao et al., 2014; Wattam et al., 2014) and then a part of it has been semantically enriched with information about speakers and topics (Nanni et al., 2019). In addition, the Canadian Parliamentary Debates (1901-present) have been standardised, enriched and distributed within the “Digging into Linked Parliamentary Data” project (Beelen et al., 2017). The period from 1947 to 2017 is instead covered by a dataset of Dutch and Danish party congress speeches (Schumacher et al., 2019).

As for Italian, to the best of our knowledge, the only available comprehensive study of the language of Italian politicians is the one by Bolasco (2015). He analyses the parliamentary proceedings of the Italian Chamber of Deputies in the period 1953-2008 using the TalTac2 software⁷, thus providing a lexical and statistical analysis. Another project related to our work is “Voci della Grande Guerra” whose online platform allows to explore a corpus of documents related to the first World War including samples of parliamentary proceedings and political speeches (Lenci et al., 2016). Similarly to what we present in this paper, such documents have been automatically an-

notated and then partially revised by hand (De Felice et al., 2018). Compared with these two last works, our corpus is broader, having a multi-layered semantic analysis, and completely available for download in different formats, thus open to further analysis by the research community.

3 Corpus Description

Our corpus contains the complete collection of public documents by Alcide De Gasperi, the first Prime Minister of the Italian Republic and one of the founding fathers of the European Union. It includes 2,762 documents published between 1901 and 1954, for a total of around 3,000,000 tokens. The corpus is released as raw text, as XML with a minimal set of meta-data and associated key-concepts, and as CoNLL-like format, with additional information that have been fully or semi-automatically annotated (see Section 4). Texts, date and place of publication were automatically generated starting from the PDF files used to issue the volumes edited by Il Mulino. Each document of the collection was classified manually by a group of history scholars on the basis of a two-layered hierarchy that takes into consideration whether the text was originally released in an oral or written form, and its specific genre. It is important to note that different text genres correspond to different roles covered by De Gasperi during his life: e.g. daily press when he worked as a journalist for newspapers in Trentino, speeches in institutional venues when he was a Member of the Italian Parliament.

History scholars identified also four time spans to which each document can be assigned, that characterise different periods in De Gasperi’s life. These correspond to the four volumes of the printed edition and are used to split the corpus into different periods based on the date of publication:

Vol. I : De Gasperi was a journalist and a students’ leader. He was active mainly in Trento and in the Austrian Parliament (1901 – 1918).

Vol. II : De Gasperi founded Partito Popolare, became Parliament member in Rome and then left the Italian political life for several years after opposing the Fascist regime, working at the Vatican library and as a publicist (1919 – 1942).

Vol. III : De Gasperi founded the Christian-Democratic Party, became Prime Minister

⁷<http://www.taltac.it/>

Document	Type	Number
Written documents	Monographs / Prefaces	4
	Daily press	963
	Magazines	228
	Official documents	433
Speeches	Electoral / propaganda	473
	Party conferences	188
	Institutional venues	419
Not specified	Not specified	54

Table 1: Genre labels with corresponding statistics.

and was Italian delegate at the World War II peace conference (1943 – 1948).

Vol. IV : After Christian Democracy led by De Gasperi won the first general elections of the Italian Republic, he launched a plan of reforms to reconstruct Italy including social housing, labor policy and unemployment insurance (1949 – 1954).

4 Annotated Information

The annotations included in the release are:

- Lemma and PoS: the corpus has been lemmatised and PoS-tagged using the TextPro suite (Pianta et al., 2008). The module for the lemmatization is a rule-based system, whereas the part-of-speech annotation is statistical and has been trained on the EVALITA 2007 dataset (Tamburini, 2007) following the EAGLES tagset (Monachini, 1996).
- Person and place names: named entities have been tagged using the NER module included in TextPro and trained on the I-CAB corpus (Magnini et al., 2006). Geopolitical entities (GPEs) have also been geo-referenced using Nominatim⁸ (Clemens, 2015). The number of person and place names per volume is provided in Table 2.

After running the automatic modules, the output was uploaded in the ALCIDE platform (Moretti et al., 2016) and, through its navigation interface, we identified annotations that were systematically wrongly tagged, and fixed them manually. An evaluation of the automatic annotation is reported in Section 5.

In addition to the annotations previously mentioned, each document is assigned to a set of key-

⁸<https://nominatim.openstreetmap.org/>

concepts, that is a weighted list of n-grams representing the most important concepts of a text, automatically extracted using KD (Moretti et al., 2015).

5 Annotation Evaluation

We evaluated the quality of the automatic annotation produced by TextPro modules on a subset of our corpus. Indeed, since these modules were developed to perform best on contemporary texts, and typically trained on news, it is important to assess to what extent they can be reliably used on Italian documents of the XX Century in the political domain. To this end we manually annotated a gold standard made of documents written by De Gasperi between 1906 and 1911 for a total of 8,872 tokens. We chose texts belonging to the first period of De Gasperi’s life because they are the oldest in the corpus and therefore the most linguistically different from the texts used for training the modules. Results of the evaluation are compared with the ones obtained by TextPro on contemporary texts.

5.1 Lemmatization

Table 3 shows TextPro accuracy obtained on our gold standard compared with the ones reported in Aprosio and Moretti (2018) and calculated on the Universal Dependencies (UD) test set for Italian (Bosco et al., 2013). The drop of 0.7 points in accuracy is mainly due to some repeated anomalies of the module in the lemmatization of definite and indefinite articles (which are lemmatized using the labels “det” and “indet”, instead of singular masculine forms “il” and “uno”) and to the non-recognition of truncated words, such as “far”, “bel”, “andar”, “vuol”, not common in contemporary texts. Other sources of errors are the presence of obsolete terms, e.g. “libello”, “soziale”, “donde”, and the use of preterite (*passato remoto*, e.g. “andò”, “apparve”), a grammatical tense not very frequent in contemporary news. Most of previously mentioned anomalies have been fixed through a set of rules applied after data processing: after this correction, accuracy has risen to 0.97.

5.2 PoS Tagging

The presence of obsolete words, truncated forms and preterite verbs leads to errors also in the PoS tagger of TextPro. However, for this module the impact is less evident than for lemmatization: as

VOL I		VOL II		VOL III		VOL IV	
PER	GPE	PER	GPE	PER	GPE	PER	GPE
4,126	6,168	2,890	2,956	3,018	4,324	5,701	6,308
Gesù Cristo	Trento	Gesù Cristo	Italia	Palmiro Togliatti	Italia	Pietro Nenni	Italia
Augusto Avancini	Alto Adige	Mussolini	Roma	Pietro Nenni	Trieste	Palmiro Togliatti	Europa
Karl Lueger	Trentino	Leone XIII	Germania	Marshall	Russia	Tito	Trieste

Table 2: Occurrences of PER and GPE per volume, with three top-frequent entities for each category.

	UD Test Set	De Gasperi Corpus
	Accuracy	Accuracy
Lemma	0.96	0.89

Table 3: Comparison of lemmatization performance on the Italian UD test set and on our gold standard.

reported in Table 4, on De Gasperi’s documents the performance drop is only 0.1 points accuracy with respect to the results obtained on the UD test set. Table 5 gives details on the number and distribution of errors per grammatical category. Categories registering the higher quantity of mistaken tags are nouns, proper nouns, verbs and adjectives. Most mistakes concerning nouns are due to words capitalised to show formal respect towards highest representatives of the State or of the Church (e.g. “Vescovo”) and German common nouns that all have the initial capital letter.

	UD Test Set	De Gasperi Corpus
	Accuracy	Accuracy
PoS	0.96	0.95

Table 4: Comparison of PoS tagging performance on the Italian UD test set and on our gold standard.

Grammatical Category	#errors	%errors
Adjectives	62	15.54
Adverbs	24	6.02
Conjunctions	6	1.50
Demonstrative Adjectives	8	2.01
Prepositions	10	2.51
Pronouns	12	3.01
Relative Pronouns	1	0.25
Articles	11	2.76
Nouns	94	23.56
Proper Nouns	91	22.81
Verbs	73	18.30
Acronyms	6	1.50
Foreign Terms	1	0.25

Table 5: PoS-tagging errors per category.

5.3 Persons and GPEs

In Table 6 the performance of automatic recognition of persons (PER) and geo-political entities

(GPE) in De Gasperi’s documents is compared with the scores TextPro obtained in the EVALITA 2007 campaign (Speranza, 2007), when trained and tested on a newswire corpus. The tool shows a drop in performance on our gold standard only in the recognition of persons’ names (-0.16 F1 points), whereas place names seem to be more stable (+0.1 F1 points). In both categories, precision has decreased more than recall: to improve it, we manually checked the named entities detected by the automatic module in the whole corpus removing the wrong ones. We also verified the latitude and the longitude retrieved with Nominatim for all the GPEs assigning new correct coordinates to about 6% of them. Errors were mainly related to places that no longer exist or that have changed names after the death of De Gasperi, (e.g. “Prussia”, “Congo Belga”) and to little villages in the Trentino area (e.g. “Oltresarca”, “Termon”).

	EVALITA 2007 test set			De Gasperi corpus		
	P	R	F1	P	R	F1
PER	0.92	0.93	0.92	0.70	0.82	0.76
GPE	0.85	0.86	0.85	0.82	0.90	0.86

Table 6: Comparison of NER performance on news and on our gold standard.

6 Use Cases

The corpus has been used to perform a number of pilot studies, which have confirmed the potential of this kind of resource and could represent a starting point for further developments (Sprugnoli et al., 2016). Three of these studies are described in this Section.

A first analysis has been carried out with the goal of studying De Gasperi’s rhetoric strategy through his use of verb tenses, considered as an important marker of temporality (Sprugnoli et al., 2018). This study is based on the paradigm proposed by Chilton (2004), who includes time among the three axes of the political discourse together with space and modality.

We run the morphological analyzer included in TINT NLP Suite to recognise the tenses of all

verbs of the corpus. We then merge them into present, past and future tense and compare the distribution of the three classes across the four volumes. We observe that there is an evident difference between the use of verb tenses before and after 1943. Indeed, in the first two volumes past tenses are more frequently used, with a highly statistically significant difference with respect to volumes III and IV ($p < 0.001$ using Wilcoxon signed-rank test). On the other hand, after 1943 De Gasperi uses more present and future tense, again with high statistical significance. This can be explained by the fact that the last volumes contain many press reports describing the programmatic commitment of Christian Democracy as well as letters and telegrams sent by De Gasperi as Minister of Foreign Affairs, where the development of prospective collaborations is proposed. The last volume discusses also the reforms to be adopted for the reconstruction of the newly born Italian Republic and those about the forthcoming creation of a European Community. In general, after 1943 we observe a shift of focus from past events to the contemporary and future dimension.

A second analysis related to temporality deals with cited persons, which were linked to a Dbpedia entry using the Wiki Machine (Palmero Aprosio and Giuliano, 2016). Through this link, each person is associated with a *dbo:birthDate* and *dbo:deathDate* and then to a Past or Present label, again using the document date as a reference. Persons are considered part of the past if the referent was dead before the document publication time. Using the classification algorithm described in (Palmero Aprosio et al., 2017) we further assign a semantic category to each mention. A comparative analysis shows that contemporary persons are generally more cited than past ones, but also that the category of persons mentioned in the document changes significantly across the volumes: while in Volume I cited persons include politicians but also religious figures and artists, this range of figures decreases over time, with almost exclusively political figures mentioned in Volume IV. As an example, we report in Fig. 1 and Fig. 2 the top-cited persons in Vol. I and IV respectively: while in the early documents Beethoven, Dante and Nietzsche are highly cited, persons mentioned in the late documents include exclusively politicians and religious figures, all from present time or recent past. With reference to the previously cited

dimensions in Chilton (2004), this shift should be seen in the light of De Gasperi’s effort after 1943 to justify past and present policy, using mentioned persons to build a national ideology.

A third analysis focused on how temporal information is expressed in De Gasperi’s documents (Speranza and Sprugnoli, 2018). To explore this aspect we manually annotated ten newspaper articles, published in 1914 and related to the outbreak of the Great War, following the It-TimeML guidelines (Caselli et al., 2011). This resource has been used in the EVENTI task organized within EVALITA 2014 (Caselli et al., 2014) and is freely available online. The average number of annotated events and temporal relations in the documents written by De Gasperi is higher than in contemporary newspaper articles annotated following the same guidelines, whereas the density of temporal expressions is comparable. Other differences concern the type of events, temporal expressions and temporal relations present in the historical texts. For example, De Gasperi frequently uses events expressing personal opinions about the topics covered in the articles. The high presence of speculations influences the temporal structure of the texts: in many cases events are not ordered chronologically but presented as simultaneous with respect to the time of writing. Moreover, temporal expressions are mainly non-specific or fuzzy: a characteristic that is less evident in other corpora of contemporary texts, and that may be related to the more speculative nature of political texts.

7 Conclusions

In this paper we present the release of the corpus of Alcide De Gasperi’s public writings, including 2,762 documents and around 3 million tokens. We make available raw texts, XML files having a small set of metadata and key-concepts and CoNLL-like files with lemma, PoS, PER, GPE annotation together with the coordinates of place names. Based on an evaluation performed on all four annotation layers, we show that their quality is good, although annotation was performed automatically and only partially revised.

This is the first freely available corpus of this kind, and we hope that it can be used to foster research in political science, corpus linguistics and history, as well as to develop and test NLP systems using data that are different from widely used contemporary news.

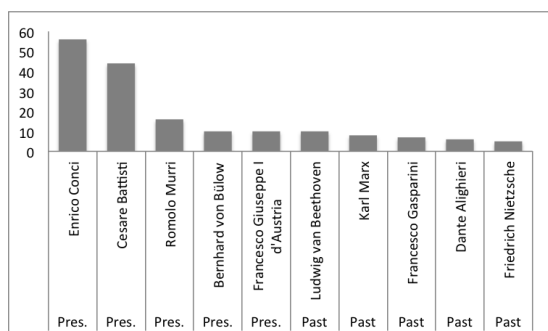


Figure 1: Past and present persons mentioned in Vol. 1.

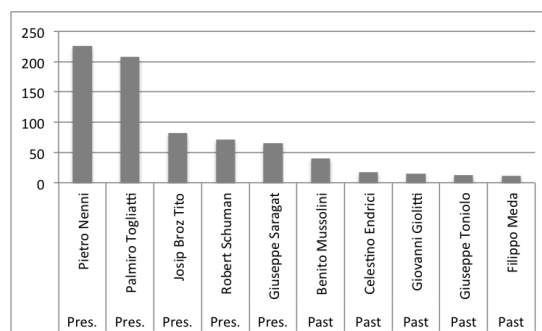


Figure 2: Past and present persons mentioned in Vol. 4.

Acknowledgments

We thank the colleagues from the Italian-German Historical Institute at Fondazione Bruno Kessler for their help in annotating De Gasperis corpus, and Edizioni Il Mulino, for giving access to the corpus and allowing its release. The project has been partially supported by Fondazione Cassa di Risparmio di Trento e Rovereto and Fondazione Cassa di Risparmio delle Province Lombarde.

References

- Kathleen Ahrens, Huiheng Zeng, and Shun-han Rebekah Wong. 2018. Using a Corpus of English and Chinese Political Speeches for Metaphor Analysis. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*.
- Alessio Palmero Aprosio and Giovanni Moretti. 2018. Tint 2.0: an All-inclusive Suite for NLP in Italian. In *Proceedings of the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018)*, Torino, Italy, December 10-12, 2018.
- Kaspar Beelen, Timothy Alberdingk Thijm, Christopher Cochrane, Kees Halvemaan, Graeme Hirst, Michael Kimmins, Sander Lijbrink, Maarten Marx, Nona Naderi, Ludovic Rheault, et al. 2017. Digitization of the Canadian parliamentary debates. *Canadian Journal of Political Science/Revue canadienne de science politique*, 50(3):849–864.
- Sergio Bolasco, 2015. *Sulla costruzione di un corpus per l'analisi automatica del linguaggio parlamentare dei leader*, chapter 5. Camera dei Deputati.
- Cristina Bosco, Montemagni Simonetta, and Simi Maria. 2013. Converting Italian Treebanks: Towards an Italian Stanford Dependency Treebank. In *7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 61–69. The Association for Computational Linguistics.
- Claire Cardie and John Wilkerson. 2008. Text Annotation for Political Science Research. *Journal of Information Technology & Politics*, 5(1):1–6.
- Tommaso Caselli, Valentina Bartalesi Lenzi, Rachele Sprugnoli, Emanuele Pianta, and Irina Prodanof. 2011. Annotating events, temporal expressions and relations in Italian: the It-TimeML experience for the Ita-TimeBank. In *Proceedings of the 5th Linguistic Annotation Workshop*, pages 143–151. Association for Computational Linguistics.
- Tommaso Caselli, Rachele Sprugnoli, Manuela Speranza, and Monica Monachini. 2014. EVENTI EVALUATION of Events and Temporal INFORMATION at Evalita 2014. In *Proceedings of the Fourth International Workshop EVALITA 2014*, pages 27–34.
- Paul Chilton. 2004. *Analysing political discourse: Theory and practice*. Routledge.
- Konstantin Clemens. 2015. Geocoding with open-streetmap data. *GEOProcessing 2015*, page 10.
- Irene De Felice, Felice Dell'Orletta, Giulia Venturi, Alessandro Lenci, and Simonetta Montemagni. 2018. Italian in the Trenches: Linguistic Annotation and Analysis of Texts of the Great War. In *Fifth Italian Conference on Computational Linguistics (CLiC-it 2018)*, pages 160–164. Accademia University Press.
- Alcide De Gasperi. 2006. Alcide De Gasperi nel Trentino asburgico. In *Scritti e discorsi politici di Alcide De Gasperi*, volume 1. Il Mulino.
- Alcide De Gasperi. 2008a. Alcide De Gasperi dal Partito popolare italiano all'esilio interno 1919-1942. In *Scritti e discorsi politici di Alcide De Gasperi*, volume 2. Il Mulino.
- Alcide De Gasperi. 2008b. Alcide De Gasperi e la fondazione della Democrazia cristiana, 1943-1948. In *Scritti e discorsi politici di Alcide De Gasperi*, volume 3. Il Mulino.
- Alcide De Gasperi. 2009. Alcide de Gasperi e la stabilizzazione della Repubblica 1948-1954. In *Scritti*

- e discorsi politici di Alcide De Gasperi*, volume 4. Il Mulino.
- Marco Guerini, Danilo Giampiccolo, Giovanni Moretti, Rachele Sprugnoli, and Carlo Strapparava. 2013. The new release of CORPS: A corpus of political speeches annotated with audience reactions. In *Multimodal Communication in Political Speech. Shaping Minds and Social Action*, pages 86–98. Springer.
- Graeme Hirst, Yaroslav Riabinin, and Jory Graham. 2010. Party status as a confound in the automatic classification of political speech by ideology. In *Proceedings of the 10th International Conference on Statistical Analysis of Textual Data (JADT 2010)*, pages 731–742.
- Alessandro Lenci, Nicola Labanca, Claudio Marazzini, and Simonetta Montemagni. 2016. Voci della Grande Guerra An Annotated Corpus of Italian Texts on World War I. *Italian Journal of Computational Linguistics*, pages 101–108.
- Bernardo Magnini, Emanuele Pianta, Christian Girardi, Matteo Negri, Lorenza Romano, Manuela Speranza, Valentina Bartalesi Lenzi, and Rachele Sprugnoli. 2006. I-CAB: the Italian Content Annotation Bank. In *LREC*, pages 963–968.
- Stefano Menini, Federico Nanni, Simone Paolo Ponzetto, and Sara Tonelli. 2017. Topic-based agreement and disagreement in US electoral manifestos. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2938–2944.
- Monica Monachini. 1996. ELM-it: EAGLES specifications for Italian morphosyntax lexicon specification and classification guidelines. Technical report, Centre National de la Recherche Scientifique Paris, France.
- Giovanni Moretti, Rachele Sprugnoli, and Sara Tonelli. 2015. Digging in the Dirt: Extracting Keyphrases from Texts with KD. In *Proceedings of the Second Italian Conference on Computational Linguistics (CLiC-it 2015)*.
- Giovanni Moretti, Rachele Sprugnoli, Stefano Menini, and Sara Tonelli. 2016. ALCIDE: Extracting and visualising content from large document collections to support Humanities studies. *Knowledge-Based Systems*, 111:100–112.
- Federico Nanni, Stefano Menini, Sara Tonelli, and Simone Paolo Ponzetto. 2019. Semantifying the UK Hansard (1918-2018). In *Proceedings of JCDLI9*.
- Alessio Palmero Aproso and Claudio Giuliano. 2016. The Wiki Machine: an open source software for entity linking and enrichment. *ArXiv e-prints*, September.
- Alessio Palmero Aproso, Sara Tonelli, Stefano Menini, and Giovanni Moretti. 2017. Using Semantic Linking to Understand Persons’ Networks Extracted from Text. *Front. Digital Humanities*, 2017.
- Emanuele Pianta, Christian Girardi, and Roberto Zanolini. 2008. The TextPro Tool Suite. In *Proceedings of Language Resources and Evaluation Conference*, pages 2603–2607, Marrakech, Morocco.
- Scott Piao, Fraser Dallachy, Alistair Baron, Paul Rayson, and Marc Alexander. 2014. Developing the Historical Thesaurus Semantic Tagger. In *The Digital Humanities Congress 2014*.
- Ludovic Rheault, Kaspar Beelen, Christopher Cochrane, and Graeme Hirst. 2016. Measuring emotion in parliamentary debates with automated textual analysis. *PloS one*, 11(12):e0168843.
- Martijn Schoonvelde, Anna Brosius, Gijs Schumacher, and Bert N Bakker. 2019. Liberals lecture, conservatives communicate: Analyzing complexity and ideology in 381,609 political speeches. *PloS one*, 14(2):e0208450.
- Gijs Schumacher, Daniel Hansen, Mariken ACG van der Velden, and Sander Kunst. 2019. A new dataset of Dutch and Danish party congress speeches. *Research & Politics*, 6(2):2053168019838352.
- Manuela Speranza and Rachele Sprugnoli. 2018. Annotation of Temporal Information on Historical Texts: a Small Corpus for a Big Challenge. *Formal Representation and the Digital Humanities*, page 203.
- Manuela Speranza. 2007. EVALITA 2007: The Named Entity Recognition Task. In *Proceedings of the EVALITA 2007 Workshop on Evaluation of NLP Tools for Italian*, pages 66–68, Rome, Italy.
- Rachele Sprugnoli, Giovanni Moretti, Sara Tonelli, and Stefano Menini. 2016. Fifty years of european history through the lens of computational linguistics: the de gasperi project. *IJCol-Italian journal of computational linguistics*, 2(2):89–100.
- Rachele Sprugnoli, Giovanni Moretti, and Sara Tonelli. 2018. Temporal Dimension in Alcide De Gasperi: Past, Present and Future in Historical Political Discourse. In *AIUCD 2018 - Book of Abstracts*, pages 77–80.
- Carlo Strapparava, Marco Guerini, and Oliviero Stock. 2010. Predicting Persuasiveness in Political Discourses. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC’10)*, pages 1342–1345.
- Fabio Tamburini. 2007. Evalita 2007: The Part-of-Speech Tagging Task. *Intelligenza artificiale*, 4(2):57–73.

- Matt Thomas, Bo Pang, and Lillian Lee. 2006. Get out the vote: Determining support or opposition from Congressional floor-debate transcripts. In *Proceedings of the 2006 conference on empirical methods in natural language processing*, pages 327–335. Association for Computational Linguistics.
- Stephen Wattam, Paul Rayson, Marc Alexander, and Jean Anderson. 2014. Experiences with Parallelisation of an Existing NLP Pipeline: Tagging Hansard. In *LREC*, pages 4093–4096.
- Lori Young and Stuart Soroka. 2012. Affective news: The automated coding of sentiment in political texts. *Political Communication*, 29(2):205–231.
- Bei Yu, Stefan Kaufmann, and Daniel Diermeier. 2008. Classifying party affiliation from political speech. *Journal of Information Technology & Politics*, 5(1):33–48.

Reflexives, Impersonals and Their Kin: a Classification Problem

Kledia Topciu

Università degli Studi Di Siena

Via Roma 56

I-53100 Siena (Italy)

kledia.topciu@student.unisi.it

Cristiano Chesi

NETS - IUSS

P.zza Vittoria 15

I-27100 Pavia (Italy)

cristiano.chesi@iusspavia.it

Abstract

Despite the fact that true reflexives always require a local antecedent, attempting an automatic referential resolution is often far from trivial: in many languages, reflexives are morphologically indistinguishable from impersonals and both particles are sensitive to the syntactic structure in a non-trivial sense. Focusing on Italian, we annotated part of the Repubblica Corpus to attempt an automatic classification of the reflexive and impersonal *si* constructions. In this preliminary study we show that the accuracy of the automatic classification methods that do not use any relevant structural information are rather modest. A thoughtful discussion of the structural analysis required to distinguish among different contexts is provided, in the end suggesting that these structural configurations are not easily recoverable using a purely distributional approach.

1. Introduction

The non-triviality of reflexive/impersonal constructions in Italian is exemplified in (1):

- (1) a. Ada_i si_i presentò.
 A_i SI_i introduced_{3-SG-PAST}
 ‘A. introduced herself.’
 b. $Si_{i/*j}$ presentò Ada_i .
 $SI_{i/*j}$ introduced_{3-SG-PAST} A_i
 ‘A. introduced herself.’
 c. $Si_{i/j}$ presentò ad Ada_i .
 $SI_{i/j}$ introduced_{3-SG-PAST} to A_i
 ‘S/He introduced him/herself to A.’

- d. pro_i $Si_{i/*j}$ tolse la giacca_i.
 pro_i $SI_{i/*j}$ took_{3-SG-PAST} off the jacket
 ‘S/He took off the jacket.’
 e. Il compagno_j di Ada_i $si_{i/j}$ presentò.
 The friend_j of A_i $SI_{i/j}$ introduced_{3-SG-PAST}
 ‘A.’s friend introduced her/him-self.’
 f. Riconosciuto il compagno_j di Ada_i ,
 pro_k $si_{i/*j/k}$ presentò.
 Recognized_{3-SG-P.PART} the friend_j of A_i ,
 pro_k $SI_{i/*j/k}$ introduced_{3-SG-PAST}.
 ‘Once s/he recognized A.’s friend,
 s/he introduced her/him-self.’
 g. $Si_{generic}$ pensa sempre a salvarsi la pelle.
 $SI_{generic}$ thinks always to save_{INF-REFL} the skin
 ‘We always think about saving our own skin.’

Expecting the co-referential DP to be always “immediately to the left” of the reflexive form quickly leads to wrong predictions: if this generalization might seem sufficient in (1a) this is bluntly wrong in (1b), where we need to assume an empty referent (*pro*, Rizzi 1986) before the reflexive (see §1.1). Moreover, we should accept that the coreferential DP can be placed sometimes to the right of the predicate (structurally speaking, *pro* and post-verbal subject options are related, Belletti 2002); in this case, the (focalized/dislocated) post-verbal subject is a good candidate, (1b). Being “the closest DP” is however not a sufficient condition as suggested by the examples (1c-d). Hence, the null subject hypothesis as well as a structural analysis unravelling the role of each DP surrounding the predicate is requested, for the identification of the correct local binding domain (1e-f). Last but not least, a proper classification of the predicate admitting a reflexive or an impersonal pronoun is needed (1g). Under this perspective, we decided to run a little experiment to verify the consistency of a “usage-based” approach (Tomasello 2003) in this specific context and consider whether the “structural

analysis” (Chomsky 1995; 2008) can be proved to be an outdated approach for the classification of the distinct kinds of *si*. In the remaining part of this introduction we will present the (possibly outdated) structural analyses proposed for reflexive (§1.1) and impersonal (§1.2) clitic *si*. We will then present our experiment consisting of the annotation of a small fragment of the Repubblica Corpus (Baroni et al. 2004) that we used to train and test a set of Machine Learning classification algorithms (§2). Results presentation (§3) and their discussion (§4) will follow.

1.1 The reflexivization configuration

A popular structural analysis of reflexives is the unaccusative one: under this perspective, the subject of reflexives is an underlying object (just like the subject of unaccusatives) which has to raise to the subject position for Case reasons (reflexive morphology absorbs its Case). Two main variants of this approach are discussed in the literature: a lexical and a syntactic one. The lexical version predicts that the external argument is absorbed in the lexicon (Marantz 1984 and Grimshaw 1990), while the syntactic one proposes that the external argument is present in syntax via the reflexive clitic *se* (Kayne 1988, Pesetsky 1995, Sportiche 1998).

A different analysis is proposed by Reinhart & Siloni (1999, 2005): reflexives should be unergative entries since unaccusativity tests (e.g. *ne* cliticization, (2b)) fail with reflexive constructions:

- (2) a. Ne sono arrivati tre.
 of+them_{cl} are arrived three
 ‘Three of them arrived.’
 b. *Se ne sono vestiti tre.
 SI of+them_{cl} are dressed three
 ‘Three of them got dressed.’

Since the internal argument only can be cliticized and the reflexive verb fails the *ne* test, we conclude that the subject of the reflexives is an external argument, unlike the subject of unaccusatives. Another test helping us to tease apart external from internal argument structures is reduced relatives modification: when the modification is implemented via past participle, this does not allow for predicates with an external argument. The reduced relative in (3a) contains a reflexive predicate, while the one in (3b) is an impossible cliticization of a transitive reflexive past participle.

- (3) a. Il bicchiere rotti ieri apparteneva a mio
 nonno.
 the glass broken-him/herself yesterday
 belonged to my grandfather

- b. *L'uomo lavatosi ieri è mio nonno.
 the man washed-him/herself yesterday is
 my grandfather

A robust evidence supports the idea that the subject of reflexive verbs patterns with the subject of unergatives, hence confirming its external argument nature (but see Pescarini 2015:42ff).

Kayne (1975) observes that reflexives occur in environments where transitive verbs are disallowed, e.g. in French causative constructions: when the verb embedded under the causative verb *faire* ‘make’ is a transitive verb (4a), its subject must be introduced by the preposition *a* ‘to’; when the lower verb is intransitive or reflexive, its subject cannot be introduced by *a* (4b/c).

- (4) a. Je ferai laver Jean *(a) Luc.
 I make_{FUT} wash Jean to Luc.
 ‘I will make Jean wash Luc.’
 b. Je ferai courir (*a) Jean.
 I make_{FUT} Jean run.
 ‘I will make Jean run.’
 c. Je ferai se laver (*a) Jean.
 I make_{FUT} SE wash Jean.
 ‘I will make Jean wash himself.’

When the lower verb is reflexive, its subject appears without the preposition, exactly like the subject of unergative verbs. Therefore, reflexive verbs are not transitive entries either.

Reinhart & Siloni (2005) suggest that these reflexive constructions are unergative entries derived from their transitive alternate by a reduction operation targeting the internal argument (identified with the external one). They take verbal reflexivization even further and propose a *lexicon-syntax parameter*: arity operations (on θ -roles) can apply either to the syntax or to the lexicon. Reflexivization is essentially the same phenomenon cross-linguistically, that is, two available θ -roles are assigned to the same syntactic argument, or, better said, the operation of reflexivization takes two θ -roles and forms one complex θ -role.

The distinctions follow from two different modes of operation: a lexical mode and a syntactic one. Languages such as Hebrew, English, Russian and Dutch have the parameter set to “lexicon”, while in Romance languages, Greek and German the “syntax” value of the parameter is set. In the syntactic option (which is relevant here), what is to become a reflexive verb leaves the lexicon with the same number of θ -roles, which need to be assigned, as the basic verbal entry. Since the clitic itself cannot be viewed as an argument (the lack of Case blocks its merge), the “extra” θ -role has to be explained by an arity reduction operation.

In conclusion, an automatic classification algorithm, attempting at identifying the typology of the *si* reflexive pronoun, should necessarily have access to the subcategorization verbal frame and postulate an arity-reduction as suggested by (Reinhart & Siloni 2005). If this information is not available as lexical resource, we might try to rely on structural cues to infer the correct argument structure (as in Merlo & Stevenson 2001, Basili et al 1997 or Ienco et al. 2008). On the other hand, if statistical cues would be available, annotating them overtly would be unnecessary.

1.2 Impersonal *si* constructions

- (5) a. In Italia *si* mangia troppo.
In Italy *si* eats_{3rdSG} too much
'In Italy, people eat too much.'
b. In Italia *pro* mangia troppo.
In Italia *pro* reads_{3rd-SG} a lot
'In Italia he/she reads a lot'

As for the argumental status of *si*, there is a large disagreement in the linguistic community: Cinque (1988) proposes the existence of two different *si* items: the presence of *si* is usually restricted to finite clauses, however, it is also permitted in certain untensed clauses, namely in

- (6) Non *essendosi* ancora scoperto il colpevole...
 not being_{GERUND}-SI yet discovered_{P-PART-SG-MASC} the culprit_{SG-MASC}
 ‘Not having yet discovered the culprit...’
- (7) Sembra non *essersi* ancora scoperto il colpevole ...
 seems_{3RD-SG} not being-SI yet discovered_{P-PART-SG-MASC} the culprit_{SG-MASC}
 ‘It seems it hasn’t yet been discovered the culprit.’

Dobrovie-Sorin (1998, 1999) argues that it is not necessary to postulate this: according to her, what Cinque calls a +arg *si* is actually a middle passive Accusative *si*. The only Nominative *si* is Cinque's -arg *si*. She argues that *si* is not licensed in non-finite clause because it is a Nominative clitic and, in Italian, Nominative clitics are not allowed in non-finite clauses. Only transitive and unergative Aux-to-Comp and Raising structures allow *si* as Accusative. Dobrovie-Sorin tries to unify all the uses of SE in Romance languages and assumes that *si* is not a special lexical item that absorbs a θ -role or Case. Her analysis accounts for special cases, such as Romanian, which has *si* constructions but doesn't have Nominative clitics. Italian *si* constructions, on the other hand, rely either on Nominative (8) or Accusative (which also includes reflexive configurations) (9).

In (8), *si* is an anaphor and if we assume a restricted theory of binding, the anaphoric status of the clitic is transferred to its trace. The indexing configuration corresponds to a single argument, the Theme. On the other hand, the *si* in (9) is not an anaphor and therefore imposes no relation between the subject and object positions; it binds an empty category in the subject A-position.

modern Italian there are two reflexive *si* constructions: a *passive* one and an *impersonal* one (the reader should refer to Pescarini 2015 for a more detailed discussion of a richer classification). The first one, exemplified in (10b), is characterized by the cancelation of the subject (10a) and the transformation of the direct object into the grammatical subject (triggering agreement); the derived grammatical subject can occur also in the canonical preverbal position (10c):

- (10) a. Il preside ha consegnato i diplomi.
The dean has awarded the diplomas
b. *Si* sono consegnati i diplomi.
SI_{generic} are awarded the diplomas
'Diplomas got awarded'
c. I diplomi *si* consegnano (agli studenti).
the diplomas *SI_{generic}* awarded
(to the students)
'Diplomas are getting awarded
(to the students)'

This construction is only possible with (di)transitive predicates, since the promotion of the object to the grammatical subject role is only available when a direct object is available.

On the other hand, the impersonal version of *si* does not induce the promotion of the internal argument to the grammatical subject role and in fact this construction is available without any verbal class restriction:

- (11) a. *Si* guarda la partita
SI_{generic} watches the game
'We watch the game'
b. *Si* dorme
SI_{generic} sleeps
'We sleep'
c. *Si* cade
SI_{generic} falls
'We fall'

In sum, with the impersonal *si* construction, the subcategorization verbal frame (i.e. the verbal argumental structure) could help in isolating the passive *si* construction, but not the impersonal one. As for reflexive *si*, the full argument structure must be identified and then either the passive strategy (deletion and promotion) or the impersonal one (simple deletion) considered. As a consequence of the null subject option in Italian, the difference between impersonal and passive *si* is often blurred.

2. Materials and methods

From Repubblica Corpus (Baroni et al 2004), we extracted all contexts in which the “*si*” lemma was present: 2.737.558 contexts are returned by the simple query including a left and right context of maximum 8 words around the *si* + predicate cluster; each left and right context was cut at full stops, colons, semi colons, exclamative and question marks, whenever those were found within the 8 tokens context. The tagset used in the Repubblica Corpus neither distinguishes among reflexive and various types of impersonal forms (“CLI/*si*” is the generic tag used) nor among different verbal classes with respect to their argumental structure (only VB for “be”, VH for “have”, and VV for other verbs are included). We then decided to annotate manually the first 2.000 contexts returned by our query (0,07% of the total) using the following scheme much simplified with respect to the structural asymmetries revealed by the discussion in §1: **I** (impersonal), **L** (local, DP immediately preceding “*si*” is the correct one), **PV** (post-verbal: the first DP after the predicate following “*si*” is the correct co-referent) and **LM** (the DP immediately preceding, in the hierarchical sense, the reflexive “*si*” is the correct one, but such DP is “modified” by a PP or a relative clause) and **A** (the referent is not present/retrievable in the extracted context; these are in the great majority pro-drop cases, in just two cases the referent was lexically realized outside the context isolated). Both authors annotated independently the corpus and discussed about the disagreement cases (less than 1% of the sample) in order to find an agreement in the annotation. Table 1 indicates the distribution of the classes across the annotated corpus fragment, while Table 2 exemplifies the classification. Due to the simplicity of this classification (that essentially focus on the identification of the reflexive antecedent, if present/necessary), we would expect a better performance compared to any richer classification, which is apparently necessary according to the structural analysis previously discussed.

annotation	# of contexts	%
I	332	16.6
L	994	49.7
LM	417	20.8
PV	183	9.15
A	74	3.7

Table 1. Distribution of the annotated categories across the sample.

annotation	example
I	<i>si è deciso di ridurre il deficit</i> <i>we decided to reduce the deficit</i>
L	<i>[i fedeli]_i si_i sono tuttavia sciolti</i> <i>the faithfuls, nevertheless, split up</i>
LM	<i>[il vertice di Dublino]_i si_i è dimostrato</i> <i>the Dublin summit proved to be ...</i>
PV	<i>nel cortile si_i stendono [le stuoie]_i</i> <i>in the courtyard the mats unfolded</i>
A	<i>per 16 anni si_i è occupato dei processi</i> <i>for 16 years [he] took care of the trials</i>

Table 2. Sample annotation using 5 categories.

2.1 Classifiers descriptions

Under the “usage-based” approach the disambiguation (i.e. the interpretation of the correct referent, if necessary) of the distinct *si* constructions should be possible on the basis of the purely statistical distribution of the (implicit) features across the corpus (Tomasello 2003 and related works). To test this hypothesis we created a set of classifiers using the Weka environment (Frank et al 2016). 4 different classifiers are used including the original extracted context of maximum 8 words before and after the clitic *si* + predicate cluster (Table 3): pure Bag-of-Words (BoW) approach was used for the first two classifiers, one with only the left context included, the other with both left and right context; then we manipulated the left context classifier substituting the words with their POS (classifier C3-POS-L) and with a more coarse set of POS tags (C4-CPOS-L). POS and CPOS annotation are obtained using a free online tool (*ItaliaNLP REST API*, Cimino & Dell’Orletta 2016).

Class. ID	Approach	Context
C1-BOW-L	BoW	Left context
C2-BOW-LR		Left & Right context
C3-POS-L	POS	Left context
C4-CPOS-L	CPOS	Left context

Table 3. Classifier description

2.2 Classification algorithms

Given the baseline classification of 49.7% of accuracy, obtained by choosing always the reflexive local class (L classification), we compared Naïve Bayesian algorithms (i.e. NaïveBayes, *n.bayes* in table 4, and NaïveBayesMultimodal, *n.bayes.mul.* in table 4) with a decision tree-based algorithm (i.e. *J48*) and then with both 3 layers convoluted (with LSTM layer; *conv.net* in table 4) and simple recurrent

neural networks using Weka wrappers for Deeplearning4j 1.5.13 (*srnn.net* in table 4) for a total of 5 classifiers. We run our experiments within Weka 3.8.3 environment with CUDA 10.1 GPU nVIDIA support. Word embeddings are built using a larger fragment of left and right contexts (+/-10 words at most, breaking the left/right context at full stops) extracted from Repubblica corpus including the “si” seed (first 1.000.000 sentences returned using the publicly available Sketch Engine search interface).

3. Results

The results of the classification tests are reported in table 4. The accuracy indicates the rate of correct classifications and the standard deviation running 10 experiments with cross-fold validation (standard deviation is indicated) and the significance is expressed with respect to the baseline: ♡ indicates that the accuracy is significantly better than baseline, ♢ significantly worse and no sign means no significant difference (pair-wise comparison using corrected resampled T-Test, Witten & Frank 2005).

Class. ID	Algorithm	Accuracy (SD)	Sign.
<i>baseline</i>		<i>49.70%</i>	
C1-BOW-L	n.bayes	56.95% (2.79)	♡
	n.bayes.mul.	54.28% (2.03)	♡
	J48	58.34% (2.48)	♡
	conv.net	51.88% (1.44)	♡
	srnn.net	39.63% (11.79)	♢
C2-BOW-LR	n.bayes	49.21% (3.40)	
	n.bayes.mul.	51.61% (1.17)	♡
	J48	48.66% (2.53)	
C3-POS-L	conv.net	49.77% (0.41)	
	srnn.net	39.05% (12.77)	♢
	n.bayes	54.49% (2.35)	♡
	n.bayes.mul.	53.26% (1.99)	♡
	J48	60.76% (2.97)	♡
C4-CPOS-L	conv.net	57.58% (1.98)	♡
	srnn.net	43.52% (7.17)	♢
	n.bayes	59.96% (2.85)	♡
	n.bayes.mul.	50.89% (1.03)	♡
	J48	61.49% (3.08)	♡
	conv.net	49.70% (0.25)	
	srnn.net	44.20% (6.17)	♢

Table 4. Classification accuracy results

In both left and left-right context classifiers, BoW approach (C1-BOW-L and C2-BOW-LR) is clearly not sufficient to solve the classification problem; the introduction of a right context (C2-BOW-LR) significantly reduces the performance of the classifier. Notice that in almost 10% of the cases the availability of the referent is post-verbal (PV classification). Decision trees (J48), overall, perform better ($M=58.34\%$ $SD=2.48$) but this performance represents a significant improvement only with C1-BOW-L and C4-CPOS-L classifiers. None of the deep learning approaches (conv.net and srnn.net) are significantly better than decision trees (in some cases SRNs perform significantly worse). The best absolute performance in obtained substituting words with coarse POS (C4-CPOS-L). In this case J48 obtains the best accuracy ($M=61.49\%$ $SD=3.08$).

4. Discussion

In this paper, we discussed the nature of some *si* constructions in Italian, suggesting that, despite their apparent simplicity, their structural intricacies require a deep syntactic analysis for identifying correctly the typology of the clitic in various contexts and retrieve, when necessary, a proper referent. Also using a simplified set of five classes (I = impersonal; L = local immediately preceding coreferential DP; PV = local, immediately post-verbal coreferential DP; LM = local preceding coreferential DP but with prepositional phrase or relative clause modification; A = absent referent), we demonstrated that, using an annotated sample of the Repubblica corpus, no classifier has exceeded the performance of 61.49% of accuracy. This is well below any human reasonable performance (as suggested by the 99% agreement in classification between annotators). These results, even though still based on a small fragment of the Repubblica Corpus, extend Chesi & Moro (2018) original considerations using a wider dataset and more advanced ML algorithms. These results showed that neither the algorithms used nor the extension of the context (both left and right) helped in classifying correctly the instances of “*si*” when the referent had to be retrieved non-locally or in impersonal “*si*” cases. Replacing the words with their POS mildly helped in improving the performance of some classifiers (especially using the coarse tagset), with decision tree classifier (J48) obtaining the best performance (on average) across the tests.

Given the poor performance of the classifiers tested, we concluded that the “usage-based” intuition is not sufficient here to account for the acquisition of the discriminative capabilities any

Italian native speaker owns and that enable her/him to identify correctly the relevant referent both pre- and post-verbally, even in the case of complex subjects (referent DPs modified by prepositional phrases or relative clauses), as well as its unnecessary (in generic/impersonal readings) or its recovery in case of pro-drop. We might expect then that a richer syntactic annotation could help to boost the automatic classification results in accordance with the structural analysis summarized in §1.1 and §1.2: first, a verbal subcategorization specification properly describing the predicate argument structure could be useful, then a correct analysis of the subject phrase structure, including agreement cues should be used, as well as a richer classification of temporal/modal adverbials/modifiers.

As suggested by an anonymous reviewer, information structure, which is largely obliterated in written texts, is expected to disambiguate between reflexive and impersonal constructions: for instance, non-dislocated preverbal subjects (L(M) in our classification) should be ruled out in impersonal constructions (see Raposo & Uriagereka 1996); moreover, non-focalized (or right-dislocated) postverbal subjects (PV in our classification) should be ruled out in reflexive constructions. Then, despite the fact that prosody/information structure cannot be assessed within a corpus-based study, we might expect an improvement of the classifiers performance considering some relevant features associated to these configurations: e.g. post-verbal subject annotation in connection with the verbal class and adverbials placement between the subject and verb indicating a dislocated subject.

A follow up of this study should test these predictions and, possibly, extend the study to the whole Repubblica corpus, confirming (or disconfirming) our preliminary results that suggest we cannot avoid a deep structural analysis of these constructions to classify (and interpret) them correctly.

References

- Baroni, Marco, Silvia Bernardini, Federica Comastri, Lorenzo Piccioni, Alessandra Volpi, Guy Aston, and Marco Mazzoleni. 2004. Introducing the La Repubblica Corpus: A Large, Annotated, TEI (XML)-compliant Corpus of Newspaper Italian. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC 2004)*.

- Basili, Roberto, Maria Teresa Pazienza, and Michele Vindigni. 1997. Corpus-driven unsupervised learning of verb subcategorization frames. *Congress of the Italian Association for Artificial Intelligence*. Springer, Berlin, Heidelberg.
- Belletti, Adriana. 2002. Aspects of the low IP area. Forthcoming in *The structure of IP and CP. The Cartography of Syntactic Structures*, vol. 2, L. Rizzi (ed.). New York: Oxford University Press.
- Burzio, Luigi 1992. On the morphology of reflexives and impersonals. *Theoretical analyses in Romance linguistics*. Amsterdam: Benjamins, 399-414.
- Chesi, Cristiano, & Moro, Andrea 2018. Il divario (apparente) tra gerarchia e tempo. *Sistemi intelligenti*, 30(1), 11-32.
- Chomsky, Noam 1995. *The minimalist program*. Cambridge, MA: MIT press.
- Cimino, Andrea, Dell’Orletta, Felice. 2016. “Building the state-of-the-art in POS tagging of Italian Tweets”. In *Proceedings of EVALITA ’16, Evaluation of NLP and Speech Tools for Italian*, 7 December, Napoli, Italy.
- Cinque, Guglielmo 1988. On si constructions and the theory of arb. *Linguistic inquiry*, 19(4), 521-581.
- Dillon, Brian, Alan Mishler, Shayne Sloggett, and Colin Phillips. 2013. Contrasting intrusion profiles for agreement and anaphora: experimental and modeling evidence. *J. Mem. Lang.* 69, 85–103.
- Dobrovie-Sorin, Carmen. 1998. Impersonal se constructions in Romance and the passivization of unergatives. *Linguistic Inquiry*, 29(3), 399-437.
- Frank, Eibe, Mark A. Hall, and Ian H. Witten. 2016. The WEKA Workbench. Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques", Morgan Kaufmann, Fourth Edition, 2016.
- Grimshaw, Jane. 1990. *Argument Structure*. MIT Press, Cambridge, MA.
- Ienco, Dino, Serena Villata, and Cristina Bosco. 2008. Automatic extraction of subcategorization frames for Italian. In *LREC08*, pp. 2094-2100. European Language Resources Association (ELRA)
- Marantz, Alec. 1984. *On the Nature of Grammatical Relations*. MIT Press, Cambridge.
- Merlo, Paola and Stevenson, S Suzanne, 2001. Automatic verb classification based on statistical distributions of argument structure. *Computational Linguistics*, 27(3), pp.373-408.
- Pescarini, Diego. 2015. *Le costruzioni con si. Italiano, dialetti, lingue romanze*. Roma: Carocci.
- Pesetsky, David. 1995. *Zero Syntax*. MIT Press, Cambridge, MA
- Raposo, Eduardo & Juan Uriagereka. 1996. Indefinite SE. *Natural Language and Linguistic Theory* 14: 749—810.
- Reinhart, Tania, & Siloni, Tal. 2005. The lexicon-syntax parameter: Reflexivization and other arity operations. *Linguistic inquiry*, 36(3), 389-436.
- Rizzi, Luigi. (1986). Null objects in Italian and the theory of 'pro'. *Linguistic inquiry*, 17(3), 501-558.
- Salvi, Giampaolo 2018. La formazione della costruzione impersonale in italiano. *Linguistica: Revista de Estudos Linguísticos da Universidade do Porto*, 3, 13-37.
- Sportiche, Dominique. 1998. *Partitions and atoms of clause structure: Subjects, agreement, Case and clitics*. New York: Routledge.
- Tomasello, Michael. 2003. *Constructing a language: A usage-based theory of language acquisition*. Cambridge, MA: Harvard University press.
- Witten, Ian, H. and Eibe Frank 2005. *Data Mining: Practical machine learning tools and techniques*. 2nd edition Morgan Kaufmann, San Francisco.

Annotation and Analysis of the PoliModal Corpus of Political Interviews

Daniela Trotta Università di Salerno dtrotta@unisa.it	Sara Tonelli FBK satonelli@fbk.eu	Alessio Palmero Aprosio FBK aprosio@fbk.eu	Annibale Elia Università di Salerno elia@unisa.it
--	--	---	--

Abstract

English. In this paper, we present the first available corpus of Italian political interviews with multimodal annotation, consisting of 56 face-to-face interviews taken from a political talk show. We detail the annotation scheme and we present a number of statistical analyses to understand the relation between these multimodal traits and language complexity. We also exploit the corpus to test the validity of existing studies on political orientation and language use, showing that results on our data are not as clear-cut as on English ones.¹

1 Introduction

In the context of a political interview, the host, typically a journalist, acts as a representative of the audience. This means that, if a politician manages to convince or deal with the criticism that the host addresses, then her/his trustworthiness, reliability and credibility will be easily established. In this situation, a politician is judged not only based on one's arguments and rhetorical choices, but also on the attitude, self-confidence, and in general on an overall convincing behaviour. For example, if a politician seems to be conversationally dominant and manages interruptions to a satisfactory degree, it is more likely that the host, and therefore the audience, will be convinced by the arguments put forward by the interviewee. For this reason, analysing the combination of verbal and non-verbal elements in a political interview could be very interesting for scholars in political science and communication science, and in general to study consensus mechanisms. In this light, we present the first multimodal corpus of political

interviews in Italian, and analyze how the combination of verbal and non-verbal elements can shed new light into political agendas and politicians' attitude. By 'multimodal' we mean that the corpus is composed of manual transcriptions of interviews broadcast on TV and annotated with information not only about the linguistic structure of the utterances but also about non-verbal expressions².

The corpus, which we call PoliModal, addresses the need to make up for the lack of Italian linguistic resources for political-institutional communication and is annotated in XML following the standard for the transcriptions of speech TEI Guidelines for Electronic Text Encoding and Interchange³. In all transcripts, interviewers, interviewees and other guests' turns have been enriched with the manual annotation of non-lexical and semi-lexical aspects such as breaks, interruptions, false starts, overlaps, interjections, etc. Furthermore, additional linguistic traits related to language complexity, use of pronouns and persons' mentions have been automatically tagged, enabling an in-depth analysis of speakers' attitude and communication strategy. In this work we present not only the corpus, which is made freely available at the link <https://github.com/dhfbk/InMezzoraDataset>, but also an analysis that, combining verbal and non-verbal elements, shows how these traits contribute to making an interview more or less convincing.

2 Related work

In recent years, political language has received increasing attention, especially in the Anglo-Saxon

¹Copyright ©2019 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

²According to (Allwood, 2008): "The basic reason for collecting multimodal corpora is that they provide material for more complete studies of 'interactive face-to-face sharing and construction of meaning and understanding' which is what language and communication are all about".

³P5: Guidelines for Electronic Text Encoding and Interchange. See more <https://tei-c.org/release/doc/tei-p5-doc/en/html/TS.html#TSSAPA>

and American world, where it is possible to have free access to speech transcriptions from government portals and personal foundation websites, e.g. White House portal, William J. Clinton Foundation, Margaret Thatcher Foundation. This has fostered research on political and media communication and persuasion strategies (Guerini et al., 2010; Esposito et al., 2015).

However, not all languages are well represented in this kind of studies. According to LRE Map⁴ there are currently 24 monolingual corpora for Italian, two of which concern spoken language, i.e. VoLIP (Alfano et al., 2014) and LUNA corpus (Dinarelli et al., 2009), and one multimodal, named ImagAct-ItalWorNet-Mapping (Bartolini et al., 2014); no entry includes an Italian corpus for the political domain. Furthermore, researchers in Italian politics have mainly focused on political communication in the verbal modality, evaluating monological discourse (Bolasco et al., 2006; Cedroni, 2010; Longobardi, 2010; Catellani et al., 2010; Bongelli et al., 2010; Zurloni and Anolli, 2010; Sprugnoli et al., 2016; Moretti et al., 2016) to study a politician's lexical, textual or rhetorical patterns. An exception is the work by Salvati and Pettorino (2010), that diachronically analyses some of the suprasegmental aspects of Berlusconi's speeches from 1994 to 2010. The corpus, however, is not available for further studies.

Concerning political corpora developed specifically for conversation analysis, Bigi et al. (2011) present a multimodal corpus of political debates at the French National Assembly, on May 4th, 2010 and introduce an annotation scheme for a political debate dataset which is mainly in the form of video and audio annotations. Navarretta and Paggio (2010) deal with the identification of interlocutors via speech and gestures in annotated televised political debates in British and American English. Other papers have focused primarily on visual aspects (gaze, gestures, facial expressions) of communicative interaction during political talk shows or parliamentary speeches (D'Errico et al., 2010).

The most similar approach to ours is presented in Koutsombogera and Papageorgiou (2010). The authors analyse a Greek multimodal corpus of 10 face-to-face television interviews focusing on non-verbal aspects in order to study the attempts of

persuasion and interruption during political interviews. Their work, however, is mainly aimed at studying the strategies for conversational dominance, and annotate specific traits accordingly. Our work, instead, is more general, includes a different set of tags and integrates also automatic linguistic features.

3 Description of the PoliModal corpus

The PoliModal corpus includes the transcripts of 56 TV face-to-face interviews of 14 hours - taken from the Italian political talk show "In mezz'ora in più" broadcast from 24 September 2017 to 14 January 2018. The show follows a fixed format, with interviews conducted by a journalist, Lucia Annunziata, to a guest, typically a prominent figure in the political or cultural scene. A secondary guest may participate as well, usually a second journalist to comment on the debate. Each interview is done in the same limited time frame, 30 minutes, and no audience is present, so that applause and any other type of reactions are not included in the corpus.

The audio signal has been transcribed using a semi-supervised speech-to-text methodology (Google API + manual correction). All hesitations, repetitions and interruptions of the original interview have been included. The output has been further segmented into turns, and punctuation has been added, mainly to delimit sentence boundaries when they were not ambiguous.

It is important to note that, even if transcription seems to be an objective task, it involves a certain degree of interpretation. Indeed, the inclusion of the punctuation necessary to make the writing comprehensible, as well as the selection of non-verbal messages and non-verbal expressions (interjections, laughter, unfinished words, etc.) are interpretative choices aimed at revealing a sense.⁵ Therefore, in the case of ambiguous sentences, they have been identified manually, mainly looking at the context of the enunciation. According to (Ducrot, 1995), in fact, it is not possible to understand a communicative act without knowing the context in which it occurs. The context is therefore essential to choose one of the possible interpretations of ambiguous expressions.

⁴LRE Map is a mechanism intended to monitor the use and creation of language resources by collecting information on both existing and newly-created resources, free available at <http://lremap.elra.info/>

⁵As (Portelli, 1985) reminds us: "La punteggiatura serve sia a scandire il ritmo che a gerarchizzare sintatticamente il discorso; non sempre le due funzioni coincidono, per cui trascrivendo si è costretti spesso optare per l'una a danno dell'altra"

In PoliModal, annotation has been done using XML as markup language and following the TEI standard for Speech Transcripts in terms of utterances. The linguistic resource has currently 100,870 tokens and includes interviews to politicians covering all the Italian political spectrum (from the extreme right movement Casa Pound to the liberal and progressive Partito Radicale). Beside politicians, also a small number of people with different backgrounds (students, academics, judges, economists, etc.) has been interviewed and is therefore included in the corpus.

For each interview the following information was manually annotated and is included in the XML resource file:

(a) **metadata**: these include useful information for a quick identification of transcriptions, for example the tools used for the transcription, a link to the interview, the owner account, the title of the talk show, the date of airing, the guests, etc.

(b) **pause**: this tag is used to mark a pause either between or within utterances. Speakers differ very much in their rhythm and in particular in the amount of time they leave between words, so the following element is provided to mark occasions where the transcriber judges that a speech has been paused, irrespective of the actual amount of silence. Several studies have converged on the conclusion that we alternate between planning speech and implementing our plans. Indeed, as shown in (Henderson et al., 1966), participants to interviews typically show a cycle of hesitation and fluency, although the ratio of speech to silence varies among speakers.

(c) **vocal**: with this tag we mark any vocalized but not necessarily lexical phenomenon, for example non-lexical expressions (i.e. burp, click, throat, etc.) and semi-lexical expressions (i.e. ah, aha, aw, eh, ehm etc.). These traits have been associated with the fact that linguistic planning is very cognitively demanding, and it is difficult to plan an entire utterance at once (Lindsley, 1975). Therefore, hesitation pauses and similar vocal phenomena may be useful to perform a careful lexical retrieval, since past studies (Levelt, 1983) found that pauses occurred more often before low-frequency words than before high frequency ones.

(d) **del**: this tag covers different phenomena of speech management, specifically false starts, repetitions and truncated words. Since they are marked in the TEI Guidelines as ‘editorially deleted’, the

corresponding tag is **del**. We include these in our annotation since several past studies (Simone, 1990; Bazzanella, 1992; Tannen, 1989) highlighted their importance in spontaneous speech, mentioning in particular the role of repetitions in controlling the in-progress textual design of speech (Voghera, 2001).

(e) **overlap**: this phenomenon is present when the speaker conveys (in a verbal or non-verbal manner) that he/she is about to finish his/her turn and the co-locutor starts speaking so that there is a slight overlap of utterances. Overlaps can be competitive, when the overlapper disrupts the speech and can be perceived as intrusive by dominating the conversation, and cooperative, when the goal of the overlapper is to maintain the flow of the turns and add to the conversation with further comments (Truong, 2013).

4 Corpus Analysis

In this section, we analyse several linguistic dimensions that can be either automatically extracted or derived from the corpus annotation, and that can contribute to better understand typical traits of political communication.

4.1 Statistics of Non-Verbal Traits

We first group the politicians in our corpus into political parties, and then analyse those that are represented by least 3 politicians: Forza Italia, a conservative center-right political party (3 interviews), Lega Nord, a right-wing political party often targeting immigrants (5 interviews), Movimento 5 Stelle, a populist citizens’ movement (3 interviews) and Partito Democratico, a moderate centre-left political party (9 interviews). An overview of the distribution of non-verbal traits in the PoliModal corpus for each party is reported in Fig. 1. Although the graph shows some differences in the frequency of occurrences, they are not statistically significant, also because of the relatively small number of interviews considered in the study. Also, the standard deviation for the averages tends to be high, showing high differences among interviewees of the same party. For example, politicians of Lega Nord make on average more pauses, but the range goes from 0.286 per turn (Roberto Maroni) to 0 (Luca Zaia). Similarly, non-lexical and semi-lexical expressions, marked as vocal, are on average more frequent for PD politicians, but range from 1.25 per turn (En-

rico Letta) to 0.10 (Matteo Renzi). These results show that differences pertain more to single persons and conversational style than to political orientation. An exception is given by overlaps, for which the three politicians of M5Stelle (Alessandro Di Battista, Luigi Di Maio, Giancarlo Cancelleri) all show a frequency above average, suggesting that it may be connected with the communication strategy of the members of Movimento.

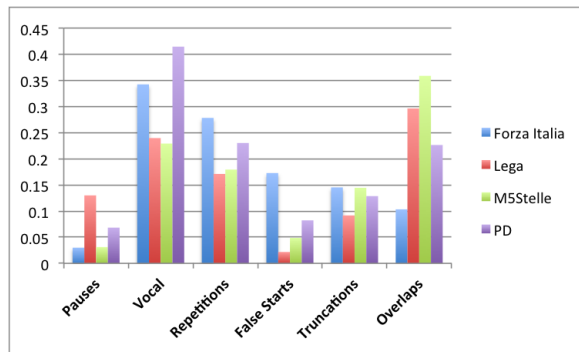


Figure 1: Distribution of traits per political party (avg. number of occurrences per turn).

4.2 Political orientation and Language Use

A second analysis we carry out is related to existing works about the use of linguistic features related to political orientation. In particular, a recent study by Schoonvelde et al. (2019) has analysed more than 380,000 speeches from five different Parliaments, and has proven that ideologically conservative politicians use a less complex language than liberal ones (this result is however less clear for economic left-right ideology). Since these findings were not tested on Italian political documents, we carry out a comparison using the collected transcripts. In order to analyse the complexity of the language used by each politician we computed the type-token ratio and the average lexical density, i.e. the number of content words divided by the total number of tokens. We do not take into account the Gulpease index (Lucisano and Piemontese, 1988), which is the de-facto standard metric of readability in Italian, because it was meant for written documents and heavily relies on sentence length, a boundary that is not always present in transcripts.

Fig.2 shows the average type-token ratio and conceptual density per political party. There are almost no variations among the parties, with small standard deviations. This comparison suggests

that in our case the hypothesis by Schoonvelde et al. (2019) is not confirmed, with the three highest ttr values belonging to politicians from three different parties: Forza Italia (Mariastella Gelmini, 0.87 ttr), Lega Nord (Matteo Salvini, 0.82) and PD (Michele Emiliano, 0.82).

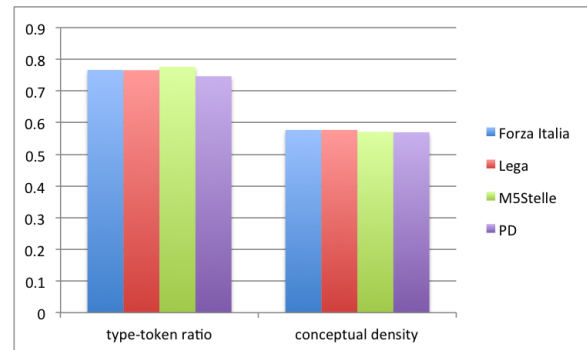


Figure 2: Avg. ttr and conceptual density per political party

A second hypothesis we want to test is the one introduced in the work by Cichocka et al. (2016), where the authors show that Republican presidents used a higher proportion of nouns than Democratic presidents, while there were no reliable differences in the use of verbs or adjectives. The authors suggest that, compared to liberals, conservative politicians are more inclined to use parts of speech that stress clarity and predictability (such as nouns) and reduce uncertainty and ambiguity (such as verbs or adjectives). We therefore compute the average number of nouns, adjectives and verbs per political party and compare them. Similar to the previous analysis, averages are all in the same range and there is no statistically significant difference among parties. However, some of the results are in line with Cichocka et al.'s study, with PD showing a slightly lower number of nouns on average (and Valeria Fedeli being the politician with the lowest noun ratio, 0.16). Also, Matteo Salvini and Luigi di Maio are the politicians with the highest use of nouns, 0.22 per token on average. A further evidence in favour of these results are the statistics obtained on the use of content words, in particular on the percentage of nouns, verbs, adverbs and adjectives, reported in Fig 3. We consider the five politicians with the highest number of turns in the corpus (see Table 1): Alessandro Di Battista (Movimento 5 Stelle), Carlo Calenda (PD), Matteo Renzi (PD), Angelino Alfano (Popolo della Libertà), Matteo

Salvini (Lega). The figure confirms that Matteo Salvini is the politician using the most nouns on average, in line with the findings by Cichocka et al. (2016). Carlo Calenda, instead, is the politician that on average uses most verbs and adverbs, conveying more uncertainty and ambiguity than all the other politicians including Matteo Renzi.

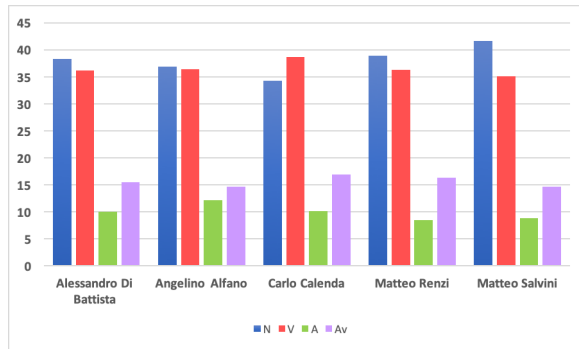


Figure 3: Use of nouns, verbs, adjectives and adverbs for each politician (% over all content words)

The fact that the two studies considered do not find a clear confirmation in our corpus, where the differences among the parties are rather blurred, may have three possible explanations: *i*) this corpus may be too small to test the above hypotheses. Its expansion is indeed already in progress; *ii*) the hypotheses do not actually hold in our case, i.e. in the Italian political scene it is not true that liberals use more complex language and tend to use less nouns than conservatives; or *iii*) the four parties considered cannot be straightforwardly divided into liberals and conservatives, and there are different positions inside the same party.

4.3 Relation between verbal and non-verbal traits

A third analysis is aimed at studying the correlation between non-verbal traits and language complexity. We therefore focus on the interviews that have a minimal length of 50 turns. The list of politicians and corresponding count of annotated traits is reported in Table 1. Again, for complexity we consider type-token ratio and conceptual density.

We perform an analysis of the correlation between language complexity and the six non-verbal traits manually annotated in the interviews, normalised by the number of turns uttered by each politician. While type-token ratio (TTR) does not correlate with any of the manual traits, we found

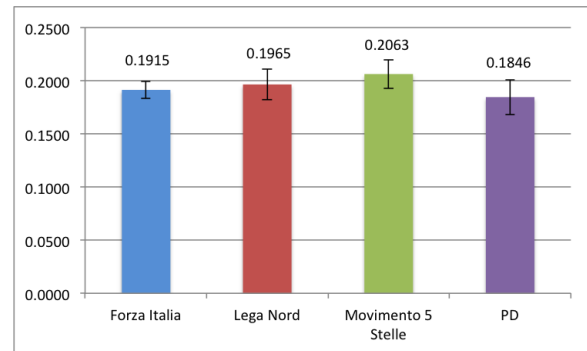


Figure 4: Avg. nouns per political party

that lexical density shows a moderate negative correlation with repetitions ($n=13$, $r=-0.51$), truncations ($r=-0.46$) and non-lexical and semi-lexical expressions ($r=-0.43$). On the contrary, it has a moderate positive correlation with the average number of pauses ($r=0.49$). This result suggests that, among the manual traits, pauses are used as a linguistic device and are an indicator of a good control of the conversation. Therefore, they are more often used by politicians showing a high lexical density, i.e. the ability to convey concepts in a concise way, which is crucial especially during TV interviews. The other manually annotated traits, instead, seem to be more frequent in speeches that are less organised, for which the management of the discourse is less efficient.

Among the politicians considered in this study, Carlo Calenda makes on average the highest number of pauses (0.27 per turn on average, with a lexical density of 0.579), followed by Giulio Tremonti (0.16 pauses per turn, 0.585 lexical density).

5 Conclusions

In this work, we present PoliModal, the first freely-available multimodal corpus of political interviews, manually annotated with six non-verbal traits. The corpus covers 56 interviews, where each guest is associated with a role (for non politicians) or a political party. We also present a first statistical analysis of the traits and their association with language complexity and with the speakers' political orientation.

In the future, we plan to start from the annotated material not only to extend the corpus, but also to investigate other aspects of political communication. For example, the choice to note non-verbal expressions is motivated by the will to study

Guest	Turn	Repetition	FalseStart	Truncation	Overlap	Pause	Non-lexical	Semi-lexical
Alessandro Di Battista	203	24	14	34	76	19	9	66
Carlo Calenda	137	10	13	1	48	37	1	34
Matteo Renzi	187	40	19	69	25	0	3	16
Walter Veltroni	55	16	12	10	11	0	2	8
Simone Di Stefano	91	20	5	15	23	0	0	4
Pierluigi Bersani	92	30	0	20	15	1	14	24
Angelino Alfano	100	17	3	3	31	9	2	22
Giulio Tremonti	56	8	0	0	14	9	2	6
Matteo Orfini	67	10	0	0	21	1	2	8
Luigi Di Maio	74	14	0	14	32	0	4	11
Matteo Salvini_1	57	13	0	11	19	3	2	14
Matteo Salvini_2	86	19	3	3	30	13	7	19
Pier Carlo Padoan	67	5	1	7	13	8	13	21

Table 1: Corpus statistics related to the 13 interviews included in our study

the strategies of persuasion used by the speakers. According to Poggi (2005), persuasion strategies are multimodal constructs because politicians – specifically in televised political interviews – attempt to persuade their supporters not only by their discursive style and argumentative speech, but also through their personality and their interactional behaviour. In the context of a political interview, persuasion is related to conversational dominance, i.e. a speaker’s tendency to control the other speaker’s conversational actions over the course of an interaction (Itakura, 2001), which is made evident through the kind of non-verbal expressions annotated in our corpus.

Finally, since at the moment only one annotator has performed the transcription, segmentation and tagging task, we plan to compute inter-annotator agreement in the near future. The annotation task addressed so far falls – from a qualitative point of view – in the first of the general types identified by (Mathet et al., 2015), in which the subjective interpretation is limited. Indeed, it deals with the “identification of units” (Krippendorff, 2018), in which the annotator, given a written or spoken text, must identify the position and boundary of linguistic elements (e.g. identification of prosodic or gestural units, topic segmentation). We therefore expect agreement to be at least fair, but we plan to measure it using standard metrics.

References

- Iolanda Alfano, Francesco Cutugno, Aurelio De Rosa, Claudio Iacobini, Renata Savy, and Miriam Voghera. 2014. Volip: a corpus of spoken italian and a virtuous example of reuse of linguistic resources. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC’14)*, Reykjavik, Iceland, may. European Language Resources Association (ELRA).
- Jens Allwood. 2008. Multimodal Corpora. In Lüdelling, A. & Kytö, and M., editors, *Corpus Linguistics. An International Handbook*, pages 207–225. Mouton de Gruyter.
- Roberto Bartolini, Valeria Quochi, Irene De Felice, Irene Russo, and Monica Monachini. 2014. From synsets to videos: Enriching italwordnet multimodally. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC’14)*, Reykjavik, Iceland, may. European Language Resources Association (ELRA).
- Carla Bazzanella. 1992. Aspetti pragmatici della ripetizione dialogica. In Giovanni Gobber, editor, *Linguistica pragmatica*, volume XXIV of *Atti SLI*, pages 433–454, Roma. Bulzoni.
- Brigitte Bigi, Cristel Portès, Agnès Steuckardt, and Marion Tellier. 2011. Multimodal annotations and categorization for political debates. In *ICMI Workshop on Multimodal Corpora for Machine learning*, pages 1–4.
- Sergio Bolasco, Nora Galli de’Paratesi, and Luca Giuliano. 2006. *Parole in libertà: un’analisi statistica e linguistica dei discorsi di Berlusconi*. Manifestolibri.
- Ramona Bongelli, Ilaria Riccioni, and Andrzej Zuczkowski. 2010. Certain-uncertain, true-false, good-evil in italian political speeches. In *International Workshop on Political Speech*, pages 164–180. Springer.
- Patrizia Catellani, Mauro Bertolotti, and Venusia Covelli. 2010. Counterfactual communication in politics: Features and effects on voters. In *International Workshop on Political Speech*, pages 75–85. Springer.

- Lorella Cedroni. 2010. Politolinguistics. Towards a New Analysis of Political Discourse. In *International Workshop on Political Speech*, pages 220–232. Springer.
- Aleksandra Cichocka, Michał Bilewicz, John T Jost, Natasza Marrouch, and Marta Witkowska. 2016. On the grammar of politics or why conservatives prefer nouns. *Political Psychology*, 37(6):799–815.
- Marco Dinarelli, Silvia Quarteroni, Sara Tonelli, Alessandro Moschitti, and Giuseppe Riccardi. 2009. Annotating spoken dialogs: From speech segments to dialog acts and frame semantics. In *Proceedings of SRS� 2009, the 2nd Workshop on Semantic Representation of Spoken Language*, pages 34–41, Athens, Greece, March. Association for Computational Linguistics.
- Oswald Ducrot. 1995. Les modificateurs déréalisants. *Journal of pragmatics*, 24(1-2):145–165.
- Francesca D’Errico, Isabella Poggi, and Laura Vincze. 2010. Discrediting body. a multimodal strategy to spoil the others image. In *International Workshop on Political Speech*, pages 181–206. Springer.
- Fabrizio Esposito, Pierpaolo Basile, Francesco Cutugno, and Marco Venuti. 2015. The Comp-WhoB Corpus: Computational construction, annotation and linguistic analysis of the white house press briefings corpus. *Proceedings of CLiC-it*.
- Marco Guerini, Danilo Giampiccolo, Giovanni Moretti, Rachele Sprugnoli, and Carlo Strapparava. 2010. The new release of Corps: A corpus of political speeches annotated with audience reactions. In *International Workshop on Political Speech*, pages 86–98. Springer.
- Alan Henderson, Frieda Goldman-Eisler, and Andrew Skarbek. 1966. Sequential temporal patterns in spontaneous speech. *Language and Speech*, 9(4):207–216.
- Hiroko Itakura. 2001. Describing conversational dominance. *Journal of Pragmatics*, 33(12):1859–1880.
- Maria Koutsombogera and Harris Papageorgiou. 2010. Multimodal indicators of persuasion in political interviews. In *International Workshop on Political Speech*, pages 16–29. Springer.
- Klaus Krippendorff. 2018. *Content analysis: An introduction to its methodology*. Sage publications.
- Willem JM Levelt. 1983. Monitoring and self-repair in speech. *Cognition*, 14(1):41–104.
- James R Lindsley. 1975. Producing simple utterances: How far ahead do we plan? *Cognitive Psychology*, 7(1):1–19.
- Ferdinando Longobardi. 2010. Linguistic factors in political speech. In *International Workshop on Political Speech*, pages 233–244. Springer.
- Pietro Lucisano and Maria Emanuela Piemontese. 1988. Gulpease. Una formula per la predizione della difficoltà dei testi in lingua italiana. *Scuola e Città*, 3:57–68.
- Yann Mathet, Antoine Widlöcher, and Jean-Philippe Métivier. 2015. The unified and holistic method gamma (γ) for inter-annotator agreement measure and alignment. *Computational Linguistics*, 41(3):437–479.
- Giovanni Moretti, Rachele Sprugnoli, Stefano Menini, and Sara Tonelli. 2016. ALCIDE: Extracting and visualising content from large document collections to support Humanities studies. *Knowledge-Based Systems*, 111:100–112.
- Costanza Navarretta and Patrizia Paggio. 2010. Multimodal behaviour and interlocutor identification in political debates. In *International Workshop on Political Speech*, pages 99–113. Springer.
- Isabella Poggi. 2005. The goals of persuasion. *Pragmatics & Cognition*, 13(2):297–335.
- Alessandro Portelli. 1985. *Biografia di una città: storia e racconto: Terni 1830-1985*. Einaudi.
- Luisa Salvati and Massimo Pettorino. 2010. A diachronic analysis of face-to-face discussions: Berlusconi, fifteen years later. In *International Workshop on Political Speech*, pages 65–74. Springer.
- Martijn Schoonvelde, Anna Brosius, Gijs Schumacher, and Bert N Bakker. 2019. Liberals lecture, conservatives communicate: Analyzing complexity and ideology in 381,609 political speeches. *PloS one*, 14(2):e0208450.
- Raffaele Simone. 1990. *Fondamenti di linguistica*. Laterza Bari.
- Rachele Sprugnoli, Giovanni Moretti, Sara Tonelli, and Stefano Menini. 2016. Fifty years of european history through the lens of computational linguistics: the de gasperi project. *IJCOL*, pages 89–100.
- Deborah Tannen. 1989. Interpreting interruption in conversation. *Gender and discourse*, pages 53–83.
- Khiet P. Truong. 2013. Classification of cooperative and competitive overlaps in speech using cues from the context, overlapper, and overlappee. In *Proceedings of INTERSPEECH*.
- Miriam Voghera. 2001. Teorie linguistiche e dati di parlato. In Sornicola R. Stromboli C. Albano Leoni F., Stenta Krosbakken E., editor, *Dati empirici e teorie linguistiche*, volume XXXIII of *Congresso Internazionale di Studi della Società di linguistica italiana*, pages 75–95., Roma. Bulzoni.
- Valentino Zurloni and Luigi Anolli. 2010. Fallacies as argumentative devices in political debates. In *International Workshop on Political Speech*, pages 245–257. Springer.

Analyses of Literary Texts by Using Statistical Inference Methods

Mehmet Can Yavuz

Computer Science and Engineering Department, Sabancı University, Tuzla
Management Information Systems Department, Kadir Has University, Cibali

Physics Department, Boğaziçi University, Bebek

İstanbul, Türkiye

mehmetyavuz@sabanciuniv.edu

Abstract

If a road map had to be drawn for Computational Criticism and subsequent Artificial Literature, it would have certainly considered Shakespearean plays. Demonstration of these structures through text analysis can be seen as both a naive effort and a scientific view of the characteristics of the texts. In this study, the textual analysis of Shakespeare plays was carried out for this purpose.

Methodologically, we consecutively use Latent Dirichlet Allocation (LDA) and Singular Value Decomposition (SVD) in order to extract topics and then reduce topic distribution over documents into two-dimensional space. The first question asks if there is a genre called Romance between Comedy and Tragedy plays. The second question is, if each character's speech is taken as a text, whether the dramatic relationship between them can be revealed.

Consequently, we find relationships between genres, also verified by literary theory and the main characters follow the antagonisms within the play as the length of speech increases. Although the results of the classification of the side characters in the plays are not always what one would have expected based on the reading of the plays, there are observations on dramatic fiction, which is also verified by literary theory. Tragedies and revenge dramas have different character groupings.

1 Introduction

If a road map had to be drawn for Computational

Criticism (Moretti, 2013) and subsequent Artificial Literature, it would have certainly considered Elizabethan drama. In particular, Shakespearean texts are the most outstanding examples of dramatic fiction. Demonstration of these structures through text analysis can be seen as both a naive effort and a scientific view of the characteristics of the texts. In this study, the textual analysis of Shakespeare plays was carried out for this purpose.

To begin with, "the First Folio" is the printed material in which all Shakespeare's works are brought together for the first time, (Synder, 2001). The edition of 1623 was directed by two actors from the group called King's Men. King's Men is the ensemble that Shakespeare is also a member of. Half of the 36-play collection had never been published anywhere before. The Folio was also printed in Quarto form. These prints took their names from the way the books were folded. It is known that the First Folio has 800 prints, 233 of them have reached today. In the First Folio, Shakespearean plays are typically divided into three groups: Comedies, Tragedies, and Histories. Romance is the genre that hybridizes Comedy and Tragedy, developed at the beginning of the 17th Century. At the end of his career, he wrote four romances: Pericles, Cymbeline, The Winter's Tale and The Tempest. "The First Folio" groups Cymbeline with Tragedies; and The Winter's Tale and The Tempest together with Comedies. The reason for this may be that The Winter's Tale and The Tempest began as tragedies and then turned to comedies, and Cymbeline started as a comedy and ended as a tragedy.

Shakespeare's two tragedies Macbeth and Othello are two very good examples of a true tragedy and a revenge tragedy. Tragedies are designed as the struggle of the main characters and the opposing characters who create obstacles for the main character. The protagonist is generally the main

Copyright © 2019 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

character that the audience sympathizes with. Although not sympathetic, Macbeth is a protagonist and the opposing characters are antagonists: Duncan and Banquo. Similarly, there is also antagonism in revenge drama and the main theme is revenge. The antagonist or protagonist seeks revenge for an imaginary or real injury. Iago the antagonist gets his revenge provoking Othello, the protagonist, against his wife.

Computerized analysis of literary texts, in other words computational criticism is a new and promising field, (Ramsay, 2011). Pioneering works aim to answer critical questions by using Natural Language Processing (NLP) methods. It is of interest to create fictional texts with the help of computer in the developing artificial literature along with these studies. In this study, we make a computational analysis of Shakespearean texts. There are basically two questions we're trying to answer. The first is if the genres in Shakespeare's theater texts can be classified by computer. Secondly, if the sentences in which the characters speak are taken as texts, can antagonisms be revealed? I tried to find answers to both with the same unsupervised learning technique.

In recent years, NLP methods have been developing rapidly and text analysis methods are getting more advanced. Topic Modeling articles are among the top cited articles. An unsupervised topic modelling algorithm is used in this study. It is able to generate latent topics in which each document is a mixture. Having the latent topic distribution, by using dimension reduction algorithm, each document is mapped onto two dimensional coordinates without losing intrinsic characteristics.

1.1 Related Works

Digital Humanities field lets researchers discuss quantitative methods in literary and cultural studies (Clement et al., 2008; Crane, 2006). "Drametrics" is a field that deals with quantitative analysis of the literary genre of drama (Romanska, 2015). Digital Shakespeare studies also have gotten attention since the 2000, (Hirsch, 2017; Mueller, 2008). The studies includes issues from digital archives to authorship analysis, (Vickers, 2011; Evert, 2017). Besides, machine learning based text analyses are also carried out for genre classifications, (Ardunuy, 2004; Hope, 2010; Schoch, 2016; Underwood, 2013; Yu, 2008). Informa-

tion theoretical approaches are also successfully applied, (Rosso, 2009). In literature, structural elements are quantified, such as the *dramatis personae* as well as scene structures; and applications are developed to further increase analysis (Dennerlein, 2015; Krautter, 2018; Schmidt, 2019; Trilcke, 2015; Wilhelm, 2013; Xanthos, 2016).

In order to analyze a literary text, we would like to use unsupervised topic modeling. Although there are linear-algebraic models such as Non-Negative Matrix Factorization (Lee, 1999), probabilistic models are more reliable and capable of representing true distributions of topics. Probabilistic Latent Semantic Analysis (Hoffman, 1999) and Latent Dirichlet Allocation (Blei, 2003) are the two major unsupervised topic modeling algorithms. Although both allow us to classify texts according to topic distribution, Latent Dirichlet Allocation as a generative model has a proven superiority over competitors. Principal Component Analysis (Jolliffe, 2002), Linear Discriminant Analysis (Brown, 2000) or Non-Negative Matrix Factorization (NMF) techniques are all dimension reduction algorithms, along side Singular Value Decomposition (Golub, 1970). The last algorithm we use is K-Means Clustering algorithm, a well known clustering algorithm that minimize variance within clusters (Lloyd, 1982).

2 Theory

In this study, we will use text analysis to investigate genres and antagonisms in Shakespearean plays. By using Latent Dirichlet Allocation (LDA), document distributions over topics are generated. Firstly, optimum number of topics will be obtained for LDA with grid search optimization and then dimension reduction algorithm, truncated Singular Value Decomposition (tSVD) will map these documents into a two-dimensional plane and graphed.

In the following sections, generating topics with LDA algorithm and dimension reduction by tSVD algorithm are explained. The aim of using tSVD algorithm is to express each text with two floating numbers while preserving the latent topic properties. Thus, classification can be made depending on the distances between each text in the new two-dimensional feature space. At the last step, we use a clustering with Euclidean distance. Theoretical section is kept brief and explanatory due fact that the main focus is on experimental results.

2.1 Latent Dirichlet Allocation (Blei, 2003)

LDA is a generative statistical model that explains why certain parts of the data are similar based on an observation set. LDA assumes that observations are generated by latent variables, or latent topics. Thus, each document is a mixture of topics and each topic is a distribution over words and each word is drawn from the mixture. The observations are frequency statistics of each document, so called the document-term matrix. The method is called the bag-of-words approach and intends to reflect how important a word is in a document. Thus, topics are identified on the basis of term co-occurrence, the topics-term matrix, and each document is assumed to be characterized by a particular set of topics, the document-topics matrix. Topics, mixtures and other variables are all hidden and need to be predicted from the observation data, the document-term matrix. In Figure 1, plate notation of LDA is represented. In the plate notation, there are $N \times D$ different variables that represent observations. There are K total topics and D total documents.

All at once, α and η are parameters of the prior distributions over θ and β respectively. θ_d the distribution of topics for document d (real vector of length K). β_k is the distribution of words for topic k (real vector of length V). $z_{d,n}$ is the topic for the n^{th} word in the d^{th} document. $w_{d,n}$ the n^{th} word of the d^{th} document. Only gray shaded circles are the observed variables. The rest of the white circles would be inferred by using Variation Inference. The topic for each word, the distribution over topics for each document, and the distribution of words per topic are all latent variables in this model. By this formulation, similarities can be introduced between documents.

The model contains both continuous and discrete variables. θ_d and β_k are vectors of probabilities. $z_{d,n}$ is an integer in $\{1, \dots, K\}$ that indicates the topic of the n^{th} word in the d^{th} document. $w_{d,n}$ is an integer in $\{1, \dots, V\}$ which indexes over all possible words.

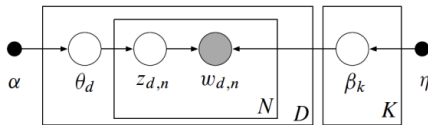


Figure 1: Plate notation representing the LDA model.

2.2 SVD (Golub, 1970)

If data has a large number of features, reduce it into a subset of features that are the most relevant to the prediction problem. SVD breaks any A matrix into a multiplication of three matrices so that,

$$A = USV' \text{ which} \quad (1)$$

$$UU' = I \text{ and } VV' = I \quad (2)$$

S is a diagonal matrix that consists of r singular values. r is the rank of A . Truncated SVD is a reduced rank approximation. All singular values are equated to zero except for the largest k , and largest singular values are the first k columns of U and V . The dimensions of truncated SVD are $[uxk] * [kxk] * [kxv]$. Since A matrix is approximated by k dimensions, there is a dimension reduction between matrix multiplications. A descriptive subset of the data is called T , which is a dense summary of the matrix A ,

$$T = US_k \quad (3)$$

S_k denotes k largest singular values, which is the number of reduced features. Each feature can be expressed by a percentage of variance, the reason behind this is choosing only the most significant ones.

2.3 K-means Clustering (Lloyd, 1982)

The K-Means clustering algorithm separates a group of equal variance samples from data by minimizing the sum-of-squares within clusters. The number of clusters needs to be pre-determined.

3 Experiments²

We included two evaluations in our experiments. The first is whether or not the genre of Romance can be distinguished computationally by computer. In order to carry out this experiment, each tragedy, comedy and romance is treated as a different document; and is processed by LDA. Afterwards, for the document-topic distribution matrix, the number of topics is reduced to two by means of dimension reduction algorithm, tSVD. Similarly, in the second evaluation, the lines of each character were treated as a text and the document-subject matrix was reduced to two after processing it with LDA. Two different type of tragedies are considered: Macbeth and Othello. Thus, three different

²In Python, Scikit-learn library used for LDA, tSVD and GridSearch functions.

experiments and optimization were conducted for these two evaluations.

3.1 Dataset and Preprocess

Two preprocesses were performed for each set of documents. Primary, stop-words were removed from the dictionary. These stop-words were created for both the usual English and Elisabethan English. The number of stop words is 1144. The characteristic of these words is that they often appear in every text. The secondary process is the expression of texts with word frequencies and the creation of the document-term matrix. Thus, each text could be expressed in a dictionary size fixed-length vector. Concatenations of these vectors creates the document-term matrix.

3.2 Optimization

In order to find the right topic number, we need an optimization. Since the subjects/topics are latent variables, there is no right number of topics. Grid-search optimization over topic numbers is carried out, and the highest log-likelihood is the optimal settings. In all three experiments, the values between 6 and 12 were tried three times and drawn in Figure 2. Thus for example, for Macbeth, 3 experiments were conducted for a certain topic number. The LDA function that we called for the experiment was repeated up to 10 times before giving results. Thus, for example, the LDA algorithm was repeated up to 30 times in total for a certain topic number.

As an observation, as the number of topics decreases, log-likelihood increases. However, we prefer not to try less than 6 latent topics because, in literature, the number of themes/topics for Shakespearean plays is generally at least 6, ("William Shakespeare", 2015).

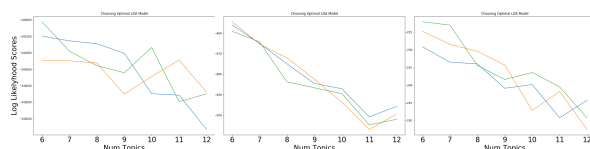


Figure 2: Optimization. Likelihood w.r.t. Topics Numbers. Tragedies-Comedies, Macbeth, Othello, respectively.

4 Discussion

4.1 Tragedy-Comedy-Romance

In Figure 3, documents consisting of Tragedy-Comedy-Romance plays are represented. The document-topic distribution matrix is reduced to two dimensions, and graphed. More than half of variances is explained by these two components. Even in three dimensions, the clustering does not change. The plays that are shown in red are Comedies, the blues are Tragedies and the greens are Romances according to the First Folio.

In the upper left corner, the majority of the Comedies are clustered, and likewise in the lower right corner Tragedies are clustered. In the middle of these two clusters, three plays, "All's Well That Ends Well", "Measure for Measure" and "Troilus and Cressida" are placed known as problem plays. Some critics also includes "Timon of Athens" which is a neighbor of other problem plays, (Snyder, 2001). Thus, in the middle of the two clusters, there is a gray zone in which problem plays are placed. An interesting fact is, although "All's Well That Ends Well" and "Measure for Measure" are grouped as Comedies in the First Folio, they are much closer to tragedies. An unexplained fact is that Coriolanus and Othello are also placed in this gray zone. Another question in this grouping is "Romeo and Juliet". As a tragedy that has comedy elements is placed thematically very close to the Comedies cluster.

Another important distinction is that these three Romances are clustered within the Tragedies. According to this analysis, the genre of Romance is not different from tragedy.



Figure 3: Genre classification of Tragedies, Comedies and Romance

4.2 Macbeth

After the analysis, the characters of Macbeth clearly demonstrate Antagonist/Protagonist relations as graphed in Figure 4. There are two basic clusters in the tragedy of Macbeth. The first is the protagonists, led by Macbeth and Lady Macbeth. The second is the antagonists, who are the murdered king and Macduff who suspects foul play. In the plot, protagonists are shown in blue and antagonists in red. Lady Macbeth stands at the bottom left corner, since Lady Macbeth doesn't have much to talk except to Macbeth. Macbeth's himself is closer to the red cluster. He has relations with red clusters as a new King. Macduff, who is suspicious and kills Macbeth in the last scene, is in the center of the red cluster. Lady Macduff is also in this cluster. The murdered King Duncan is also at the center of this cluster. However, there is also a misclassification. Siward is in the blue cluster. However, Siward and Macbeth have a clash in which Siward is killed. Other than that, the witches who oracles, are in the opposite cluster of Macbeth. Other characters may not be fully explained due to their small and ambiguous roles. Apart from these two clusters, there is a top left green cluster. The main character of this cluster is Banquo. This character is Macbeth's brave and noble companion. But he had no idea about Macbeth's machinations until he is killed. Tragedy of Macbeth has a very clear separation between clusters. The distance between clusters is also meaningful. The reds are between green and blues. The greens are actually closer to reds rather than Macbeth's evil cluster.

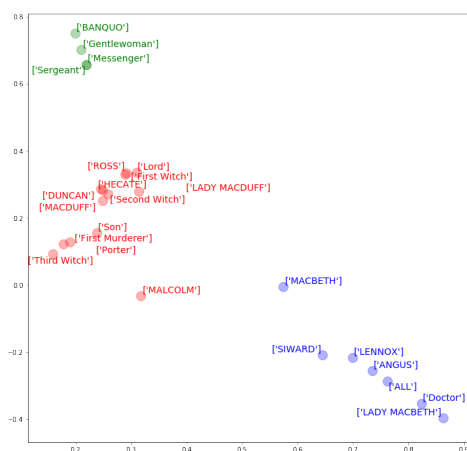


Figure 4: Characters of the play Macbeth are represented.

4.3 Othello

The characters of the Othello play are shown in the Figure 5 in accordance with the analysis. I give Othello as an example of revenge tragedies. Unlike a true tragedy, Macbeth, the Othello play does not have antagonist/protagonist clusters in the Figure 5. Iago is a single character who sets traps to get revenge on Othello. Throughout the play, Iago misleads Othello for reasons and purposes that only he and the reader know. Othello kills his beloved wife in a crisis of jealousy.

There are three different colored clusters shown. The red set consists of the main people of the play. Blue and green clusters belong to side characters and antagonisms are computationally ambiguous. The main characters of the red cluster at the bottom right, Othello, Emilia, Iago and Cassio have spoken almost the same subject because of the frequency of their dialogue with each other. Therefore, a conflict between them is not visible. But Iago is shown in the lower right corner because he shows his true intention in his monologues. Therefore, Othello is a negative example for the methodology we developed. Characters such as the Duke of Venice and the Senator are mentioned in the top left corner and are in fact extremely outside the plot. Shown from the green cluster, Bianca is again outside the plot as Cassio's lover.

In Othello, there are interesting observations on revenge tragedies. In revenge tragedies of Shakespeare, a lonely character shows him/herself differently and his/her true intentions remain hidden. Thus, the clear difference from tragedies, is their dramatic structure.

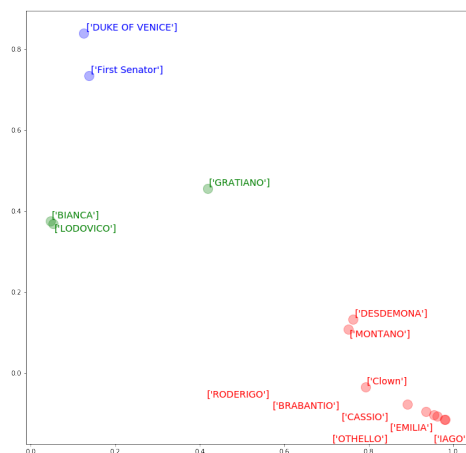


Figure 5: Characters of the play Othello are represented.

5 Conclusion

The classification of genres shows us that the method we use provides successful quantitative information for the differentiation of genres. The length of the texts can be mentioned among the reasons for this success. Positioning the plays between Tragedy and Comedy is much discussed in the literature theory. The Romance genre hybridizes Tragedy and Comedy elements. Instead of mapping the Romance genre in between, the algorithm mapped four "Problem Plays" in a region between Tragedies and Comedies. Another interesting finding is that Romance cannot be distinguished from Tragedies. The method used shows that the reason for some literary discussion is at the same time quantitative. The method classifies Romances within the Tragedies. In the light of theoretical discussions, of course, there may be a genre called Romance, but we have not been able to quantify this difference yet.

There are also some results from our experiments on the two tragedies we have chosen. I intentionally choose a tragedy and a revenge play, although Macbeth clearly shows antagonisms. This is mainly due to the frequency of conversations within these clusters. For example, Macbeth and Lady Macbeth are always aware of each others true intentions. Dialogues within these clusters are always compatible with each other. Therefore, the cluster forms. There is a group subjectivity, also verified computationally. The war scene at the end of Macbeth can clearly be observable from the clusters. Two clusters to clash are formed through out the play, which is quantifiable. On contrary, Iago who hides his true intention from everyone, has apparently always agreed with Othello. On the contrary, Iago never shares his intentions with anyone in the play. His intentions are shared through monologues. Thus, he could not form a cluster. He is a lonely character. That is why, algorithm fails to find an antagonisms. From this point of view, we can say that the method forms clusters of characters that agree with each other. The dramatic structure of revenge plays cannot be revealed by the method we proposed. Our method is successful when finding the clusters. We carried out a similar analysis for the play Hamlet, another type of revenge plays. Hamlet distinguished himself in a different cluster, as a lonely character with Lord Polonius who is responsible for spying on Hamlet. Lord Polonius is a similar character with Iago in

terms of hiding their true intentions.

The dramatic fiction in Shakespeare's texts is shown to a certain extent. The advantage of the proposed pipeline is using non-linearity over a linear layer. Instead of directly clustering the document-term matrix, a powerful representation of each document in a feature space is generated by LDA. After generating document-topic matrix, a linear layer of dimension reduction, tSVD, that extracts principal directions or principal axes in which the document-topic matrix have the largest variance.

I think that these naive efforts on the way to Artificial Literature also have a positive effect. The production of a play is possible with the knowledge of authorship for humans and even for Shakespeare. By authoring knowledge, we mean, for example, how to write a play from dramatic perspective. It is firstly introduced by Aristotle to shed light on present-day methods. It would be possible to reverse engineering them for artificial literature. Going from a quantitative analysis to plays would be possible. Therefore, as we analyze literary pieces, especially texts in dialogue form can help us verify critical questions and theories. From these analyses, going back to the literary text generation becomes possible.

Acknowledgments

This work was supported by grant 12B03P4 of Boğaziçi University.

The author would like to thank Muhittin Mungan for suggesting Master of Science thesis as his advisor and Meltem Gürle Mungan for her kind opinion. The author would also like to thank actor Güneş Yakın, for talks together on Shakespeare.

References

- Ardanuy, M. C., & Sporleder, C. (2014, April). Structure-based clustering of novels. In *Proceedings of the 3rd Workshop on Computational Linguistics for Literature (CLFL)* (pp. 31-39).
- Blei, D. M., Ng, A. Y. & Jordan, M. I. (2003). Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3, 993–1022. doi: <http://dx.doi.org/10.1162/jmlr.2003.3.4-5.993>
- Brown, M. T., & Wicker, L. R. (2000). Discriminant analysis. In H. E. A. Tinsley & S. D. Brown (Eds.), *Handbook of applied multivariate statistics and mathematical modeling* (pp. 209-235). San Diego, CA, US: Academic

- Press. <http://dx.doi.org/10.1016/B978-012691360-6/50009-4>
- Clement, T., Steger, S., Unsworth, J. and Uszkalo, K. (2008). *How not to read a million books*. Available online at <http://people.brandeis.edu/~unsworth/hownot2read.html>
- Crane, G. (2006). *What do you do with a million books?* D-Lib Magazine. Available online at <http://www.dlib.org/dlib/march06/crane/03crane.html>
- Dennerlein, K. (2015). Measuring the average population densities of plays. A case study of Andreas Gryphius, Christian Weise and Gotthold Ephraim Lessing. *Semicerchio. Rivista di poesia comparata* LIII: 80–88.
- Evert, Thomas & Proisl, & Jannidis, Fotis & Reger, Isabella & Pielström, Steffen & Schöch, Christof & Vitt, Thorsten. (2017). Understanding and explaining Delta measures for authorship attribution. *Digital Scholarship in the Humanities*. 32. 4-16. 10.1093/lc/fqx023.
- Hofmann, T. (1999). Probabilistic latent semantic analysis. *Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence* (p./pp. 289–296), .
- Hope, J., & Witmore, M. (2010). The Hundredth Psalm to the Tune of "Green Sleeves": Digital Approaches to Shakespeare's Language of Genre. *Shakespeare Quarterly*, 61(3), 357-390. Retrieved from <http://www.jstor.org/stable/40985589>
- Hirsch, B., & Craig, H. (2014). "Mingled Yarn": The State of Computing in Shakespeare 2.0. In T. Bishop, & A. Huang (Eds.), *The Shakespearean International Yearbook* (Vol. 14: Special Section, Digital Shakespeares, pp. 3-35). United Kingdom: Ashgate Publishing Limited.
- Golub, G. H.; Reinsch, C. (1970). "Singular value decomposition and least squares solutions". *Numerische Mathematik*. 14 (5): 403–420. doi:10.1007/BF02163027. MR 1553974.
- Jolliffe, I. (2002). *Principal component analysis*. New York: Springer Verlag.
- Krautter, B. (2018). Quantitative microanalysis? Different methods of digital drama analysis in comparison. *Book of Abstracts, DH 2018*. Mexico-City, Mexico, pp. 225-228.
- Lee, Daniel Seung, H.. (1999). Learning the Parts of Objects by Non-Negative Matrix Factorization. *Nature*. 401. 788-91. 10.1038/44565.
- Lloyd, S.P. (1982). Least squares quantization in PCM. *IEEE Trans. Information Theory*, 28, 129-136.
- Mueller, Martin. (2008). Digital Shakespeare, or towards a literary informatics. *Shakespeare*. 4. 284-301. 10.1080/17450910802295179.
- Moretti, F. (2013). *Distant reading*. Verso Books.
- Rosso, Osvaldo & Craig, Hugh Moscato, Pablo. (2009). Shakespeare and other English Renaissance authors as characterized by Information Theory complexity quantifiers. *Physica A: Statistical Mechanics and its Applications*. 388. 916-926. 10.1016/j.physa.2008.11.018.
- Ramsay, S. (2011). *Reading Machines: Toward an Algorithmic Criticism*. University of Illinois Press.
- Romanska, M. (2015). Drametrics: what dramaturgs should learn from mathematicians. In Romanska, M. (ed.), *The Routledge Companion to Dramaturgy*. Routledge, pp. 472-481.
- Schöch, Christof. (2016). Topic Modeling Genre: An Exploration of French Classical and Enlightenment Drama. *Digital Humanities Quarterly*. <http://doi.org/10.5281/zenodo.166356>
- Schmidt, T., Burghardt, M., Dennerlein, K. & Wolff, C. (2019). Katharsis – A Tool for Computational Drametrics. In *Book of Abstracts, DH 2019*.
- Snyder, S. (2001). The genres of Shakespeare's plays. In M. De Grazia S. Wells (Eds.), *The Cambridge Companion to Shakespeare* (Cambridge Companions to Literature, pp. 83-98). Cambridge: Cambridge University Press. doi:10.1017/CCOL0521650941.006
- Trilcke, P., Fischer, F. and Kampkaspar, D. (2015). Digital Network Analysis of Dramatic Texts. *Book of Abstracts, DH 2015*. Sidney, Australia
- Underwood, T., Black, M.L., Auvil, L., & Capitanu, B. (2013). Mapping mutable genres in structurally complex volumes. 2013 *IEEE International Conference on Big Data*, 95-103.
- Vickers, Brian. (2011). Shakespeare and Authorship Studies in the Twenty-First Century. *Shakespeare Quarterly*. 62. 106-142. 10.1353/shq.2011.0004.
- William Shakespeare. (2015, August 21). *New World Encyclopedia*, . Retrieved 12:11, September 16, 2019 from http://www.newworldencyclopedia.org/p/index.php?title=William_Shakespeare&oldid=990237.
- Wilhelm, T., Burghardt, M., and Wolff, C. (2013). "To See or Not to See" - An Interactive Tool for the Visualization and Analysis of Shakespeare Plays. In R. Franken-Wendelstorf, E. Lindinger, and J. Sieck (Eds.), *Kultur und Informatik: Visual Worlds & Interactive Spaces*. Glückstadt: Verlag Werner Hülsbusch, pp. 175–185.
- Yu, B. (2008). An evaluation of text classification methods for literary study. *Literary and Linguistic Computing* 23(3): 327-343.
- Xanthos, A., Pante, I., Rochat, Y and Grandjean, M. (2016). Visualising the dynamics of character networks. *Book of Abstracts, DH 2016*. Kraków, Poland, pp. 417-419.

Neural Semantic Role Labeling using Verb Sense Disambiguation

Domenico Alfano

Eustema S.p.A.

d.alfano@eustema.it

Roberto Abbruzzese

Eustema S.p.A.

r.abbruzzese@eustema.it

Donato Cappetta

Eustema S.p.A.

d.cappetta@eustema.it

Abstract

The Natural Language Processing (NLP) community has recently experienced a growing interest in Semantic Role Labeling (SRL). The increased availability of annotated resources enables the development of statistical approaches specifically for SRL. This holds potential impact in NLP applications. We examine and reproduce the Marcheggiani's system and its individual components, including its annotated resources, parser, classification system, the features used and the results obtained by the system.

Then, we explore different solutions in order to achieve better results by approaching to Verb-Sense Disambiguation (VSD). VSD is a sub-problem of the Word Sense Disambiguation (WSD) problem, that tries to identify in which sense a polysemic word is used in a given sentence. Thus a sense inventory for each word (or lemma) must be used.

Finally, we also assess the challenges in SRL and identify the opportunities for useful further research in future.

1 Introduction

One of the fields where AI is gaining great importance is the NLP. Nowadays, NLP has many applications: search engines (semantic/topic search rather than word matching), automated speech translation, automatic summarization, etc.

Therefore, there are many sub-tasks for natural language applications that have already been studied. An example is the syntactic analysis of the words of a sentence. The object of this research

study is the realization of a system able to perform SRL.

A SRL system does nothing more than take a set of input phrases and, for each of them, it starts to determine the various components that could play a semantic role. A component of a proposition that plays a semantic role is defined as constituent. Once the possible candidates are determined, Machine Learning techniques are used to label them with the right role.

This task becomes important for advanced applications where it is also necessary to process the semantic meaning of a sentence. Moreover, all this applications have to deal with ambiguity.

Ambiguity is the term used to describe the fact that a certain expression can be interpreted in more than one way.

In NLP, ambiguity is present at several stages in the processing of a text or a sentence, such as: tokenization, sentence-splitting, part-of-speech (POS) tagging, syntactic parsing and semantic processing. Semantic ambiguity is usually the last to be addressed by NLP systems, and it tends to be one of the hardest to solve among all types of ambiguities mentioned.

For this type of ambiguity, the sentence has already been parsed and, even if its syntactic analysis (parse tree) is unique and correct, some words may feature more than one meaning for the grammatical category they were tagged with.

Usually this difference in meaning is associated to syntactic properties. In order to overcome these issues, this research study approaches to the VSD task. The majority of the systems used in the VSD task are based on Machine Learning techniques (Witten, 2011).

We approach both the tasks by following two different solutions.

2 Related Work

2.1 SRL Approaches

Until recently, state-of-the-art Semantic Role Labeling (SRL) systems relied on complex sets of lexico-syntactic features (Pradhan, 2005) as well as declarative constraints (Punyakanok, 2008). Neural SRL models, instead, exploit induction capabilities of neural networks, largely eliminating the need for complex "hand-made" features. Recently, it has been shown that an accurate span-based SRL model can be constructed without relying on syntactic features (Jie Zhou, 2015). In particular, Roth and Lapata (Roth and Lapata, 2016) argue that syntactic features are necessary for the dependency-based SRL and show that performance of their model degrades dramatically if syntactic paths between arguments and predicates are not provided as an input.

Recent studies (Luheng He, 2018) propose an end-to-end approach for jointly predicting all predicates, arguments spans, and the relations between them. The model makes independent decisions about what relationship, if any, holds between every possible word-span pair, and learns contextualized span representations that provide rich, shared input features for each decision.

2.2 WSD Approaches

An overview of the most used techniques and features for WSD was also conducted, based on the systems evaluated at the SensEval3. The most common learning algorithms (Witten, 2011) used at SensEval3 are the following:

- The Naive Bayes algorithm, which estimates the most probable sense for a given word w based on the prior probability of each sense and the conditional probability for each of the features in that context.
- The Decision List algorithm (Yarowsky, 1995), which builds a list of rules, ordered from the highest to the lowest weighted feature. The correct sense of the word is determined by the first rule that is matched.
- The Vector Space Model algorithm, which considers the features of the context as binary values in a vector. In the training phase, a centroid is calculated for each possible sense of the word. These centroids are then com-

pared with vectors of features from testing examples using the cosine function.

- Support Vector Machines, the most widely used classification technique in WSD at SensEval3 (Agirre, 2004); (Lee, 2004); (Villarejo, 2004), is a classification method that finds the maximal margin hyperplane that best separates the positive from the negative examples. In the particular case of WSD, this has to be slightly tuned for multiple class classification. Usually, methods like one-against-all are used, which lead to the creation of one classifier per class.

The most commonly used features used by the systems proposed and presented at SensEval3 can be divided as follows:

- Collocations: n -grams (usually bi-grams or tri-grams) around the target word are collected. The information stored for then-grams is composed by the lemma, word-from and part-of-speech tag of each word.
- Syntactic dependencies: syntactic dependencies are extracted among words around the target word. The relations most commonly used are subject, object, modifier. However, depending on the system, other dependencies might also be extracted.
- Surrounding context: single words in a defined window size are extracted and used in a bag-of-words approach.
- Knowledge-Based information: Some systems also make use of information such as WordNet's domains, FrameNet's syntactic patterns or annotated examples, among others.

3 Data

The dataset used is the CoNLL 2009 Shared Task built on the CoNLL 2008 task which has been extended to multiple languages. The core of the task was to predict syntactic and semantic dependencies and their labeling.

Data was provided for both statistical training and evaluation, in order to extract these labelled dependencies from manually annotated Treebanks such as the Penn Treebank for English, the Prague Dependency Treebank for Czech and similar Treebanks for Catalan, Chinese, German, Japanese and

Spanish languages, enriched with semantic relations. Great effort has been dedicated in providing the participants with a common and relatively simple data representation for all the languages, similar to the 2008 English data. Role-annotated data makes it available for many research opportunities in SRL including a broad spectrum of probabilistic and machine learning approaches.

We have introduced the dataset associated with SRL; we are now prepared to discuss the approaches to automatic SRL and VBS.

4 Metrics

For many of these subtasks there are standard evaluations techniques and corpora. Standard evaluation metrics from information retrieval include precision, recall and a combined metric called F_1 measure (Jurafsky, 2000).

Precision is a measure of how much of the information that the system returned is correct, also known as accuracy. Recall is a measure of how much relevant information the system has extracted from text, thus a measure of the system's coverage. The F_1 measure balances recall and precision.

A corpus is often divided into three sets: training set, development set and testing set. Training set is used for training systems, whereas the development set is used to tune parameters of the learning systems, and selecting the best model. Testing set is used for evaluation. Cross-corpora evaluation is used in some tasks, for which a fresh test set different from the training corpora is used for evaluation.

In this case, F_1 measure is computed as the harmonic mean of Precision and Recall.

5 Semantic Role Labeling

The model architecture for SRL is inspired from the one ideated by Marcheggiani et al., 2017 (Marcheggiani, 2017) based on the following three components.

Then, a table with all the hyperparameter values will be shown.

5.1 Word Representation

The word representation component builds from a word w_i in a sentence w a word representation x_i . Each word w is represented as the concatenation of four vectors:

- A randomly initialized word embedding $x_{re} \in R^{d_w}$.
- A pre-trained word embedding $x_{pe} \in R^{d_w}$.
- A randomly initialized part-of-speech tag embedding $x_{pos} \in R^{d_p}$.
- A randomly initialized lemma embedding $x_{le} \in R^{d_l}$ that is only active if the word is one of the predicates.

Then, it has been used the Predicate-Specific Encoding. Specifically, when identifying arguments of a given predicate, the authors added a predicate-specific feature to the representation of each word in the sentence by concatenating a binary flag to the word representation. The flag is set as 1 for the word corresponding to the currently considered predicate, it is set as 0 otherwise. In this way, sentences with more than one predicate will be re-encoded by Bidirectional LSTMs multiple times.

5.2 Encoder

Recurrent neural networks (RNN) (Elman, 1990), more precisely, Long Short-Term Memory (LSTM) networks (Hochreiter and Schmidhuber, 1997) are one of the most effective ways to model sequences. Formally, the LSTM is a function that takes as input the sequence and returns a hidden state. This state can be regarded as a representation of the sentence from the start to the position i , or, in other words, it encodes the word at position i along with its left context.

Bidirectional LSTMs make use of two LSTMs: one for the forward pass, and another for the backward pass. In this way the concatenation of forward and backward LSTM states encodes both left and right contexts of a word.

In this case, the Bidirectional Long-Short Term Memory (BiLSTM) Encoder takes as input the word representation x_i and provides a dynamic representation of the word and its context in a sentence.

5.3 Role Classifier

The goal of the classifier is to predict and label arguments for a given predicate.

The basic role classifier takes the hidden state of the top-layer bidirectional LSTM corresponding to the considered word at position i and uses it to estimate the probability of the role r .

However, since the context of a predicate in the

sentence is highly informative for deciding if a word is its argument and for choosing its semantic role, the authors provides the predicate’s hidden state as another input to the classifier.

Finally, it has been proven advantageous to jointly embed the role r and predicate lemma l using a non-linear transformation: *ReLU* (Vinod Nair and Geoffrey Hinton, 2010) that is the rectilinear activation function. In this way each role prediction is predicate-specific, and at the same time it has expected to learn a good representation for roles associated to in frequent predicates.

5.4 Hyperparameters

In the following table the hyperparameter values.

Hyperparameter	Value
English word embeddings	100
POS embeddings	16
Lemma embeddings	100
LSTM hidden states	512
Role representation	128
Output lemma representation	128
BiLSTM depth	4
Learning rate	.001

Table 1: Hyperparameter values.

6 Verb-Sense Disambiguation

In order to improve the results obtained from the Marcheggiani’s SRL model, two solutions will be presented:

- Multi-Task Learning: by sharing representations between related tasks (VBS), we can enable our model to generalize better on our primary task (SRL).
- Babelfy: usage of a pre-trained model that helps to disambiguate sentences and verbs.

In the first solution the two models run in parallel. In the second solution, since the SRL model uses as input the Babelfy’s output, the two models run sequentially.

6.1 Multi-Task Learning Solution

In Machine Learning we typically care about optimizing a particular metric. In order to do this, we generally train a single model to perform our desired task, then fine-tune and tweak this model until its performance no longer increases.

Even if it is possible to achieve generally acceptable performance, in this way we could miss information that might help us to optimize the relevant metric. Specifically, information deriving from the training signals of related tasks.

We can consider multi-task learning as a form of inductive transfer. Inductive transfer can help to improve a model by introducing an inductive bias, defining a model as preferable with respect to other hypotheses.

Furthermore, the Verb Sense Disambiguation model has been created; following, a brief explanation of the model. We use the same Word Representation and Encoder of the Marcheggiani’s system explained in sections 5.1, 5.2.

The output of the Encoder is used to predict the sense of the verb by applying the Softmax activation function.

Model	P	R	F_1
Lei (2015)	-	-	86.6%
FitzGerald (2015)	-	-	86.7%
Roth and Lapata (2016)	88.1%	85.3%	86.7%
Marcheggiani (2017)	88.7%	86.8%	87.7%
SRL+VSD Model	88.65%	86.62%	87.6%

Table 2: Multi-Task Learning Results.

As Table 2 shows, performance worsens in terms of Precision and Recall.

Therefore, we have a lower value in term of F_1 score, which, as already mentioned above, is the harmonic mean of Precision and Recall.

For this reason another solution was developed in order to improve the results on both Precision and Recall and then of F_1 .

6.2 Babelfy Solution

Babelfy (Navigli, 2014) is both a multilingual encyclopedic dictionary and a semantic network which connects concepts and named entities in a very large network of semantic relations called Babel synsets. Each Babel synset represents a given meaning and contains all the synonyms which express that meaning.

Specifically, Babelfy performs the tasks of multilingual Word Sense Disambiguation and Entity Linking.

Extracted senses have been used as input of the SRL Model, by replacing the randomly initialized

lemma embedding $x_{le} \in R^{d_l}$ of the word representation of 5.1

Model	P	R	F_1
Lei (2015)	-	-	86.6%
FitzGerald (2015)	-	-	86.7%
Roth and Lapata (2016)	88.1%	85.3%	86.7%
Marcheggiani (2017)	88.7%	86.8%	87.7%
SRL + Babelfy	88.96%	86.87%	87.9%

Table 3: Babelfy Results.

In this case we can observe improvements in all fields. This improvement is not so significant (Reimers and Gurevich, 2017) because LSTM-based models tend to be significantly sensible to initialization, for this reason 0.2% improvement in a small dataset like CoNLL2009 may not be a satisfactory increase.

Moreover, this results shows that improving the VSD task determines improvements in SRL task.

7 Conclusions

The realized work represents the development of a complete system for the Semantic Role Labeling, an important tool to be used in advanced Natural Language Processing applications.

A system of SRL alone is not very useful and it necessarily must be included in a wider application, for example a Question&Answering system or a Neural Machine Translation system.

In conclusion, as all the new applications of natural language processing must be able to handle semantic information if they want to have good performances, this type of system can be considered a valuable solution to achieve such performances. The statistical analysis of the errors registered by the system, developing from this analysis new algorithms in order to correct such errors, is another aspect to be considered in the evaluation of this system.

8 Future Works

As for future works we could certainly try to develop a new Semantic Role Labeling model architecture trying to discover approaches related to models based on Attention.

Attention (Bahdanau, 2015) is one of the main innovations for machine translation based on neu-

ral networks, the key idea that allowed neural networks to overcome classics translation models.

The main obstacle for the sequence-to-sequence learning is the need to compress all the information contained in the original sequence into a prefixed vector. Attention alleviates this problem (Luong, 2015), allowing the decoder to look again at the list of hidden states corresponding to the original sequence, whose weighted average is used as input from the decoder in addition to the compressed vector representation.

An interesting effect of attention (Vaswani, 2017) is the possibility to observe, superficially, the operating mechanisms inside the model: the attention makes visible which parts of the input have proved important for a certain output, thanks to the weights applied to get the average of the incoming sequence.

Another future research activity could be the examination of the abovementioned models under different languages, such as Italian.

References

- Witten, I. H., E. Frank, and M. A. Hall. 2011. *Data Mining: Practical Machine Learning Tools and Techniques* (3 ed.).
- Sameer Pradhan, Kadri Hacioglu, Wayne H. Ward, James H. Martin, and Daniel Jurafsky. 2005. *Semantic role chunking combining complementary syntactic views*. In Proceedings of CoNLL.
- Vasin Punyakanok, Dan Roth, and Wen-tau Yih. 2008. *The importance of syntactic parsing and inference in semantic role labeling*. Computational Linguistics 34(2):257–287.
- Jie Zhou and Wei Xu. 2015. *End-to-end learning of semantic role labeling using recurrent neural networks*. In Proceedings of ACL.
- Michael Roth and Mirella Lapata. 2016. *Neural semantic role labeling with dependency path embeddings*. In Proceedings of ACL.
- Luheng He, Kenton Lee, Omer Levy, Luke Zettlemoyer. 2018. *Jointly Predicting Predicates and Arguments in Neural Semantic Role Labeling*. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics.
- Yarowsky D. 1995. *Unsupervised word sense disambiguation rivaling supervised methods*. In Proceedings of the 33rd annual meeting on Association for Computational Linguistics. 189–196.
- Agirre E., Aldabe I., Lersundi M., Martinez D., Pociello E., Uria L. 2004. *The Basque Lexical-Sample Task*. Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text, Barcelona, Spain, Association for Computational Linguistics (2004) 1–413.
- Lee, Y.K., Ng, H.T., Chia, T.K. 2004. *Supervised Word Sense Disambiguation with Support Vector Machines and Multiple Knowledge Sources*. Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text, Barcelona, Spain, Association for Computational Linguistics (2004) 137–14014.
- Villarejo L., Marquez L., Agirre E., Martinez D., Magnini B., Strapparava C., McCarthy D., Montoyo A., Suarez A. 2004. *The "Meaning" System on the English All-Words Task*. Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text, Barcelona, Spain, Association for Computational Linguistics (2004) 253–256
- Jurafsky D. and Martin J. H. 2000. *Machine Translation*. In Speech and Language Processing. Prentice Hall.
- Diego Marcheggiani, Anton Frolov and Ivan Titov. 2017. *A simple and accurate syntax-agnostic neural model for dependency-based semantic role labeling*.
- Jeffrey L. Elman. 1990. *Finding structure in time*. Cognitive Science 14(2):179–211.
- Sepp Hochreiter and Jurgen Schmidhuber. 1997. *Long short-term memory*. Neural Computation 9(8):1735–1780.
- Vinod Nair and Geoffrey Hinton. 2010. *Rectified Linear Units Improve Restricted Boltzmann Machines*. ICML.
- A. Moro, A. Raganato, R. Navigli. 2014. *Entity Linking meets Word Sense Disambiguation: a Unified Approach*. Transactions of the Association for Computational Linguistics (TACL), 2, pp. 231–244, 2014.
- Reimers and Gurevich. 2017. *Reporting Score Distributions Makes a Difference: Performance Study of LSTM-networks for Sequence Tagging*.
- Bahdanau D., Cho K., and Bengio Y. 2015. *Neural Machine Translation by Jointly Learning to Align and Translate*. In ICLR 2015.
- Luong M.-T., Pham H. and Manning C. D. 2015. *Effective Approaches to Attention-based Neural Machine Translation*. In Proceedings of EMNLP 2015.
- Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A. N., Polosukhin I. 2017. *Attention Is All You Need*. In Advances in Neural Information Processing Systems.

Kronos-it: a Dataset for the Italian Semantic Change Detection Task

Pierpaolo Basile

University of Bari A. Moro
Dept. Computer Science
E. Orabona 4, Italy

pierpaolo.basile@uniba.it

Giovanni Semeraro

University of Bari A. Moro
Dept. Computer Science
E. Orabona 4, Italy

giovanni.semeraro@uniba.it

Annalina Caputo

ADAPT Centre
Dublin City University
Dublin, Ireland

annalina.caputo@dcu.ie

Abstract

This paper introduces Kronos-it, a dataset for the evaluation of semantic change point detection algorithms for the Italian language. The dataset is automatically built by using a web scraping strategy. We provide a detailed description about the dataset and its generation, and four state-of-the-art approaches for the semantic change point detection are benchmarked by exploiting the Italian Google n-grams corpus.

1 Background and Motivation

Computational approaches to the problem of language change have been gaining momentum over the last decade. The availability of long-term and large-scale digital corpora, and the effectiveness of methods for representing words over time, are the prerequisite behind this interest. However, only few attempts have focused on the evaluation, due to two main issues. First, the amount of data involved limits the possibility to perform a manual evaluation and, secondly, to date no open dataset for the diachronic semantic change has been made available. This last issue has roots in the difficulties of building a gold-standard for detecting the semantic change of terms in a specific corpus or language. The result is a fragmented set of data and evaluation protocols, since each work in this area has used different evaluation datasets or metrics. This phenomenon can be gauged from (Tahmasebi et al., 2019), where it is possible to count at least twenty different datasets used for the evaluation. In this paper, we describe how to build a dataset for the evaluation of semantic change point detection algorithms. In particular, we adopt a

web scraping strategy for extracting information from an online Italian dictionary. The goal of the extraction is to build a list of lemmas with a set of change points for each lemma. The change points are extracted by analysing information about the year in which the lemma with a specific meaning is observed for the first time. Relying on this information we build a dataset for the Italian language that can be used to evaluate algorithms for the semantic change point detection. We provide a case study in which four different approaches are analysed using a unique corpus.

The rest of the article is organised as follows: Section 2 describes how our dataset is built, while Section 3 provides details about the approaches under analysis and the evaluation. Finally, Section 4 closes the paper and provides possible future work.

2 Dataset Construction

The main goal of the dataset is to provide for each lemma a set of years which indicate a semantic change for that lemma. Some dictionaries provide historical information about meanings, for example the year in which each meaning is observed for the first time. The main problem is that generally these dictionaries are not digitally available or they are in a format that is not machine readable.

Regarding the Italian language, the dictionary “Sabatini Coletti”¹ is available on-line. It provides for some lemmas the year in which each meaning was observed for the first time. For example, taking into account the entry for the word “imbarcata” from the dictionary, we capture its original meaning “Group of people who gather to find each other, to leave together”, and other two meanings: 1) “Acrobatic manoeuvre of an air-plane” introduced in 1929; and 2) “fall in love” introduced in 1972.

Copyright ©2019 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

¹https://dizionari.corriere.it/dizionario_italiano/

We setup a web scraping algorithm able to extract this information from the dictionary. In particular, the extraction process is composed of several steps:

1. Downloading the list of all lemmas occurring in the online dictionary with the corresponding URL. We obtain a list of 34,504 lemmas;
2. For each lemma, extracting the section of the web page containing the definition with the list of all possible meanings. We obtain a final list of 34,446 definitions;
3. For each definition, extracting the year in which that meaning was introduced. For a given lemma, we are not able to assign the correct year to each of its meaning, but we can only extract a year associated with the lemma. This happens because the dictionary does not follow a clear template for assigning the year to each meaning. Although associating the year of change to the definition of the meaning is not useful for the purpose of our evaluation, it could help to understand the reason behind the semantic change. We plan to fix this limitation in a further release of the dataset. In the rest of the paper we call change point (CP) each pair (lemma, year);
4. Removing those change points that are expressed in the form “III sec.” (*III century*) because they refer to a broad period of time rather than to a specific year.

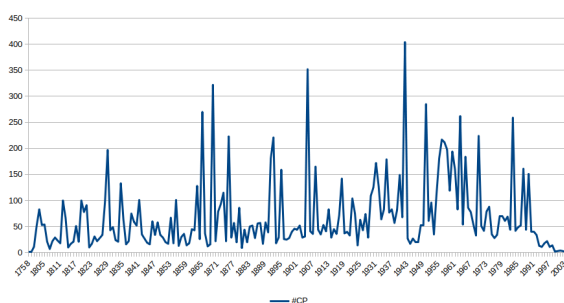


Figure 1: The distribution of change points over time.

The final dataset² contains 13,818 lemmas and 13,932 change points. The average change points for lemma is 1.0083 with a standard deviation of 0.0924. The maximum number of change points

²<https://github.com/pippokill/kronos-it>

for lemma is 3 and the number of lemmas with more than one change point is 113. The oldest reported change point is 1758, while the most recent one is 2003; this suggests that the dictionary is outdated and it does not contain more recent meanings.

The dataset is provided in textual format and reports for each row the lemma followed by a list of years, each one representing a change point. For example:

```
enzima 1892
monopolistico 1972
tamponare 1886 1950
elettroforesi 1931
fuoricorso 1934
```

The low number of change points for lemma reflects the fact that generally, the first meaning has no information about the year it first appeared in or that its time period is expressed in the form of century. This means that all the other meanings are additional meanings introduced after the main one. However, there are some more recent words for which the first year associated with that entry corresponds to the year in which the word is observed for the first time. Unfortunately, it is not easy to automatically discern the two cases.

Finally, we report the distribution of change points over time in Figure 1. The years with a peak are 1942, 1905 and 1869 with respectively 404, 352 and 322 change points.

3 Evaluation

For the evaluation we adopt our dataset as gold-standard and the Italian Google n-grams (Michel et al., 2011) as corpus³.

Google n-grams provides n-grams extracted from the Google Books project. The corpus is composed of several compressed files. Each file contains tab-separated data, each line has the following format: *ngram TAB year TAB match_count TAB volume_count NEWLINE*. For example:

```
parlare di pace e di 2005 4 4
parlare di pace e di 2006 3 3
parlare di pace e di 2007 7 7
parlare di pace e di 2008 2 2
parlare di pace e di 2009 4 4
```

The first line tells us that in 2005, the 5-grams “*parlare di pace e di*” occurred 4 times overall, in 4 distinct books.

³<http://storage.googleapis.com/books/ngrams/books/datasetsv2.html>

In particular, we use the 5-grams corpus and we limit the analysis to words that occur at least in twenty 5-grams. Moreover, we lowercase words and filter out all words that do not match the following regular expression: $[a-z\grave{e}\grave{a}\grave{i}\grave{o}\grave{u}]^+$. We limit our analysis to the period [1900-2012].

In order to build the context words by using 5-grams, we adopt the technique described in (Ginter and Kanerva, 2014). Given a 5-gram $(w_1, w_2, w_3, w_4, w_5)$, it is possible to build eight pairs: (w_1, w_2) (w_1, w_3) \dots (w_1, w_5) and (w_5, w_1) (w_5, w_2) \dots (w_5, w_4) . Then, for each pair (w_i, w_j) , a sliding window method also visits (w_j, w_i) by obtaining 16 training examples from each 5-gram.

We investigate four systems for representing words over time and then we apply a strategy for extracting change points from each technique. Finally, we evaluate the accuracy of each approach by using our dataset as gold standard.

3.1 Representing words over time

We adopt four techniques for representing words over time. The first strategy is based only on word co-occurrences, the other three exploit Distribution Semantic Models (DSM). In particular, the techniques are:

Collocation. This approach is very simple and it is used as baseline. The idea is to extract for each word and each time period the set of relevant collocations. A collocation is a sequence of words that co-occur more often than what would be expected by chance. We extract the collocation by analysing the word pairs extracted from 5-grams and score each word pair using the Dice score:

$$dice(w_a, w_b) = \frac{2 * f_{ab}}{f_a + f_b} \quad (1)$$

where f_{ab} is the number of times that the words w_a and w_b occur together and f_a and f_b are respectively the number of times that w_a and w_b occur in the corpus. Since the Dice score is independent of the corpus size, it is possible to build for each word and each time period a list of collocations by considering only the collocations occurring in a specific period of time. In order to consider only a restricted number of collocations, we take in account only the collocations with a Dice value above 0.0001. For each word and each

time period we obtain a list of collocations with the associated Dice score. For example, a portion of the list of collocations for the word *pace* (*peace*) in the period 1980-1984 is reported as follows:

```
pace guerra 0.007223173
pace giustizia 0.0068931305
pace trattati 0.0067062946
pace trattative 0.006033537
```

Temporal Random Indexing (TRI). TRI (Jurgens and Stevens, 2009) is able to build a word space for each time period where each space is comparable to one another. In each space, a word is represented by a dense vector and it is possible to compute the cosine similarity between word vectors across time periods. In order to build comparable word spaces, TRI relies on the incremental property of the Random Indexing (Sahlgren, 2005). More details are provided in (Basile et al., 2014) and (Basile et al., 2016).

Temporal Word Analogies (TWA). This approach is able to build diachronic word embeddings starting from independent embedding spaces for each time period. The output of this process is a common vector space where word embeddings are used for computing temporal word analogies: word w_1 at time t_i is like word w_2 at time t_j . We build the independent embedding spaces by using the C implementation of word2vec with default parameters (Mikolov et al., 2013). More details about this approach are reported in (Szymanski, 2017).

Procrustes (HIST). This approach aligns the learned low-dimensional embeddings by preserving cosine similarities across time periods. More details are available in (Hamilton et al., 2016). We apply the alignment to the same embeddings created for TWA.

All approaches are built using the same vocabulary and the same context words generated starting from the 5-grams as previously explained.

3.2 Building the time series

In order to track how the semantics of a word changes over time we need to build a time series.

A time series is a sequence of values, one for each time period, that indicates the semantic shift of that word in the specific period. In our evaluation, we split the interval [1900-2012] in time periods of five years each.

The time series are computed in different ways according to the strategy used for representing the words. In particular, the values of each time series $\Gamma(w_i)$ associated to the word w_i is computed as follow:

- Collocation: given two lists of collocations related to two different periods, we compute the cosine similarity between the two lists by considering a list as a Bag-of-Collocations (BoC). In this case each point k of the series $\Gamma(w_i)$ is the cosine similarity between the BoC at time T_{k-1} and the BoC at time T_k ;
- TRI: we use two strategies, (*point-wise* and *cumulative*), as proposed in (Basile et al., 2016). The point-wise approach captures how the word vector changes between two time periods, while the cumulative analyses captures how the word vector changes with respect to all the previous periods. In the point-wise approach, each point k of $\Gamma(w_i)$ is the cosine similarity between the word vector at time T_{k-1} and the word vector at time T_k , while for the cumulative approach the point k is computed as the cosine similarity between the average word vectors of all the previous time periods T_0, T_1, \dots, T_{k-1} and the word vector at time T_k ;
- TWA: we exploit the word analogies across time and the common vector space for capturing how a word embedding changes across two time periods as reported in (Szymanski, 2017);
- HIST: time series are built by using the pairwise similarity as explained in (Hamilton et al., 2016).

We obtain seven time series as reported in Tables 1 and 2. In particular: BoC is build on temporal collocations; TRI_{point} and TRI_{cum} are based on TRI by using respectively point-wise and cumulative approach; TWA_{int} and TWA_{uni} are built using TWA on words that are common (intersection) to all the periods (TWA_{int}) and on the union of words (TWA_{uni}). The same procedure

is used for $HITS$ obtaining the two time series $HIST_{int}$ and $HIST_{uni}$.

For finding significant change points in a time series, we adopt the strategy proposed in (Kulkarni et al., 2015) based on the Mean Shift Model (Taylor, 2000).

3.3 Metrics

We compute the performance of each approach by using Precision, Recall and F-measure. However, assessing the correctness of the change points generated by each system is a not easy task. A change point is defined as a pair (*lemma*, *year*). In order to adopt a soft match, when we compare the change points provided by a system with respect to the change points reported in the gold standard, we take into account the absolute value of the difference between the year predicted by the system and the year provided in the gold standard.

As a first evaluation (exact match), we impose the difference between the detected year and the gold standard to be less or equal than five, which is the time period span of our corpus. As a second evaluation (soft match), we impose only that the predicted year is greater or equal than the change point in the gold standard. This is a common methodology adopted in previous work.

For a fairer evaluation, we perform the following steps:

- We remove from the gold standard all the change points that are outside of the period under analysis ([1900-2012]);
- We remove from the gold standard all the words that are not represented in the model under evaluation. This operation is necessary because (1) the previous filtering step can exclude some words;(2) there are words that do not appear in the original corpus.

Since the gold standard contains lemmas and not words, we perform a lemmatization of each output by using Morph-it! (Zanchetta and Baroni, 2005).

3.4 Results

Results of Precision (P), Recall (R) and F-measure (F) are reported in Table 1. We can observe that generally we obtain a low F-measure. This is due to a large number of false positive change points detected by each system.

Γ	exact match			soft match		
	P	R	F	P	R	F
<i>BoC</i>	.0034	.0084	.0049	.0274	.0670	.0389
<i>TRI_{point}</i>	.0056	.0394	.0098	.0248	.1750	.0434
<i>TRI_{cum}</i>	.0058	.0387	.0101	.0251	.1672	.0436
<i>TWA_{int}</i>	.0034	.0009	.0015	.0165	.0046	.0072
<i>TWA_{uni}</i>	.0052	.0060	.0056	.0373	.0435	.0402
<i>HIST_{int}</i>	.0024	.0048	.0032	.0111	.02211	.0148
<i>HIST_{uni}</i>	.0022	.0066	.0033	.0118	.0356	.0177

Table 1: Results of the evaluation.

Γ	exact match			soft match		
	P	R	F	P	R	F
<i>BoC</i>	.0361	.1243	.0560	.2881	.9930	.4466
<i>TRI_{point}</i>	.0581	.2244	.0923	.2581	.9973	.4100
<i>TRI_{cum}</i>	.0610	.2308	.0959	.2617	.9979	.4146
<i>TWA_{int}</i>	.0402	.2000	.0670	.1960	.9750	.3264
<i>TWA_{uni}</i>	.0526	.1367	.0759	.3794	.9866	.5480
<i>HIST_{int}</i>	.0344	.2147	.0593	.1569	.9791	.2704
<i>HIST_{uni}</i>	.0314	.1842	.0536	.1675	.9836	.2863

Table 2: Results of the evaluation obtained by considering only common lemmas between the gold standard and the system output.

The best approach in both evaluations is *TRI_{cum}*. Considering the *exact match* evaluation, the difference in performance is remarkable since generally TRI has a high recall. In the *soft match* evaluation, *TWA_{uni}* obtains the best precision, while the simple *BoC* method is able to achieve good results compared with more complex approaches such as *TWA_{int}* and *HIST*.

The results of the evaluation prove that the task of semantic change detection is very challenging; in particular, the large number of false positive drastically affects the performance.

Further analyses are necessary to understand which component affects the performance. In this preliminary evaluation, we adopt a unique approach for detecting the semantic shift. An extended benchmark is necessary for evaluating several approaches for detecting semantic change points.

The systems are built on a vocabulary that is larger than both the original dictionary and the gold standard. For that reason, we provide an additional evaluation in which we perform an ideal analysis by evaluating only lemmas that are common to the gold standard and the system output. The goal of this analysis is to measure the ability of correctly identifying change points for those

lemmas that are represented in both the gold standard and the system. Results of this further evaluation are provided in Table 2

For the *exact match* evaluation, *TRI_{cum}* obtains the best F-measure as in the first evaluation, while *TWA_{uni}* achieves a very good performance in the *soft match* evaluation.

The plot in Figure 2 reports how the F-measure increases according to the time span that we adopt in the soft match. In particular, the X-axis reports the maximum absolute difference between the year in the gold standard and the year predicted by the system. We can observe that under 20 years *TRI* provide better performance than *TWA*, and after 60 years all the approaches reach a stable F-measure value.

4 Conclusion and Future Work

In this paper, we provide details about the construction of a dataset for the evaluation of semantic change point detection algorithms. In particular, our dataset focused on the Italian language and it is built by adopting a web-scraping strategy. We provide a usage example of our dataset by evaluating several approaches for the representation of words over time. The results prove that the task of detecting semantic shift is challenging due to a large

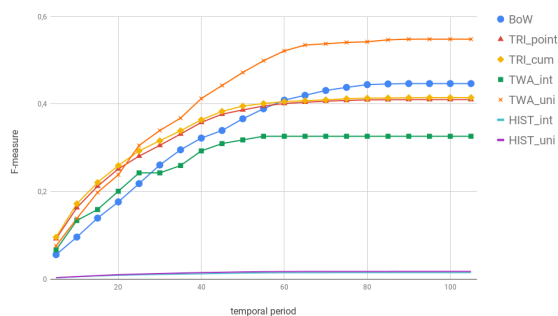


Figure 2: The plot shows how the F-measure increases according to the time span used in the soft match.

number of detected false positive. As future work, we plan to investigate further methods for building time series and detecting semantic shifts in order to improve the overall performance. Moreover, we plan to fix some issues of our extraction process in order to improve the quality of the dataset itself.

Acknowledgements

This work was supported by the ADAPT Centre for Digital Content Technology, funded under the Science Foundation Ireland (SFI) Research Centres Programme (Grant SFI 13/RC/2106) and is co-funded under the European Regional Development Fund and by the European Unions Horizon 2020 (EU2020) research and innovation programme under the Marie Skłodowska-Curie grant agreement No.: EU2020 713567. The computational work has been executed on the IT resources made available by two projects, ReCaS and PRISMA, funded by MIUR under the program “PON R&C 2007-2013”.

References

- Pierpaolo Basile, Annalina Caputo, and Giovanni Semeraro. 2014. Analysing word meaning over time by exploiting temporal random indexing. In *First Italian Conference on Computational Linguistics CLiC-it*.
- Pierpaolo Basile, Annalina Caputo, Roberta Luisi, and Giovanni Semeraro. 2016. Diachronic analysis of the italian language exploiting google ngram. *CLiC it*, page 56.
- Filip Ginter and Jenna Kanerva. 2014. Fast training of word2vec representations using n-gram corpora.
- William L Hamilton, Jure Leskovec, and Dan Jurafsky. 2016. Diachronic word embeddings reveal

statistical laws of semantic change. *arXiv preprint arXiv:1605.09096*.

David Jurgens and Keith Stevens. 2009. Event detection in blogs using temporal random indexing. In *Proceedings of the Workshop on Events in Emerging Text Types*, pages 9–16. Association for Computational Linguistics.

Vivek Kulkarni, Rami Al-Rfou, Bryan Perozzi, and Steven Skiena. 2015. Statistically significant detection of linguistic change. In *Proceedings of the 24th International Conference on World Wide Web*, pages 625–635. International World Wide Web Conferences Steering Committee.

Jean-Baptiste Michel, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K Gray, Joseph P Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, Jon Orwant, et al. 2011. Quantitative analysis of culture using millions of digitized books. *science*, 331(6014):176–182.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Magnus Sahlgren. 2005. An introduction to random indexing.

Terrence Szymanski. 2017. Temporal word analogies: Identifying lexical replacement with diachronic word embeddings. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 448–453, Vancouver, Canada, July. Association for Computational Linguistics.

Nina Tahmasebi, Lars Borin, and Adam Jatowt. 2019. Survey of computational approaches to lexical semantic change. *arXiv:1811.06278v2*.

Wayne A Taylor. 2000. Change-point analysis: a powerful new tool for detecting changes.

Eros Zanchetta and Marco Baroni. 2005. Morph-it! a free corpus-based morphological resource for the italian language. In *Proceedings of corpus linguistics*.

How do Physiotherapists and Patients talk?

Developing the *RiMotivAzione* dialogue corpus.

Andrea Bolioli[1], Francesca Alloatti[1,2], Mariafrancesca Guadalupi[1],
Roberta Iolanda Lanzi[1], Giorgia Pregnolato[3], Andrea Turolla[3]¹

¹CELI - Language Technology, Italy

²Department of Computer Science - Università degli Studi di Torino, Italy

³IRCCS Fondazione Ospedale San Camillo, Italy

Abstract

The research project *RiMotivAzione* aims at helping post-stroke patients who are following an arm and hand rehabilitation path. In this paper we present the *RiMotivAzione* corpus, the first collection of dialogues between physiotherapists and patients recorded in an Italian hospital and annotated following the RIAS annotation protocol. We describe the dataset, the methodologies applied and our first investigations on relevant features of the dialogue process. The corpus was the basis for the design of a conversational interface integrated with a wearable device for rehabilitation, to be used by the patient during the exercises that he or she may perform independently.¹

1 Introduction

In recent years, computational linguistics and medical research have started to collaborate in order to analyze the communication in the healthcare domain, in particular between clinicians and patients. From a medical perspective, linguistic analysis and dialogue modeling can be used to better understand and potentially enhance communication in different healthcare settings (Sen et al., 2017; Chang et al., 2013; Marzuki et al., 2017), as well as to identify "preclinical" or "pre-symptomatic" diseases for specific ranges of patients, e.g. discovering early linguistic signs of cognitive decline (Beltrami et al., 2018).

Natural Language Processing (NLP) technologies are also used to develop new communicative tools, e.g. virtual assistants, to alleviate the burden on medical personnel or shift to a home-based patient-centered model of care. Through mHealth (mobile health), for example, people can receive assistance at home, and monitoring devices can check the well-being of a person (Sezgin et al.,

2018). A recent review of scientific literature about Artificial Intelligence and IoT in healthcare can be found in (Shah and Chircu, 2018).

The research project *RiMotivAzione* aims at helping the patients who suffered from a stroke and are following an arm and hand rehabilitation path. The goal is to motivate the patients to follow the assigned exercises through the use of a new wearable device with motion sensors developed by the Istituto Italiano di Tecnologia (IIT), integrated with a visual App and a conversational interface. This last component guides the user through the therapeutic path proposing the exercises, giving advice and asking for feedback.

The implementation of voice technologies in the healthcare domain allows for patients with motor impairments to interact with devices through spoken language (Moore et al., 2018), while arm and hand are busy performing the assigned exercises. The interaction is seamless and spontaneous. The patient can keep up autonomously with the therapy thanks to the guidance provided by the voice assistant. The physiotherapist can monitor the patients at a distance, to evaluate their progress, and he can prevent a situation of therapy neglect by the patient, while the latter is motivated to stick to the path and he can reach his rehabilitation goals on time. Needless to say, these digital assistants are not meant to substitute the clinician.

2 Methodological Background and Related Work

As we described in the previous section, the study of communication and conversation in the medical domain is growing in the last years, as well as the introduction of conversational agents in the healthcare sector. A review of current applications and evaluation measures of conversational agents used for health-related purposes can be found, for example, in (Laranjo et al., 2018). Otherwise, there is no systematic review of scientific literature

¹Copyright © 2019 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

concerning the linguistic analysis of dialogues in healthcare. Some scientific studies describe how communication can influence clinical outcomes in the rehabilitation setting, e.g. how patient satisfaction, decision-making, and stress level correlate with physicians' communicative acts (Hall and Roter, 2012). Some researchers propose methods to detect and track topics in psycho-therapeutic conversations (Chaoua et al., 2018). Other researchers conducted an analysis of actual communicative behaviors, including nonverbal ones, between physicians and patients in rehabilitation, using transcription and coding of utterances (Chang et al., 2013).

The analysis of speech acts and conversational interaction can play a relevant role in dialogue modeling for healthcare thanks to the classification of utterances, the analysis of dialogue turns and threads, the discovery of recurrent patterns. Speech acts have been investigated in linguistics and computational linguistics for long. Specifically, the task of automatic speech act recognition has been addressed leveraging both supervised and unsupervised approaches (Basile and Novielli, 2018). Otherwise, in the healthcare domain there is still much room for investigation.

In the *RiMotivAzione* project, we deal with physiotherapy sessions in a hospital. The task is to collect and analyze para-linguistic and linguistic data, according to the aforementioned goal of the research project. In this specific setting, i.e. conversational analysis of physician-patient discourse, the most widely used method is the Roter Interaction Analysis System (RIAS). RIAS was developed as a tagset for coding medical dialogue since 1991 by Debra Roter et al. (Roter, 1991; Roter and Larson, 2002) and it has been constructed as to be viable for all kind of sessions, e.g. conversations in the oncological setting (2017), between patients and psychotherapists or even patients and pharmacists. Moreover, RIAS was originally developed to annotate audio, while we transcribed the speech and annotated the transcriptions. This is motivated by the NLP analysis we wanted to perform on the text, e.g. syntactic and semantic analysis, machine learning, automatic dialogue act classification. Other dialogue annotation schemes exist, namely (Bunt et al., 2017; Serban et al., 2017; Stolcke et al., 2000), that includes rich taxonomies of communicative functions. The ISO 24617-2 standard, for example, includes the

specification of the Dialogue Act Markup Language (DiAML), used in many annotated corpora. In *RiMotivAzione* project, we deemed RIAS as the most useful one for its specific focus on medical conversation. Even though RIAS is the closest domain tagset to annotate our corpus, some problems still emerged and they will be presented in next section.

3 Corpus Annotation

The *RiMotivAzione* corpus includes two complete cycles of physiotherapy sessions with two patients in post-stroke rehabilitation (namely, P1 and P2) and three physiotherapists (T1, T2, T3). The interviews were video recorded in IRCCS Fondazione Ospedale "San Camillo" in Venice. Each session lasted about 1 hour. The physiotherapy cycle for patient P1 included 14 sessions, while P2 took 16 sessions. Therefore the total duration of recordings is about 30 hours.

The patients were carefully selected by the doctors, since they must present some features. Above all, they had to agree to be part of the experimentation and they needed to talk in Italian. In an environment where dialect is still strong, their ability to speak Italian was not to be treated lightly. Moreover, the patients did not have to present any issues related to aphasia. These requirements restrained the viable options to two candidates.

Both speakers were encouraged to talk freely about any topic that may have emerged. Their only constraint was the use of Italian; when people slipped into dialectal terminology (in this case, Venetian), it was explicitly marked with the <dialect> tag in the corpus. The audio tracks were transcribed and annotated following Savy's (2005) guidelines for orthographic transcription for spoken Italian, where applicable. As a pre-processing, we used two Automatic Speech Recognition (ASR) systems, i.e. Google Speech-to-Text and Nuance Transcription Engine. Automatic transcriptions were corrected manually and anonymized. Video and audio tracks have been separately saved for future projects.

Overlapping between the two speakers and pauses were not marked, as it was not relevant to our study. Similarly, any intervention in the dialogue from a third party was not transcribed since our interest was solely in the doctor and patient's linguistic behaviours. Each dialogue turn of the corpus was annotated by two different annotators

following the RIAS guidelines. All the annotators have a background in linguistics and a specific education about linguistic corpora. As a single dialogue turn may contain more than one sentence and more than one speech act, the tags assigned to each turn may be more than one.

RIAS tagset includes 29 categories divided in four macro-categories called Medical Interview Functions (MIF) that cover the majority of the exchanges between a doctor and a patient: Data Gathering, Information Exchange, Emotional Expression and Responsiveness, Partnership Building and Activation. Table 1 contains the list of categories occurring at least 200 times in the corpus, together with examples.

To the best of the authors' knowledge, the RIAS system has never been used to annotate sessions of physiotherapy until now. This means that not all of the tags applied completely to the situation, or that some tags may be under-represented compared to other studies: for instance, the tag *Concerns* was applied to few sentences, since patients in physiotherapy sessions may inherently express less concern than oncological patients.

All the categories defined in Roter et al. (2017) were used. Moreover, two more tags were added to include all the exchanges: *Unclear* and *Technical problems*. The first applied to incomplete sentences, unintelligible ones (also marked with the `<unclear>` tag), or even in cases where the sentence referred to the physical context, making the general meaning impossible to retrieve for the annotator. The second tag applied to situations where the wearable device wasn't working properly, therefore resulting in some technical issue out of the scope of the therapy.

Another issue concerns the use of irony. Specifically, Patient 2 heavily employed irony while talking to the therapist, even when the dialogue concerned his health and well-being. Irony is hard to interpret, resulting in the difficulty to assign correctly a tag to those sentences. Tag *Jokes* was used in this case, and where inappropriate, a discussion between the annotators oriented the choice.

As the annotation task was difficult and it was inherently affected by subjectivity, we measured the resulting inter-annotator agreement and we put in place strategies to solve the disagreement, in order to annotate all the dialogue turns. The agreement calculated at this stage, according to the Co-

hen's score, was promising ($k = 0.63$). In case of disagreement (about 25% of the data), the process was followed by reconciliation or a final decision by a super annotator, where the two annotators could not overcome the disagreement.

The *RiMotivAzione* corpus has been built and archived according to GDPR norms. It is not publicly available but it can be requested to the authors for research purposes.

4 Corpus Analysis

The *RiMotivAzione* corpus contains about 98778 tokens. The total number of dialogue turns is 7670: 3377 dialogue turns in P1 sessions, 4293 in P2 sessions.

In Table 2 and Table 3 we reported the number of types, tokens, the ratio between types and tokens (the Lexical Richness Index) and the number of questions for the two patients.

It is worth noticing that Lexical Richness Index ranges from 0 to 1 and it is closer to 0 in the doctors' speech, meaning that medical personnel employ a poorer vocabulary while talking to a patient. This is due to the fact that a therapist needs to stick to a protocol and cannot digress over a certain limit. On the other hand, the patient talks quantitatively less: he pronounces fewer words, and most of the time those words are simple answers to the questions posed by the clinician. The patient talks less but he can wander more across conversation topics: he may disclose some personal detail about his life or just chit chat. This behavior is actually encouraged by the therapist, since it makes the therapy session less dull and more spontaneous for both the participants (Delany et al., 2010; Edwards et al., 2004). To sum up, the doctor needs to talk a lot to instruct the patients about the exercise they need to fulfill, as well as to ask questions (mainly regarding general well-being and inquiries about the therapy itself). Meanwhile, the patient may talk less because most of the time he just has to answer short questions (such as "*Does it hurt?*"); or, when he talks more, it is about some external topic which generates an increment in the vocabulary richness index.

As the main goal of the study is to replicate the clinician's communicative style onto a conversational interface, the major interest is on how the therapists talk, rather than the patients. Patients' manner of speaking is taken into consideration when imagining all the orders or phrases

Specific RIAS code	Examples
Social talk	non vedevo l'ora di venirla a trovare.
Directions	per scendere chiudo, per salire apro la mano.
Agreements	esatto, perché lo abbiamo registrato proprio così.
Medical condition	un po', poco, fastidio più che male.
Approvals	bravissimo.
Unclear	[dialect] vara!
Therapeutic regimen	venerdi faremo la parte clinica ti farò io la scala di valutazione.
Jokes and laughter	ci vediamo domani, è più una minaccia che un invito.
Asking for understanding	vorrei portarla così, hai capito?
Checking for understanding	chiudo le dita. così?
Concerns	sei sicura che funziona?
CeQ Medical condition	a fare gli esercizi non ha dolore?

Table 1: Tags and examples of categories occurring at least 200 times in the corpus.

Parameters	Patient 1	Therapist
Types	2065	3017
Tokens	10533	39305
Lexical Richness Index	0,19	0,07
Questions	40	667

Table 2: Patient 1 corpus.

Parameters	Patient 2	Therapist
Types	2451	2406
Tokens	18233	30707
Lexical Richness Index	0,13	0,07
Questions	380	805

Table 3: Patient 2 corpus.

Word	Frequency
<i>vai</i>	1166
<i>apri</i>	432
<i>rilassa</i>	400
<i>bravissimo</i>	353
<i>mantieni</i>	314
<i>bravo</i>	288
<i>lascia</i>	199
<i>fare</i>	187
<i>prova</i>	156
<i>ottimo</i>	153

Table 4: Most frequent Verbs and Adjectives used by therapist 1.

that the user could say to the voice assistant to express his needs. Table 4 and Table 5 list the most frequent Verbs and Adjectives pronounced by the physiotherapists. Apart from "Okay", which is the most frequent word for both therapists (1231 and 1019 occurrences), both therapists often use adjectives of positive value: *bravissimo*, *bravo*, *ottimo*, *buono*. Other frequent words are mainly verbs expressed at the first plural person, such as *we do*, *we'll try*, or equivalent expressions (*let's relax*). The use of the "we" is a communication element that aims at putting on the same level the clinician and the patient; the goal is to make the patient feel more comfortable and therefore enhancing the probability of therapy adherence. At the same time, adjectives such as "good" and "very good" praise the patient's efforts, underlining the progress he is making. The psychological component is of paramount importance during phys-

iotherapy, especially for patients that suffered a stroke (Palma and Sidoti, 2019).

The quantitative analysis operated over the annotated corpus confirms the qualitative remarks made so far. In Figure 1 we present the distribution of dialogue tags, both for patients and therapists, i.e. the distribution of utterance type according to RIAS categories. We plotted on a logarithmic scale the frequencies of the tags.

Sentences annotated as *Social talk* were abundant, while those marked as *Concerns* were copious just for a patient, because he was frustrated about his health situation and the difficulties to manage the physiotherapy. During the sessions with Patient 1, the physiotherapist was able to engage a conversation about a hobby of his (motorcycles); even though this discussion topic is not relevant to the therapy, the fact that they were talking about something interesting for the patient contributed to the improvement of his med-

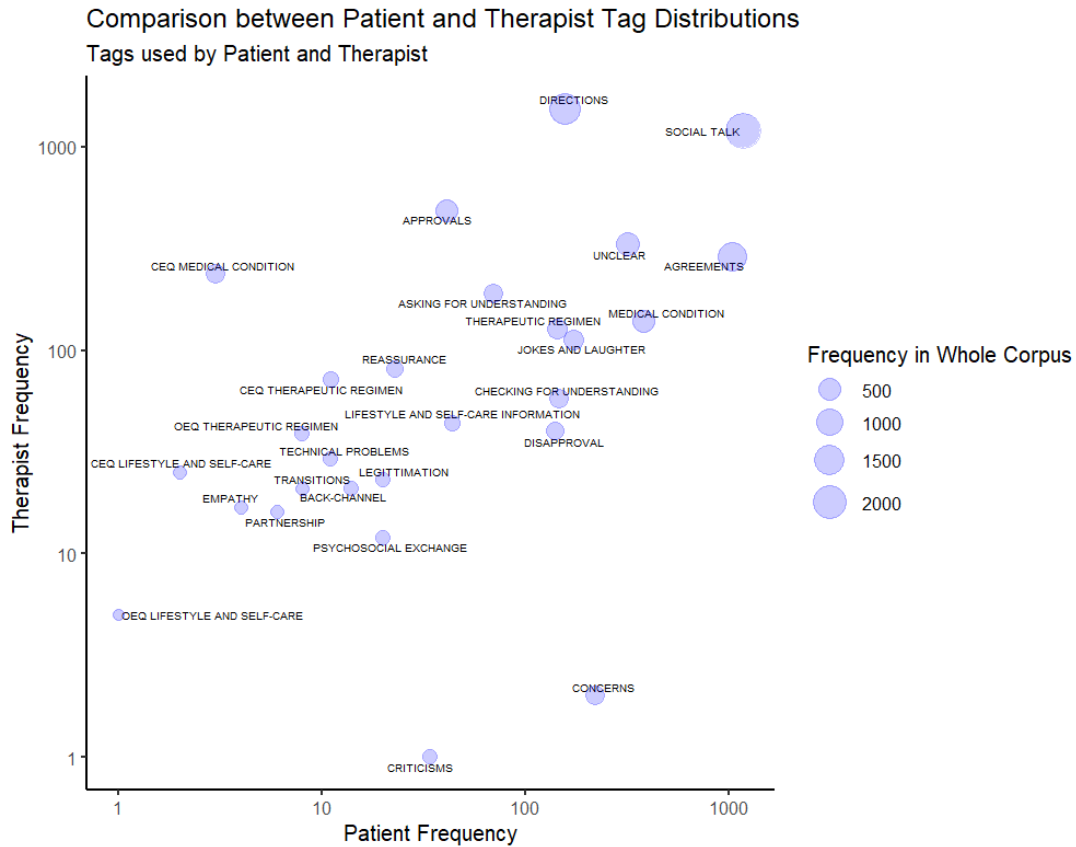


Figure 1: Distribution of dialogue tags in *RiMotivAzione* corpus

Word	Frequency
<i>vai</i>	340
<i>proviamo</i>	199
<i>apro</i>	198
<i>pronto</i>	174
<i>facciamo</i>	134
<i>attento</i>	124
<i>andare</i>	123
<i>scendere</i>	120
<i>vediamo</i>	115
<i>fare</i>	111

Table 5: Most frequent Verbs and Adjectives used by therapist 2.

ical condition (Gard and Gyllenstein, 2000).

All of these conversational elements are put in place willingly by the clinician and, even more, it is the style patients are used to. In the voice assistant design we try to mirror these strategies, providing praises when appropriate and asking questions to constantly monitor the user’s well-being. The data extracted from the transcription and the annotation represents the most frequent linguistic

behaviors emerged during the conversations. These patterns were used to build the conversational style and infrastructure of the dialogue system.

5 Conclusions and Next Steps

We created a corpus of conversations between patients and clinicians, in Italian, and we annotated the dialogue turns according to the Roter Interaction Analysis System (RIAS). This corpus was the first step in the design of a conversational interface integrated with a smart wearable device, to guide and assist the patients through the exercises assigned by the physiotherapist.

The first step in the future work will be to deepen the linguistic analysis conducted on the corpus, especially regarding the tagged dialogue acts. A stronger qualitative investigation over the data will be carried out. The second step will be to enrich the dataset: unfortunately, only two patients were deemed appropriate for the experimentation, while a corpus should contain dialogues from more speakers.

The *RiMotivAzione* corpus can be requested to

the authors for research purposes.

The system prototype will be tested in San Camillo Hospital by a set of stroke patients, following the clinical trial procedures. Thanks to the results of the test, we will produce experimental data to investigate if and how a voice assistant integrated with a wearable device can increase the effectiveness of the therapy.

6 Acknowledgments

RiMotivAzione is a two-year Research and Innovation project supported by POR FESR 2014-2020 Regione Piemonte. The partners are Koiné Sistemi, CELI, IRCCS Fondazione Ospedale San Camillo, Synesthesia, Istituto Italiano di Tecnologia (IIT) and Morecognition. We are thankful to our colleagues and project partners, in particular Paolo Ariano and Nicoló Celadon.

References

- P. Basile and N. . Novielli. 2018. Overview of the evalita 2018 italian speech act labeling (ilisten) task. In *Proceedings of the Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2018) co-located with the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018)*.
- D. Beltrami, G. Gagliardi, Rossini Favretti, E. R., Ghidoni, F. Tamburini, and L. Calzá. 2018. Speech analysis by natural language processing techniques: a possible tool for very early detection of cognitive decline? *Frontiers in Aging Neuroscience*, 10:369.
- Harry Bunt, Volha Petukhova, David Traum, and Jan Alexandersson. 2017. Dialogue act annotation with the iso24617-2 standard. *Multimodal Interaction with W3C Standards*.
- C.L. Chang, B.K. Park, and S.S. Kim. 2013. Conversational analysis of medical discourse in rehabilitation: A study in korea. *The journal of spinal cord medicine*, 36(1):24–30.
- I. Chaoua, D. R. Recupero, S. Consoli, A. Harma, and R. Helaoui. 2018. Detecting and tracking ongoing topics in psychotherapeutic conversations. *AIH@IJCAI*, pages 97–108.
- C.M. Delany, I. Edwards, G.M. Jensen, and E. Skinner. 2010. Closing the gap between ethics knowledge and practice through active engagement: an applied model of physical therapy ethics. *Physical Therapy*, 90(7):1068–1078.
- I. Edwards, M. Jones, J. Carr, A. Braunack-Mayer, and G.M. Jensen. 2004. Clinical reasoning strategies in physical therapy. *Physical Therapy*, 84(4):312–330.
- G. Gard and A. L. Gyllenstein. 2000. The importance of emotions in physiotherapeutic practice. *Physical Therapy Reviews*, 5(3):155–160.
- J. A. Hall and D. L. Roter. 2012. Physician-patient communication. In H. A. Friedman, editor, *The Oxford Handbook of Health Psychology*. Oxford University Press.
- L. Laranjo, A.G. Dunn, H.L. Tong, A.B. Kocaballi, and al. 2018. Conversational agents in healthcare: a systematic review. *Journal of the American Medical Informatics Association*, 25(9):1248–1258.
- E. Marzuki, C. Cummins, H. Rohde, H. Branigan, and G. Clegg. 2017. Resuscitation procedures as multi-party dialogue. In *Proc. SEMDIAL 2017 (SaarDial) Workshop on the Semantics and Pragmatics of Dialogue*, pages 60–69.
- R.J. Moore, M.H. Szymanski, R. Arar, and G. J. Ren. 2018. *Studies in Conversational UX Design*. Springer.
- S. Palma and E. Sidoti. 2019. La comunicazione nei processi di cura. *COMUNIT IMPERFET*, 46(4):243–251.
- D. Roter. 1991. *The Roter method of interaction process analysis (RIAS manual)*. The Johns Hopkins University, Baltimore.
- D. Roter, S. Isenberg, and L. Czaplicki. 2017. The roter interaction analysis system: Applicability within the context of cancer and palliative care. *Oxford Textbook of Communication in Oncology and Palliative Care*, pages 717–726.
- D. Roter and S. Larson. 2002. The roter interaction analysis system (rias): utility and flexibility for analysis of medical interactions. *Patient education and counseling*, pages 128–132.
- R. Savy. 2005. *Specifiche per la trascrizione ortografica annotata dei testi in Italiano Parlato. Analisi di un dialogo*. Liguori, Napoli.
- T. Sen, M.R. Ali, M.E. Hoque, R. Epstein, and P. Duberstein. 2017. Modeling doctor-patient communication with affective text analysis. In *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 170–177.
- Iulian Vlad Serban, Ryan Lowe, Peter Henderson, and Joelle Pineau Laurent Charli and. 2017. [A survey of available corpora for building data-driven dialogue systems](https://arxiv.org/abs/1512.05742). *arXiv:1512.05742*.
- E. Sezgin, S. Yildirim, S. Ozkan-Yildirim, and E. Sumuer. 2018. *Current and Emerging MHealth Technologies: Adoption, Implementation, and Use*. Springer.
- R. Shah and A. Chircu. 2018. Iot and ai in healthcare: A systematic literature review. *Issues in Information Systems*, 19(3):33–41.

Andreas Stolcke, Klaus Ries, and Elizabeth Shriberg
Noah Coccaro. 2000. Dialogue act modeling for
automatic tagging and recognition of conversational
speech. *Computational Linguistics*, 26(3).

Standardizing Language with Word Embeddings and Language Modeling in Reports of Near Misses in Seveso Industries

Simone Bruno*, Silvia Maria Ansaldi⁺, Patrizia Agnello⁺, Fabio Massimo Zanzotto*

* University of Rome Tor Vergata, fabio.massimo.zanzotto@uniroma2.it

⁺ INAIL, {s.ansaldi,p.agnello}@inail.it

Abstract

Standardizing technical language has always been a strong necessity of the technological society. Today, Natural Language Processing as well as the widespread use of computerized document writing can give a tremendous boost in reaching the goal of standardizing technical language. In this paper, we propose two methods for standardizing language. These methods have been applied to the dataset of near misses, collected during the inspections at Major-Accident Hazard (MAH) Industries.¹

1 Introduction

Standardizing technical language has always been a strong necessity of the technological society. Artifacts, objects, measures and so on should have a clear name and a clear description in order to assure mutual understanding, which leads to the reach of important goals in building and controlling machines. However, language standardization has always the same problem: language is a *social phenomenon* (de Saussure, 1916). Hence, whenever a group gather for designing or using a technical object, this group can develop a specific sub-language or just adapt the shared technical language. This adapted sub-language can be then effectively used to refer to parts of this technical object. It is sufficient that group members agree upon this language and the mutual understanding occur. Yet, the language used by the specific group may prevent the others to understand what is written.

Nowadays, Natural Language Processing as well as the widespread use of computerized document writing can give a tremendous boost in reaching the goal of standardizing technical language. Language in use can be captured and, then,

analyzed. Technical people can be invited to use a standardized dictionary with writing suggestions.

This paper discusses two different methods of standardizing technical languages, which have been applied to a dataset of near misses coming from the inspections at Major-Accident Hazard (MAH) industries, named also “Seveso” industries. . The first method aims to help a standardization agency to propose the standard language for writing these reports. We proposed to analyze language in use by word embedding similarity such that the standardization agency can propose a language that is close to the one used. The second method aims to reduce the use of unnecessary synonyms in compiling reports of near misses. In fact, using unnecessary synonyms may result in confusing the report. For this problem, we propose to use a combination of language modeling derived from the CBOW model of the word2vec (Mikolov et al., 2013) along with a classical cosine similarity using word embeddings. We experimented with a dataset of anonymized reports of near misses from Seveso Industries, which INAIL has institutionally collected.

The rest of the paper is organized as follows. Section 2 describes the application scenario and the dataset. Section 3 shortly reports on the models used in this study and proposes the two tasks. Section 4 reports on a preliminary analysis of the possible results of the system. Finally, Section 5 draws some conclusions and proposes further investigations.

2 Background

2.1 Scenario

The European “Seveso” Directive deals with the control of major-accident hazards involving dangerous substances, which can cause toxic clouds, fire, or explosion with consequences to people, as-

¹Copyright © 2019 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

Ref: 66	Data (Date): 2007-02-15	
Titolo (Title):	Trasudamento OCD da serbatoio di stoccaggio OCD	
Descrizione (Description):	Durante le operazioni di riempimento del serbatoio K2 da nave cisterna, si è notato un leggero trasudamento di OCD per corrosione del mantello (sottospessore localizzato mantello serbatoio) a quota 6 metri circa lungo il lato ovest. Uno degli operatori addetto ai controlli durante la scarica della nave ha evidenziato l'evento. L'operazione di scarica della nave cisterna è stata fermata. Non si sono avuti rilasci, a meno del leggero trasudamento.	
Sistemi tecnici critici (Critical Technical System):	serbatoio	
Sostanza (Substance):	olio combustibile (ocd)	
Fattori gestionali (Managing Factor)	Descrizione (Description)	Azioni pianificate (Planned Actions)
4.iv	Fallimento procedure di manutenzione e controllo.	Fuori servizio e bonifica del serbatoio.

Figure 1: Sample Report of a Near Miss within the European Seveso Directive - Italian Localization: Translation is provided for Field Names

sets and environment, also outside the establishments. All European Member States apply this Directive, which foresees periodical inspections by National Competent Authorities; in Italy, Inail is one of these authorities. During the inspection, the operator has to provide the inspectors with the list of near-misses, minor incidents, and accidents occurred in the last ten years. Near misses and minor incidents are events of losses of containment, involving dangerous substances with none or minor consequences, respectively. In Seveso industries, the registration and the analysis of near misses is strongly recommended, as they can be considered as precursors of incidents with serious consequences.

In Italy, under Seveso legislation, there are about a thousand industries, including refineries, petrochemical, and chemical. One of the pillars of the Seveso Directive is the Safety Management System SMS, whose adoption is mandatory for the establishments' operators, in order to control major accident hazards.

The Safety Management System (SMS), implemented by the establishment's operator, addresses technical measures and organizational procedures in order to guarantee human, asset and environmental safety, with a view to the prevention of major accident or the mitigation of their consequences.

In the recent inspections, the focus is often toward the study of the incidents and near misses (see Figure 1). The approach based on near-miss discussion is considered more "risk based" as it is able to single out the critical issues of the safety system.

2.2 Corpus

The dataset refers to the near misses reports provided by the operators of "Seveso" establishments. The collection of reports on near misses, hereafter referred as REP corpus, consists of 1300 documents called "operative experiences". These operative experiences span the period from 2006 to 2017 and are related to 320 plants.

Each "operative experience" tells about the events occurred in the recent past (see Figure 1 for an example). Each event is registered by the operators filling in a pre-defined form. The document contains information including the date, a title summarizing the event, a short description, the reference to failed, missing or misapplied technical or procedural barriers, those that stopped the escalation and the recovering actions, and eventually the planned actions for improving the safety.

It is out of scope of this paper to discuss the different methods used in the literature to manage near miss information for improving the safety management system. However, the common objective is to exploit the valuable information contained. (Ansaldi et al., 2018) describe a method to extract knowledge from this collection of documents, and to support foresights or intuitions about the safety of process industries. Another application has been developed for understanding if the lessons from major accidents have been fully learnt and implemented (Ansaldi et al., 2016). The issue has been addressed by looking for similarities between near misses and accident characteristics, and by evaluating their semantic distance.

Although the form of the document is the same adopted for all operators, the compiling mode

varies by the establishments and by the type of event recorded. The accuracy of the documents is not homogeneous and the interpretation of operative experience concept changes from one establishment to another; their carefulness varies on the sector activities, and often reveals the safety culture of the establishment. At a few establishments, just the releases of hazardous substance without consequences are registered. In other cases, reports include anomalies, unsafe situations, failures, and trivial errors; that is, events not directly related to major accident hazard. The documents are various, but represent truthful pictures of deviations occurred inside the establishment.

3 Methods

The overall goal is to show that existing methodologies can help in standardizing language in the specific case of reports on near misses on Seveso industries and we aim to perform this standardization with two tools: (1) analyzing similarities among words in current reports; (2) propose a methodology to help in writing these reports.

3.1 Challenges

The specific case of reports on near misses is particular for several compelling reasons. The first compelling reason is that reports are written by operators belonging to sub-communities of speakers. In fact, people working in each plant can be considered a sub-community, which shares a particular language. Hence, standardizing language of reports means also harmonize sub-languages of different sub-communities, which do not interact. This problem is particularly severe when the aim is to standardize language across the whole Seveso industries. The second compelling reason is the different background of reports' writers. Reports are in fact written by operators, which may have different knowledge, different school degree, and different cultural background. This reason makes particularly relevant the goal to help writers in compiling reports on near misses.

3.2 Enabling Tools and Methodologies

To meet the overall goal, we here experiment with standard and well-assessed models and methodologies: the notion of word embedding. In fact, the long tradition of representing word meaning in vectors is what is needed to: (1) help the standardization organism to develop a common and accept-

able language; (2) devise ways to suggest more appropriate words to writers of reports. In this study, we used two different word embeddings:

- General Language Word Embeddings (GLwe)(Cimino et al., 2018): these are word embeddings pre-trained with word2vec (Mikolov et al., 2013) on a general purpose corpus of the Italian language, that is, itWaC (Baroni et al., 2009)
- Domain-adapted Word Embeddings (Dawe): these are word embeddings obtained training word2vec (Mikolov et al., 2013) using GLwe as initialization and the REP training corpus

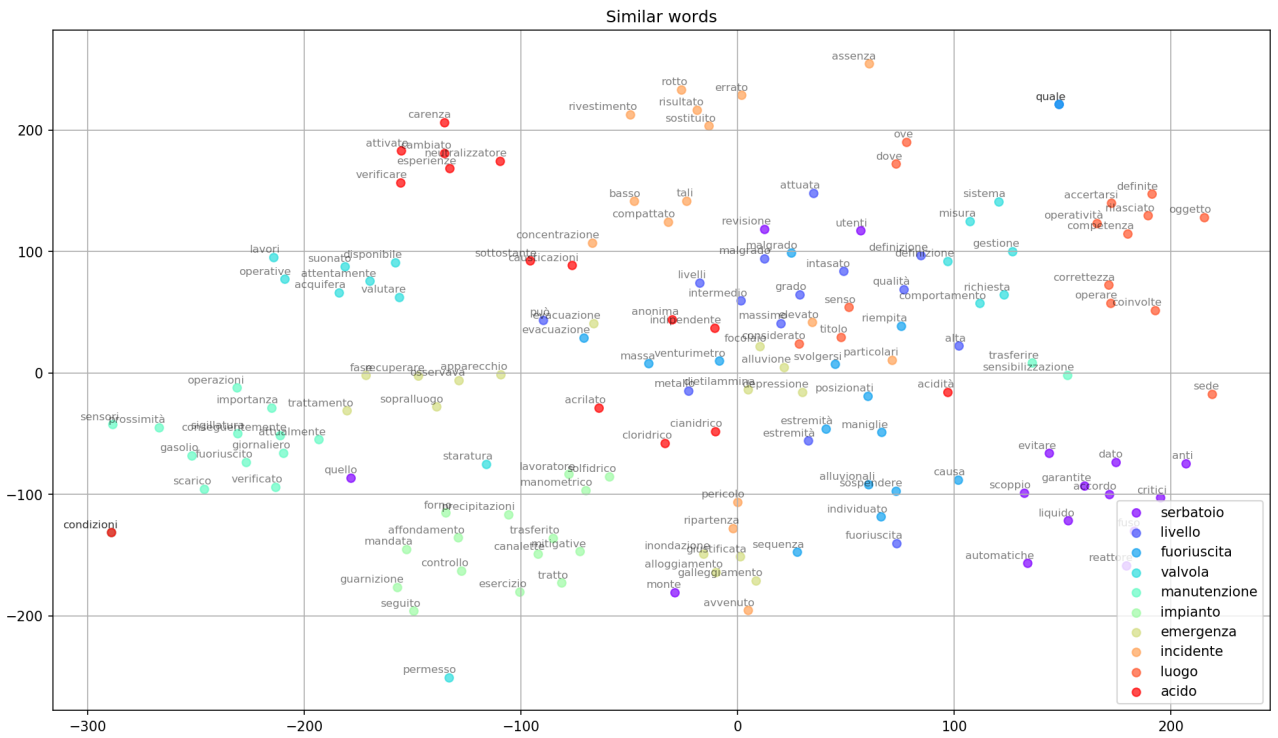
3.3 Task 1: Understanding Language of Near-Miss Reports

We aim to provide the standardization organism, that is, INAIL, the possibility to investigate the language used in these reports on near misses. The possibility we explored is to provide a visual representation of similarity computed using similarity among word embeddings. Giving this visual representation, researchers in INAIL can devise the definition of a standard language that is built on a common and shared language. This idea is similar to what has been done in the past for terminology extraction. The real added value is that similarity among terms is computed according to word embeddings.

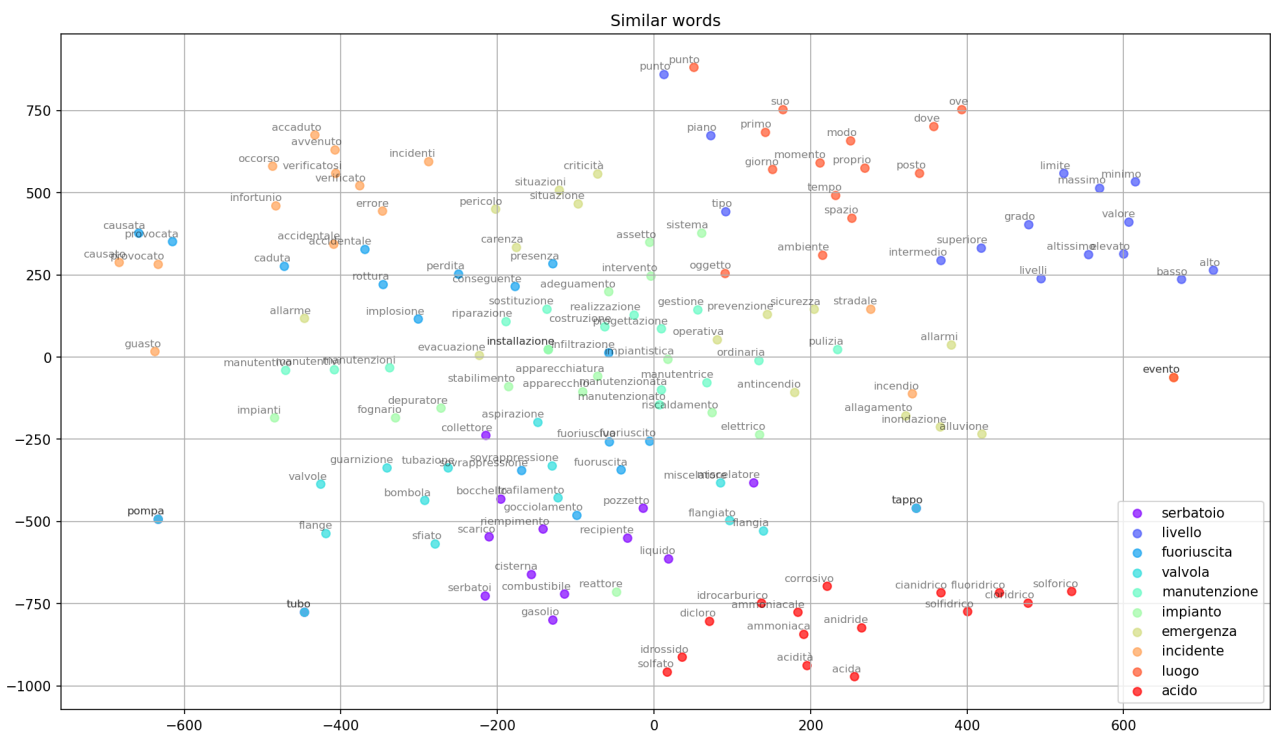
3.4 Task 2: Standardizing Report with Assisted Writing

We aim to provide a tool to assist operators while writing reports. We explored the first capability of this tool, that is, avoiding unnecessary use of synonyms while writing. In the Italian tradition, using repeating words is seen as bad writing. Hence, when writing, synonyms are used to introduce a variation. However, for technical documents, unnecessary use of synonyms in core concepts may introduce misunderstanding. Hence, we envisage a tool that helps in reducing use of synonyms.

The algorithm governing the tool works as follows. While writing a report, the algorithm accumulate words in a set W . Whenever a new content word w is added, the algorithms compute the similarity with the words in the set W . If there is a word $w' \in W$ for which the similarity $\text{sim}(w, w') = w^T w'$ is above a threshold τ , the algorithm suggests w' as a possible substitution of w . In this way, the operator is forced to



(a) General Language Word Embeddings



(b) Domain-adapted Word Embeddings

Figure 2: Similarities among Words: Studying and Understanding Technical Language with Word Embeddings

#	Text
174	La perdita non si era evidenziata al controllo dell'area effettuato preliminarmente all'inizio attività, né rilevata dal CTM presente in <i>zona</i> area (<i>sim</i> = 0.64) attività (<i>sim</i> = 0.31)
175	... Alle ore 10,15 il CT rilevava visivamente la presenza di tracce di virgin nafta miscelati con le acque di scarico e, mentre si accingeva a chiudere la valvola sul dreno di fondo colonna, improvvisamente, si sviluppava un principio d'incendio. Lo stesso CT, utilizzando le manichette di erogazione acqua già attive per il lavaggio dell'area atto a favorire il convogliamento dei reflui nel pozzetto di raccolta di raffineria, estingueva prontamente il <i>focolaio</i> incendio (<i>sim</i> = 0.46) intervento (<i>sim</i> = 0.34)
109	Necessità di prevedere un più elevato grado di protezione contro la perdita di contenimento da fondo serbatoi. La <i>fuoriuscita</i> perdita (<i>sim</i> = 0.54) contenimento (<i>sim</i> = 0.42)

Figure 3: Suggested replacements for with already used synonyms

think whether the word w' that s/he already used is similar to the word s/he is using now. In this case, w' can be used to replace w and an unnecessary synonym is avoided.

4 Experimental Results

4.1 Task 1

For the first task, we experimented with the two dictionaries: the General Language word embeddings (GLwe) and the Domain-adapted word embeddings (Dawe). Similarity spaces for the two word embeddings (see Figure 2) may help in understanding whether unnecessary synonyms are used and, hence, suggest a standardized word that should be used for a group of words.

Using the two dictionaries, we built two similarity spaces (Figure 2) obtained as follows. We selected 10 frequent words in the REP training corpus and, then, we presented in the two figures the top 15 words that are more similar to the 10 selected frequent words. The similarity spaces are built according to GLwe (Figure 2a) and according to Dawe (Figure 2b).

The Dawe similarity space (Figure 2b) gives apparently better hints on how words are used. The dictionary seems to be more tailored to the specific domain. In fact, there is an interesting groups of words such as $\{avvenuto, accaduto, occorso, verificatosi\}$ and $\{causato, provocato\}$. These groups are missing in the GLwe similarity space (Figure 2a).

4.2 Task 2

For the second task, we experimented with some sample reports. The algorithm in action is reported in Figure 3. This test has been carried out on existing reports and aimed to show that some words can be replaced with previously used words. In the report #174, the word *zona* can be replaced with the word *area*, which has been previously used. In the report #175, the word *focolaio* could be replaced with the word *incendio*. Finally, in the report #109, the word *fuoriuscita* can be replaced with the word *perdita*. However, the operator is free to accept or refuse the suggestion if this is not satisfactory.

5 Conclusion and Future Work

Standardizing language is a need of our technological society. In this paper, we investigated the possibility of using modern NLP techniques to reach this goal in the specific scenario of near misses in Seveso Industries. Initial results on the corpus provided by Inail are interesting and leave room for improvement. Future model should include the treatment of multi-word expressions by using compositional distributional semantic models (Zesch et al., 2013; Zanzotto et al., 2015), should merge distributional and ontological models, and should include a clear model for repaying knowledge producers (Zanzotto, 2019).

References

- Silvia Maria Ansaldi, Patrizia Agnello, and Paolo Bragatto. 2016. [Incidents triggered by failures of level sensors](#). *Chemical Engineering Transactions*, 53:223–228.
- Silvia Maria Ansaldi, Annalisa Pirone, Rosaria Vallerotonda Maria, Paolo Bragatto, Patrizia Agnello, and Corrado Delle Site. 2018. [How inspections outcomes may improve the foresight of operators and regulators in seveso industries](#). *Chemical Engineering Transactions*, 67:367–372.
- Marco Baroni, Silvia Bernardini, Adriano Ferraresi, and Eros Zanchetta. 2009. [The wacky wide web: a collection of very large linguistically processed web-crawled corpora](#). *Language Resources and Evaluation*, 43(3):209–226.
- Andrea Cimino, Lorenzo De Mattei, and Felice Dell’Orletta. 2018. [Multi-task learning in deep neural networks at EVALITA 2018](#). In *Proceedings of the Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2018) co-located with the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018), Turin, Italy, December 12-13, 2018*.
- Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. [Efficient estimation of word representations in vector space](#). *CoRR*, abs/1301.3781.
- Ferdinand de Saussure. 1916. *Cours de linguistique générale*. Payot, Paris.
- Fabio Massimo Zanzotto. 2019. [Viewpoint: Human-in-the-loop Artificial Intelligence](#). *Journal of Artificial Intelligence Research*, 64:243–252.
- Fabio Massimo Zanzotto, Lorenzo Ferrone, and Marco Baroni. 2015. [When the whole is not greater than the combination of its parts: A ”decompositional” look at compositional distributional semantics](#). *Comput. Linguist.*, 41(1):165–173.
- T. Zesch, I. Korkontzelos, F.M. Zanzotto, and C. Biemann. 2013. [Semeval-2013 task 5: Evaluating phrasal semantics](#). volume 2, pages 39–47. Cited By 12.

Computational Linguistics Against Hate: Hate Speech Detection and Visualization on Social Media in the “Contro L’Odio” Project

Arthur T. E. Capozzi, Mirko Lai

Valerio Basile, Fabio Poletto

Manuela Sanguinetti

Cristina Bosco

Viviana Patti

Giancarlo Ruffo

University of Turin

name.surname@unito.it

Cataldo Musto

Marco Polignano

Giovanni Semeraro

University of Bari “Aldo Moro”

name.surname@uniba.it

Marco Stranisci

ACMOS

m.stranisci@acmos.net

Abstract

The paper describes the Web platform built within the project “Contro l’odio”, for monitoring and contrasting discrimination and hate speech against immigrants in Italy. It applies a combination of computational linguistics techniques for hate speech detection and data visualization tools on data drawn from Twitter. It allows users to access a huge amount of information through interactive maps, also tuning their view, e.g., visualizing the most viral tweets and interactively reducing the inherent complexity of data. Educational courses for high school students and citizenship has been developed which are centered on the platform and focused on the deconstruction of negative stereotypes against immigrants, Roma, and religious minorities, and on the creation of positive narratives.

1 Introduction

Hate Speech (HS) is a multi-faceted phenomenon with countless nuances, a high degree of individual and cultural variation, and intersections with related concepts such as offensive language, threats, bullying and so on.

The detection of HS is a recent yet popular task that is gaining the attention of the NLP community but also that of public institutions and private companies. There are several problems connected

with this delicate task: a cultural-dependent definition, a highly subjective perception, the need to remove potentially illegal contents quickly from the Web and the connected risk to unjustly remove legal content, the partly overlapping linguistic phenomena that make it hard to identify HS. English social media texts are the most studied, but other languages, sources and textual genres are investigated as well.

“Contro l’odio”¹ is a project for countering and preventing racist discrimination and HS in Italy, in particular focused against immigrants. On the one hand, the project follows and extends the research outcomes emerged from the ‘Italian Hate Map project’ (Musto et al., 2016), whose goal was to identify the most-at-risk areas of the Italian country, that is to say, the areas where the users more frequently publish hate speech, by exploiting semantic analysis and opinion mining techniques. On the other hand, “Contro l’odio” benefits from the availability of annotated corpora for sentiment analysis, hate speech detection and related phenomena such as aggressiveness and offensiveness, to be used for training and tuning the HS detection tools (Sanguinetti et al., 2018; Poletto et al., 2017). The project brings together the competences and active participation of civil society organizations Acmos² and Vox³, and two academic research groups, respectively from the University of Bari and Turin.

This paper focuses on the technological core of the project, a Web platform that combines computational linguistics analysis with visualization techniques, in order to provide users with an inter-

Copyright © 2019 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

¹<https://controlodio.it/>

²<http://acmos.net/>

³<http://www.voxdiritti.it/>

active interface for exploring the dynamics of the discourse of hate against immigrants in Italian social media. Three typical targets of discrimination related to this topical focus are taken into account, namely Migrants, Muslims and Roma, since they exemplify discrimination based on nationality, religious beliefs and ethnicity, respectively. Since November 2018 the platform analyses daily Twitter posts and exploits temporal and geo-spatial information related to messages in order to ease the summarization of the hate detection outcome.

2 Related work

In the last few years several works contributed to the development of HS detection automatic methods, both releasing novel annotated resources, lexicons of hate words or presenting automated classifiers. Two surveys were recently published on this topic (Schmidt and Wiegand, 2017; Fortuna and Nunes, 2018). For what concerns Italian a few resources have been recently developed drawn from Twitter (Sanguinetti et al., 2018; Poletto et al., 2017) and FaceBook (Del Vigna et al., 2017), where the annotation of hateful contents also extends the simple markup of HS. A multilingual lexicon of hate words has been also developed (Bassignana et al., 2018) called Hurtlex⁴. The lexicon, originally built from 1,082 Italian hate words compiled in a manual fashion by the linguist Tullio De Mauro (De Mauro, 2016), has been semi-automatically extended and translated into 53 languages. The lexical items are divided into 17 categories such as homophobic slurs, ethnic slurs, genitalia, cognitive and physical disabilities, animals and more.

Since 2016, shared tasks on the detection of HS or related phenomena (such as abusive language or misogyny) in various languages have been organized, benefiting from the developed datasets and effectively enhancing advancements in resource building and system development both. These include in particular HatEval at SemEval 2019 (Basile et al., 2019), AMI at IberEval 2018 (Fersini et al., 2018b), HaSpeeDe and AMI at EVALITA 2018 (Bosco et al., 2018; Fersini et al., 2018a).

The project “Contro l’odio” follows and extends the research outcome emerged from the ‘Italian Hate Map project’ (Musto et al., 2016), where

⁴<http://hatespeech.di.unito.it/resources.html>

a lexicon developed within the project (Lingiarri et al., 2019) has been exploited to provide a fine-grained classification of the nature of the hate speeches posted by the users on different hate targets. In “Contro l’odio” we inherited the idea of map-based visualization to show the distribution of the hate speech, but we enhance it in two main directions: a) by creating a web platform that enables a *daily monitoring* of hate speech against immigrants in Italy and its evolution over time and space; b) by adding a level of interactivity with the results of the automatic detection of hate speech, both in terms of maps and of hate words’ inspection, which enabled interesting activities for countering hate in schools. Monitoring and countering HS is a shared goal with several recent projects, with different focuses w.r.t countries and territories monitored, targets of hate, granularity of the detection, visualization techniques provided to inspect the monitoring results. Let us mention the *CREEP* project⁵ on monitoring cyberbullying online (Menini et al., 2019), with an impact also on the Italian territory, *HateMeter*⁶, with a special focus on Anti-Muslim hatred online, the MAN-DOLA project⁷ providing a reporting infrastructure enabling the reporting of illegal hate-related speech, and the *Geography of Hate* project⁸ in the US.

3 The Contro l’odio monitoring platform

3.1 Architecture

The architecture consists of four main modules. The data collection module gathers the tweets by using the Stream Twitter API and filters them by keywords. The automatic classifier module automatically annotates the presence of HS in the filtered tweets, relying on a supervised approach. The next module stores the annotated tweet aggregating them by time and place in a database. The last module, implemented by relying on a *node.js* server, exposes the API that are requested by the front end (Figure 1).

3.2 Data Collection

We started collecting tweets from October 1st 2018 by using the Twitter’s Stream API. The

⁵<http://creep-project.eu/>

⁶<http://hatemeter.eu/>

⁷<http://mandola-project.eu/>

⁸<http://www.antiatlas.net/geography-of-hate-geotagged-hateful-tweets-in-the-united-states-en/>

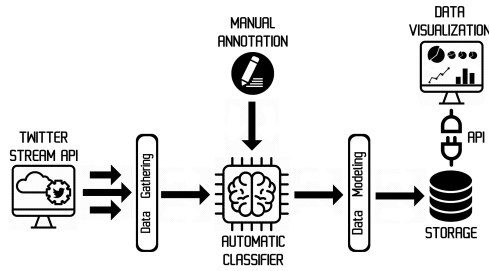


Figure 1: Architecture of the ‘Contro l’odio’ platform

streaming is filtered using the vowels as keywords and the alpha-2 code *it* as language filter. About 700,000 Italian statuses are daily gathered, but only about 17,000 are relevant for monitoring discrimination and HS against immigrants in Italy. We filtered relevant tweets by using the keywords proposed in Poletto et al. (2017), considering three typical targets of discrimination — namely migrants, Roma and religious minorities.

3.3 The Hate Detection Engine

In order to automatically label the tweets, we developed a supervised classifier to predict the presence of HS in text. The classification is binary (i.e., presence of HS vs. absence of HS). We employ a Support Vector Machine (SVM) classifier with one-hot unigram representation as feature vector. We train the classifier on the Italian Hate Speech Corpus (Sanguinetti et al., 2018, IHSC), a collection of about 6,000 tweet in the Italian language, manually annotated both by experts and crowdsourced annotators along several dimensions: hate speech, aggressiveness, offensiveness, irony, stereotype, and intensity. IHSC is particularly well suited for our scenario, since the data have been specifically collected on the topic of immigration and ethnic/religious minorities.

The following tweets are two examples of annotated tweets:

1. *#dallavostraparte non ci sono moderati, sono tutti terroristi pronti a tagliarci la testa e per questo io li odio a morte!*

#onyourside there are no moderates, they all are terrorists ready to cut our head off and for this I hate them to death!

2. *Tanto con il sole i nomadi non vengono più a scuola. Per qualcuno questa è la soluzione...*

Nomads no longer come to school when it's sunny. For some this is the solution...

In example 1, the target is “religious minorities”

and the author spreads and incites violence against Islamic people (the tweet contains hate speech). In example 2, the target is “Roma”, and the previous conditions are not detected, there’s not hate speech. By performing cross-validation experiments on such corpus, we estimate the best hyperparameters for the model: 27,642 features, learning rate *optimal*, *linear* kernel. With this settings, we record a prediction performance in cross-validation of 0.81 (0.70 for the class *hate speech*) precision and 0.81 (0.67 for the class *hate speech*) recall ($F_{avg} = 0.80 \pm 0.01$). Recently, new classification strategies base on language understanding models have been demonstrated to be suitable for the task of hate speech detection, obtaining encouraging results. As a consequence, we are considering the possibility to compare our model with a classifier base on ALBERTo (Polignano et al., 2019) as a further step for improving the performances of our hate detection engine.

It is important to note that the Italian Hate Speech Corpus has been collected in a specific time frame, from October 1st, 2016 to April 25th, 2017 (Poletto et al., 2017). The relative distribution of topics may change over time, thus we expect a performance drop when applying the model trained on IHSC to new, recent data. In order to measure this gap, we annotated 2,000 additional tweets each month for several months, collected from the Contro l’Odio pipeline (Section 3.2) and confronted the prediction of our classifier against the manual annotation. The data have been annotated in a crowdsourcing fashion, using the online platform Figure Eight⁹. The performance of the classifier trained on IHSC on the new test set, in terms of F_{avg} , degrades as the time frame moves farther from that of of IHSC: October (0.57), November (0.56), and December (0.54), 2018, and January (0.51), and February (0.47), 2019. However, we plan to reintroduce the newly annotated datasets (this experiment is currently ongoing) in the training set and re-train the model, in order to make the system more robust across time, and to keep monitoring the performance.

4 Visualizing and Interacting with Estimated Hate

4.1 Interactive Hate Maps

The main view of the dashboard is a choropleth map and allows the user to explore the spatial

⁹<https://www.figure-eight.com>

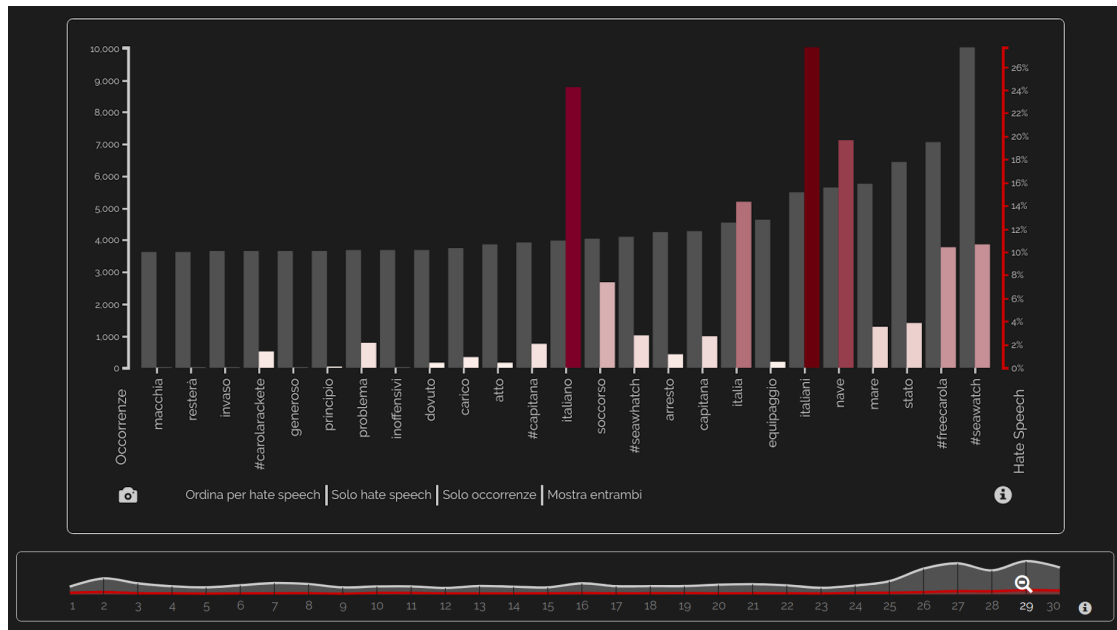


Figure 2: Word occurrences bar chart.

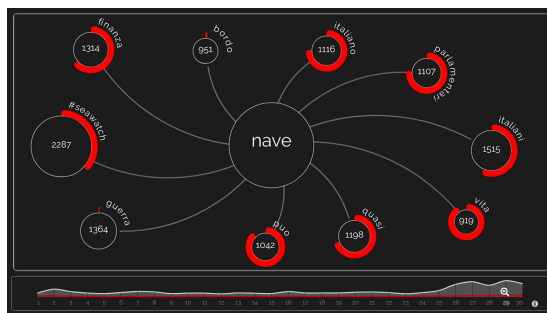


Figure 3: Word co-occurrences network.

dimension (regional and provincial level) of the dataset. The temporal dimension can be explored thanks to a time slider that also shows the trend of the total number of tweets and the percentage of HS. In figure 5 there's an example of how the choropleth map and the Dorling map appear on June 29, 2019 when the migrant and NGO themes, in a single day, become viral in the public debate¹⁰. In Figure 4 we see another example: the volume of tweets about the Roma topic in the days from 3 to 5 June 2019 has increased considerably due to some clashes in the outskirts of Rome¹¹.

¹⁰<http://www.ansa.it/sito/notizie/politica/2019/06/28/sea-watch-indagata-la-capitana.-nuovo-affondo-di-salvini-contro-lolanda-comportamento-disgustoso-991189d6-7818-48d9-b4d8-a2a7d10d31bc.html>

¹¹<http://www.ansa.it/sito/notizie/cronaca/2019/04/04/simone-il-quindicenne-di-torre-maura-contro-casapound-state-a-fa-leva-sulla-rabbia->

The liquid gauge allows the user to quickly detect the tweet volume increase, from 1,619 to 14,778, and the increase in HS rates, from 13% to 23%.

4.2 Words of Hate

Figure 2 shows another visualization: a bar chart containing the 25 words more frequently occurring in the tweets collected in the selected time period. For each word, the user can also see the average percentage of HS in tweets containing that word. As before, the example in figure 2 refers to June 29, 2019. By clicking on a word, the user can visualize additional information about it, such as the exact number of occurrences in the tweets or its co-occurrence network (figure 3).

5 Countering online hate speech in High Schools

The interactive hate maps and the 'Words of Hate' visualization settings described here have been also used within educational paths developed for citizenship and mostly targeting high school students. Such paths were focused on the dismantling of negative stereotypes against immigrants, Roma, and religious minorities, and on the creation of positive narratives to actively counteract hatred online. Since today, a team of twenty educators carried out 90 laboratories in seven different Italian regions (Piedmont, Tuscany, Liguria, Emilia

della-gente.-plauso-raggivideo_7a4bc495-bb4d-4c21-a1f7-2ecbc8422ea5.html

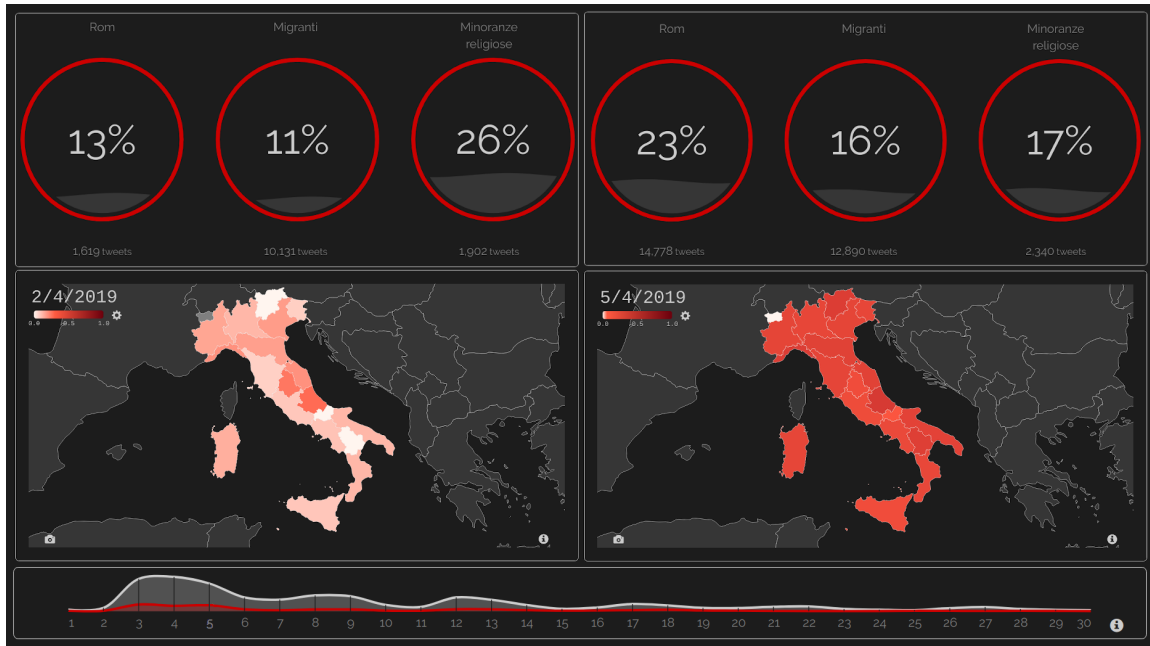


Figure 4: Choropleth map and liquid fill gauge.

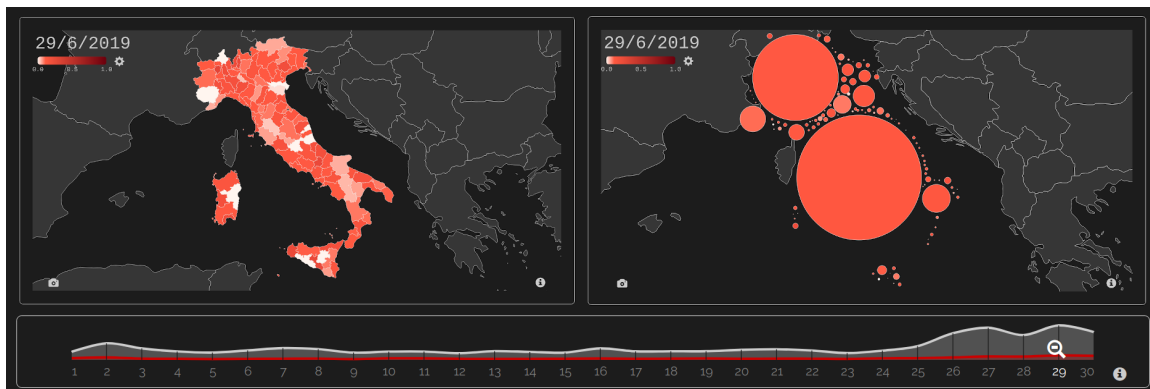


Figure 5: Choropleth map and Dorling map.

Romagna, Lazio, Friuli-Venezia Giulia, and Sardinia). At the end of the project 150 classes will be reached, and the resulting positive narratives will be published on the project website.

6 Conclusion and Future Work

In this paper we described an online platform for monitoring HS against immigrants in Italy at different levels of granularity, which uses Twitter as data source and combines HS detection and advanced visualization techniques in order to provide users with an interactive interface for the exploration of the resulted data. Another important research outcome of the project is HATE-CHECKER, a tool that automatically detects *hater users* in online social networks, which will be accessible from the platform soon. Given a target

user, the workflow that is going to be implemented in our system uses sentiment analysis techniques to identify hate speech posted by the user, and exploits a lexicon-based approach to assign to the person one or more labels that describe the nature of the hate speech she posted (e.g., racism, homophobia, sexism, etc.). A map of Italian projects and associations that spread a culture of tolerance is also under development, to allow ‘Contro l’Odio’ users to get a better understanding of the HS phenomenon and of the active forces fighting it on the Italian territory.

Acknowledgments

The work of all the authors was partially funded by Italian Ministry of Labor (*Contro l’odio: tecnologie informatiche, percorsi for-*

mativi e storytelling partecipativo per combattere l'intolleranza, avviso n.1/2017 per il finanziamento di iniziative e progetti di rilevanza nazionale ai sensi dell'art. 72 del decreto legislativo 3 luglio 2017, n. 117 - anno 2017).

References

- Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. 2019. Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in Twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 54–63. Association of Computational Linguistics.
- Elisa Bassignana, Valerio Basile, and Viviana Patti. 2018. Hurltlex: A multilingual lexicon of words to hurt. In *Proceedings of the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018)*, Torino, Italy, December 10-12, 2018., volume 2253 of *CEUR Workshop Proceedings*. CEUR-WS.org.
- Cristina Bosco, Dell'Orletta Felice, Fabio Poletto, Manuela Sanguinetti, and Tesconi Maurizio. 2018. Overview of the evalita 2018 hate speech detection task. In *Proceedings of Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2018)*, volume 2263, pages 1–9. CEUR.
- Tullio De Mauro. 2016. Le parole per ferire. *Inter-nazionale*. 27 settembre 2016.
- Fabio Del Vigna, Andrea Cimino, Felice Dell'Orletta, Marinella Petrocchi, and Maurizio Tesconi. 2017. Hate Me, Hate Me Not: Hate Speech Detection on Facebook. In *Proceedings of the First Italian Conference on Cybersecurity (ITASEC17)*.
- Elisabetta Fersini, Debora Nozza, and Paolo Rosso. 2018a. Overview of the evalita 2018 task on automatic misogyny identification (AMI). In Tommaso Caselli, Nicole Novielli, Viviana Patti, and Paolo Rosso, editors, *Proceedings of the Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2018) co-located with the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018)*, Turin, Italy, December 12-13, 2018., volume 2263 of *CEUR Workshop Proceedings*. CEUR-WS.org.
- Elisabetta Fersini, Paolo Rosso, and Maria Anzovino. 2018b. Overview of the task on automatic misogyny identification at ibereval 2018. In *Proceedings of the Third Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018) co-located with 34th Conference of the Spanish Society for Natural Language Processing (SEPLN 2018)*, Sevilla, Spain, September 18th, 2018., volume 2150 of *CEUR Workshop Proceedings*, page 214–228. CEUR-WS.org.
- Paula Fortuna and Sérgio Nunes. 2018. A survey on automatic detection of hate speech in text. *ACM Computing Surveys (CSUR)*, 51(4):85.
- Vittorio Lingiardi, Nicola Carone, Giovanni Semeraro, Cataldo Musto, Marilisa D'Amico, and Silvia Brena. 2019. Mapping twitter hate speech towards social and sexual minorities: a lexicon-based approach to semantic content analysis. *Behaviour & Information Technology*, 0(0):1–11.
- Stefano Menini, Giovanni Moretti, Michele Corazza, Elena Cabrio, Sara Tonelli, and Serena Villata. 2019. A system to monitor cyberbullying based on message classification and social network analysis. In *Proceedings of the 3rd Workshop on Abusive Language Online, co-located with ACL 2019*. Association of Computational Linguistics.
- Cataldo Musto, Giovanni Semeraro, Marco de Gemmis, and Pasquale Lops. 2016. Modeling community behavior through semantic analysis of social data: The italian hate map experience. In *Proceedings of the 2016 Conference on User Modeling Adaptation and Personalization*, pages 307–308. ACM.
- Fabio Poletto, Marco Stranisci, Manuela Sanguinetti, Viviana Patti, and Cristina Bosco. 2017. Hate speech annotation: Analysis of an italian twitter corpus. In *Proceedings of the Fourth Italian Conference on Computational Linguistics (CLiC-it 2017)*, Rome, Italy, December 11-13, 2017., volume 2006 of *CEUR Workshop Proceedings*. CEUR-WS.org.
- Marco Polignano, Pierpaolo Basile, Marco de Gemmis, Giovanni Semeraro, and Valerio Basile. 2019. Alberto: Italian bert language understanding model for nlp challenging tasks based on tweets. In *Proceedings of the Sixth Italian Conference on Computational Linguistics (CLiC-it 2019)*. CEUR.
- Manuela Sanguinetti, Fabio Poletto, Cristina Bosco, Viviana Patti, and Marco Stranisci. 2018. An Italian Twitter Corpus of Hate Speech against Immigrants. In Nicoletta Calzolari (Conference chair), Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Koiti Hasida, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, Stelios Piperidis, and Takenobu Tokunaga, editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May 7-12, 2018. European Language Resources Association (ELRA).
- Anna Schmidt and Michael Wiegand. 2017. A Survey on Hate Speech Detection using Natural Language Processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*. Association for Computational Linguistics.

Supporting Journalism by Combining Neural Language Generation and Knowledge Graphs

Marco Cremaschi, Federico Bianchi, Andrea Maurino, Andrea Primo Pierotti

Department of Computer Sciences, Systems and Communications

University of Milan-Bicocca

Viale Sarca, 336 - 20126, Milan, Italy

{marco.cremaschi, federico.bianchi, andrea.maurino}@unimib.it

a.pierotttil@campus.unimib.it

Abstract

Natural Language Generation is a field that is becoming relevant in several domains, including journalism. Natural Language Generation techniques can be of great help to journalists, allowing a substantial reduction in the time required to complete repetitive tasks. In this position paper, we enforce the idea that automated tools can reduce the effort required to journalist when writing articles; at the same time we introduce GazelLex (Gazette Lexicalization), a prototype that covers several steps of Natural Language Generation, in order to create soccer articles automatically, using data from Knowledge Graphs, leaving journalists the possibility of refining and editing articles with additional information. We shall present our first results and current limits of the approach, and we shall also describe some lessons learned that might be useful to readers that want to explore this field.

1 Introduction

Although automation is a phenomenon that is becoming more and more visible today, there are specialised jobs that require human effort to be completed. The job of a journalist is among these (Örnebring, 2010). However, recent technological progress in the field of Natural Language Generation (NLG) and the use of increasingly sophisticated techniques of artificial intelligence allow the use of software capable of writing newspaper articles almost indistinguishable from human ones. These techniques can help journalists reduce the

effort needed for repetitive tasks, such as data collection and drafting writing. The name given to this phenomenon is Automated Journalism; this new type of journalism uses algorithms to generate news under human supervision. During the past years, several newsrooms have begun to experiment this technology: Associated Press, Forbes, Los Angeles Times, and ProPublica are among the first, but adoption could spread out soon (Graefe, 2016). Automated Journalism can bring a massive change to the sector: writing news is a business that endeavours to minimise costs while maintaining maximum efficiency and full speed, and thanks to this software the above-mentioned objectives can be achieved, generating good-quality articles (van Dalen, 2012). This new technology provides many advantages: the most evident are speed and the scale of news coverage. Of course, there are also problems and limitations. One of the most relevant is the dependence from structured data (Graefe, 2016), that is the reason why sports reports, financial articles, and forecasts are the most covered topics by software: they are all domains where the complexity of the topic can be managed from software using structured data. Similar structured data are not always available in other fields. In order to generate valuable text, approaches considering data contained in the Knowledge Graphs (KGs) have recently been introduced in literature (Gardent et al., 2017; Trisedya et al., 2018).

A Knowledge Graph (KG) describes real-world entities and the relations between them. KGs are an essential source of information, and their features allow the use of this information in different contexts, such as link prediction (Trouillon et al., 2016) and recommendation (Zhang et al., 2016). Popular KGs are the Google Knowledge Graph, Wikidata and DBpedia (Auer et al., 2007). Entities are defined in an ontology and thus can be classified using a series of types. The primary element of a KG to store entities information is a

Copyright 2019 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

Resource Description Framework (RDF) triple in the format $\langle \text{subject}, \text{predicate}, \text{object} \rangle$. As RDF triples open many possibilities in Web data representation, utilising this data also in the NLG context is valuable (Perera et al., 2016). Interlinked KGs can be used to automatically extend the information relating to a given entity in an article.

In our solution, we use DBpedia, one of the fastest growing Linked Data resource that is available free of charge; it is characterised by a high number of links from the Linked Data Cloud². DBpedia is thus a central interlinking hub, an access point for retrieving information to be inserted in an article, as specified below.

Up to 2010, commercial providers in the NLG field were not popular, but in the last years few companies have started to provide this kind of services. In 2016 there were 13 companies covering this field (Drr, 2016) (e.g., AutomatedInsights³, NarrativeScience⁴). Approaches that try to integrate deep networks and text generation are now common in literature (Gardent et al., 2017). These automated tools are going to become a standard method to help journalist during the news writing process.

We shall concentrate on examples of related work in the context of lexicalization from RDF data, we shall refer to surveys from the state of the art for a more detailed overview of the field (Reiter and Dale, 1997; Gatt and Krahmer, 2018; Moussallem et al., 2018). Semantic web technologies like RDF can be used to enhance the power of current algorithms (Bouayad-Agha et al., 2012). The WebNLG challenge (Gardent et al., 2017) has been introduced to study the possibilities given by the combination of deep learning techniques and semantic web technologies. In a similar context, an approach based on Long Short-Term Memory (LSTM) networks has been proposed to generate text lexicalizations from RDF triples (Trisedya et al., 2018).

In this work, we aim to describe what is the possible automation process that can be used to help journalist in the news writing process. At the same time we describe a new prototype we have created to support journalistic activities, GazelLex (Gazette Lexicalization). GazelLex, through the use of deep learning techniques implements a

Neural Machine Translation (NMT) approach to generate articles (sentences) starting from data composed by RDF triples. GazelLex is also able to generate videos containing the images and the prominent information of the article, and to generate audio using a speech synthesis module (Figure 1). To the best of our knowledge, our prototype is the first to provide an all-in-one integrated approach to NLG with RDF triples in the context of helping journalist in writing articles.

This paper is structured as follows: in Section 2, we analyse the state-of-the-art on Natural Language Generation, showing that these methods to generate natural language are becoming popular. In Section 3 we describe our prototype, GazelLex, that combines neural methods and knowledge graphs to create soccer articles and describe how this kind of tools can be of help to journalism. In Section 4 we show a preliminary experimental analysis, while in Section 5 we provide conclusions.

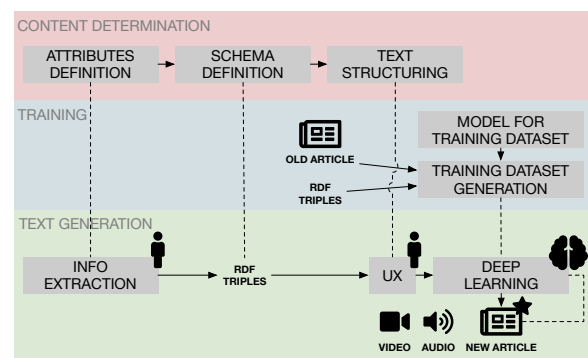


Figure 1: The workflow of our model.

2 Natural Language Generation

NLG is a “sub-field of artificial intelligence and computational linguistics that is concerned with the construction of computer systems that can produce understandable texts in English or other human languages from some underlying non-linguistic representation of information” (Reiter and Dale, 1997; Reiter and Dale, 2000). In NLG six “problems” must be addressed: **Content determination**: input data that is always more detailed and richer than what we want to cover in the text (Gatt and Krahmer, 2018) and so the aim is to filter and choose what to say. **Text structuring**: a clear text structure and the order of presentation of information are critical for readers, for this reason, pre-defining the templates is necessary. **Sen-**

²<https://wiki.dbpedia.org/dbpedia-2016-04-statistics>

³<https://automatedinsights.com/>

⁴<https://narrativescience.com/>

tence aggregation: sentences must not be disconnected. Text needs therefore to be grouped in such a way that a “more fluid and readable” text (Gatt and Krahmer, 2018) is generated. **Lexicalization:** one of the most critical phases of NLG process is how to express message blocks through words and phrases. This task is called lexicalization and concerns the actual conversion from messages to natural language. **Reference expression generation:** to avoid repetitions, selecting ways to refer to entities using different methods (such as pronouns, proper nouns, or descriptions) is essential. **Linguistic realisation:** it concerns the combination of relevant words and phrases to form a sentence.

As we stated above, lexicalization is one of the most critical and complex tasks in the NLG process. Natural language vagueness and choosing the right words to express a concept are intricate issues to manage. Looking at the state-of-the-art, we see that recent research on this topic shows that an interesting solution in these cases is based on Machine Learning (ML) (Gatt and Krahmer, 2018). Moreover, a recent challenge in the NLG field, launched and published in 2017, called WebNLG (Gardent et al., 2017) confirms the idea that not only we need to combine ML methods to generate language, but we can also use KGs to enrich sentences with additional contextual information (e.g., contextual information about a player).

3 GazelLex

In this section, we shall give an example of the NLG process in a domain specific view. As introduced previously, we developed a software, named GazelLex, that can produce soccer articles. There are two main reasons for this choice: first of all, the project was partly commissioned by an Italian newspaper publisher. Furthermore, soccer and sports, in general, are good domains to develop NLG, because they are complex enough to be challenging, yet they are easy to manage and many data exist (Barzilay and Lapata, 2005). In this scenario we focused our attention on the final output, using a solution that combines neural network with some handcrafted processes. We would like to underline that the data related to the games (e.g. number of goals, training) are extracted automatically from online services.

Our approach is divided into five tasks, in order to address the five classic NLG sub-problems (Gatt and Krahmer, 2018): in the following, for

each phase, implementation details will be provided.

3.1 Content Determination

To select the most relevant information, a handcrafted approach was chosen. To select the information to bring in the final output, we traced the most used data in soccer articles. One of the primary references was PASS, a personalised automated text system developed to write soccer articles (van der Lee et al., 2017). We took the kind of information PASS used to fill its templates and enriched them with our data fields. So we have some entities of type “TEAM”, “FORMATION”, “COACH” and some predicates like “injuryAt”, “yellowCardAt”, and “violentFoulAt”⁵. The software used this data to create triples, that algorithms used to write the article.

3.2 Text Structuring

Being a domain specific process, we developed a handcrafted template, based on real articles. Aiming to get a similar output we imitated the journalist’s job in the division of text and about information contained in each part. We also considered the text structuring approach usually developed in this domain, that uses more general information and after that a chronological order (Gatt and Krahmer, 2018). In GazelLex, it is possible to find templates (e.g., complete or short article) resulting from the process described above, but it is also possible to modify them or create new ones (Fig. 2).

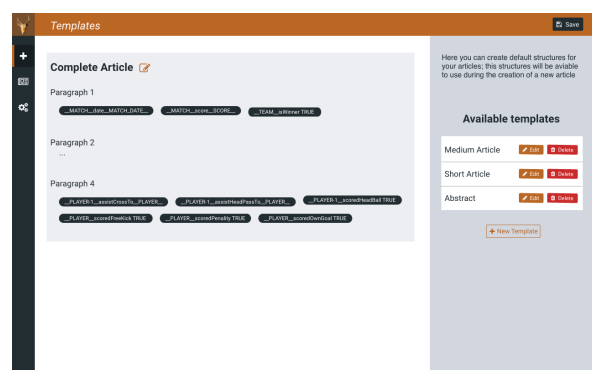


Figure 2: Page for creating and editing templates related to Text Structuring.

⁵A list of all the entities and the attributes are available here: <https://goo.gl/LonnQ5>

3.3 Sentence aggregation

In soccer data, many events could be redundant when written in an article. If a player scores a hat trick in a match writing the same sentence about each goal would be unpleasant to read while grouping them in a single sentence could be more concise and coherent. This task “focused on domain- and application-specific rules” (Gatt and Krahmer, 2018). We aggregated the RDF triples defined in the preceding section to generate a group of triples that represents the content of our news article.

3.4 Neural Lexicalization

Like we said above, we considered lexicalization like a NMT process, converting RDF data into natural language. To achieve this aim, we used a specific kind of neural network: LSTM (Hochreiter and Schmidhuber, 1997). Their recent success in NLG field is related to many advantages they provide. Compared to the traditional neural network, LSTM do not have limitations in input and output length. Furthermore, input and output are not independent, that is a vital advantage in language generation. To predict a word in a sentence it is useful to know and consider the previous one, and the hidden states of the network keep the memory about what happened in previous timesteps. In this way, LSTM can combine the previous state, the memory collected and the input, allowing dependencies to be maintained in the long term. We experimented NMT using a now widely recognized tool for neural machine translation⁶ (Klein et al., 2017). Our neural architecture is based on a standard encoder-decoder structure with 4 LSTM layers containing 200 hidden neurons on both the encoder and the decoder. Input tokenization is based on the space character (recall that our RDF triples’ elements are separated by spaces).

3.5 Reference expression generation

We used different databases to avoid redundancy and give a fluent text to the reader. Some online resources help us to create a list of possible replacements for a team or players’ name. Using DBpedia, we can find a nickname for an entity (Real Madrid players are also called *Blancos* or *Merengues*). Other resources we used are Wikidata list of soccer teams nicknames and Topend Sports database.

⁶<http://opennmt.net>

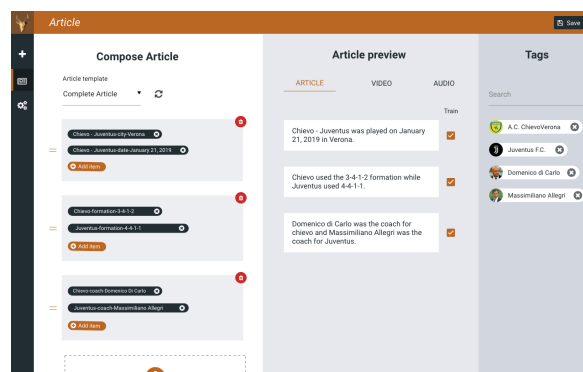


Figure 3: Page for the revision of the lexicalized triples.

4 Analysis

In the following section, we shall show some insights into our tool and on how it works. We shall present a use case, a recent soccer match, for which the generation process and the resulting text will be shown. The initial dataset for the training was created manually and consists of 4387 pairs of triples and lexicalizations. We drew inspiration from the state-of-the-art to devise the architecture of our network (Gardent et al., 2017; Trisedya et al., 2018). From our primary experiments the best performing model required two layers of bidirectional LSTM, but still, the model suffers from some limitations (outlined in the related sec.).

4.1 Use Case Exploration

To show the valid output of GazelLex, we took an example match and generated its lexicalization. We considered the football match played by Juventus F.C. and A.C Chievo on the 21st of January. Our application gathered data from an online provider and converted data in a triple format. A journalist can edit settings using a form (Figure 3): the journalist is in charge of deciding what is worth writing in the article and how it should appear to the end-user; we recall that we can also define templates for our articles (Figure 2). The final output of this process looks like the one that is shown in Figure 4. GazelLex, in order to improve the quality of the sentences and to obtain results as close to the style of the journalist as possible (i.e. style transfer), cyclically re-executes the training phase using the sentences validated by the journalist. The following is an example of lexicalization of triples relative to the use case (Table 1).

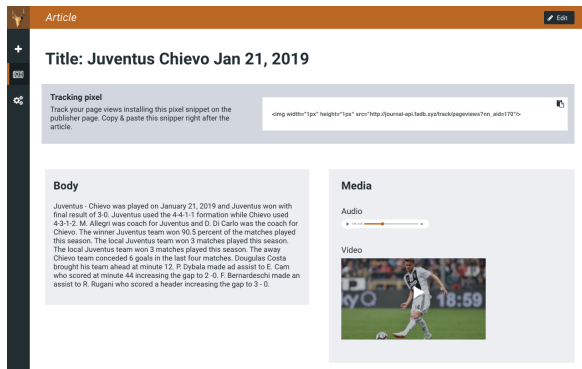


Figure 4: Lexicalization of triples from the Juventus-Chivevo football match.

Table 1: Example of lexicalization.

<i>Paulo Dybala, assistTo, Emre Can</i>
<i>Emre Can, scoredAt, 44</i>
<i>Emre Can, scoredWithScore, 0 - 2</i>
Paulo Dybala made an assist to Emre Can who scored at minute 44 increasing the gap to 0 - 2.

4.2 Current Limitations and Lessons Learned

In this section we would like to outline the current limitations of our project and also report a few lessons learned that might be useful for other researchers who are currently exploring this field. One key part of the development process comes from the definition or the selection of a good Knowledge Graph that can support the lexicalization; moreover, the definition of the new RDF predicates is a difficult process that must be done carefully to avoid errors in the next steps. Our application currently supports the lexicalization of a small set of triples (i.e., we focused on goals and final result); we decided to concentrate on this small set to generate a set of resulting sentences that can be manually inspected for quality. Our NLG model is based on a deep learning architecture, and thus some of the generated sentences are not well-formed owing to the structure of the net itself. While this is a problem that has to be solved in our settings, we have a journalist reviewing the article before it is released to the public: this allows us to have a model that is more flexible than standard pattern-based NLG, while the precision of the output can be controlled in a human-in-the-loop setting. Regarding the configuration of our model, we have replicated the state-of-the-art ex-

periments (i.e. approaches explained in (Gardent et al., 2017)) and we are currently experimenting those architectures on our domain dataset. The results are yet to be quantitatively validated and they are preliminary, but they are promising as reported by journalists. In the future, we are planning to carefully explore various architecture and consider the use of word embeddings to solve some of our current issues.

5 Conclusion

In this position paper we have analysed the future possibilities given by automated journalism. We have summarised the current state of art on this topic showing that there is an increasing interest towards automated natural language generation for the news sector. While hereby, we showed an application related to the soccer domain, the principles and the methodologies described are general, and they can be used in other fields (e.g., finance, weather reporting). We strongly believe that these tools can greatly help journalists in working on what is really important (e.g., investigation, fact checking), leaving high effort, but low value tasks to computers. The prototype we have described is a first step towards this automated process and its results are surely promising.

Acknowledgments

This research has been supported in part by the Robo-Journalism project 2018-comm25-0047 in collaboration with Instal S.r.l.⁷. Special thanks to Carlo Mattioli and Alessandra Siano for their support during the development of the project.

References

- [Auer et al.2007] Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. 2007. Dbpedia: A nucleus for a web of open data. In *The Semantic Web*, pages 722–735, Berlin, Heidelberg. Springer Berlin Heidelberg.
- [Barzilay and Lapata2005] Regina Barzilay and Mirella Lapata. 2005. Collective content selection for concept-to-text generation. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing, HLT '05*, pages 331–338, Stroudsburg, PA, USA. Association for Computational Linguistics.

⁷instal.com

- [Bouayad-Agha et al.2012] N Bouayad-Agha, G Casamayor, and L Wanner. 2012. Natural language generation and semantic web technologies. *Semantic Web Journal*.
- [Drr2016] Konstantin Nicholas Drr. 2016. Mapping the field of algorithmic journalism. *Digital Journalism*, 4(6):700–722.
- [Gardent et al.2017] Claire Gardent, Anastasia Shimorina, Shashi Narayan, and Laura Perez-Beltrachini. 2017. The webnlg challenge: Generating text from rdf data. In *Proceedings of the 10th International Conference on Natural Language Generation*, pages 124–133. Association for Computational Linguistics.
- [Gatt and Krahmer2018] Albert Gatt and Emiel Krahmer. 2018. Survey of the state of the art in natural language generation: Core tasks, applications and evaluation. *J. Artif. Int. Res.*, 61(1):65–170, January.
- [Graefe2016] Andreas Graefe. 2016. Guide to automated journalism. Technical report, Tow Center for Digital Journalism, Columbia University.
- [Hochreiter and Schmidhuber1997] Sepp Hochreiter and Jrgen Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9(8):1735–1780.
- [Klein et al.2017] Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander Rush. 2017. OpenNMT: Open-source toolkit for neural machine translation. In *Proceedings of ACL 2017, System Demonstrations*, pages 67–72, Vancouver, Canada, July. Association for Computational Linguistics.
- [Moussallem et al.2018] Diego Moussallem, Matthias Wauer, and Axel-Cyrille Ngonga Ngomo. 2018. Machine translation using semantic web technologies: A survey. *Journal of Web Semantics*, 51:1 – 19.
- [Örnebring2010] Henrik Örnebring. 2010. Technology and journalism-as-labour: Historical perspectives. *Journalism*, 11(1):57–74.
- [Perera et al.2016] Rivindu Perera, Parma Nand, and Gisela Klette. 2016. Realtext-lex: A lexicalization framework for rdf triples. *The Prague Bulletin of Mathematical Linguistics*, 106(1):45 – 68.
- [Reiter and Dale1997] Ehud Reiter and Robert Dale. 1997. Building applied natural language generation systems. *Nat. Lang. Eng.*, 3(1):57–87, March.
- [Reiter and Dale2000] Ehud Reiter and Robert Dale. 2000. *Building Natural Language Generation Systems*. Cambridge University Press, New York, NY, USA.
- [Trisedya et al.2018] Bayu Distiawan Trisedya, Jianzhong Qi, Rui Zhang, and Wei Wang. 2018. Gtr-lstm: A triple encoder for sentence generation from rdf data. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1627–1637.
- [Trouillon et al.2016] Théo Trouillon, Johannes Welbl, Sebastian Riedel, Éric Gaussier, and Guillaume Bouchard. 2016. Complex embeddings for simple link prediction. In *International Conference on Machine Learning*, pages 2071–2080.
- [van Dalen2012] Arjen van Dalen. 2012. The algorithms behind the headlines. *Journalism Practice*, 6(5-6):648–658.
- [van der Lee et al.2017] Chris van der Lee, Emiel Krahmer, and Sander Wubben. 2017. PASS: A Dutch data-to-text system for soccer, targeted towards specific audiences. In *Proceedings of the 10th International Conference on Natural Language Generation*, pages 95–104. Association for Computational Linguistics.
- [Zhang et al.2016] Fuzheng Zhang, Nicholas Jing Yuan, Defu Lian, Xing Xie, and Wei-Ying Ma. 2016. Collaborative knowledge base embedding for recommender systems. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 353–362. ACM.

Deep Bidirectional Transformers for Italian Question Answering

Danilo Croce and Giorgio Brandi and Roberto Basili

Department Of Enterprise Engineering

University of Roma, Tor Vergata

Via del Politecnico 1, 00133 Roma

{croce,basili}@info.uniroma2.it *

Abstract

English. Deep learning continues to achieve state-of-the-art results in several NLP tasks, such as Question Answering (QA). Unfortunately, the requirements of neural QA systems are very strict in the size of the involved training datasets. Recent works show that the application of Automatic Machine Translation is an enabling factor for the acquisition of large scale QA training sets in resource poor languages such as Italian. In this work, we show how these resources can be used to train a state-of-the-art deep architecture, based on effective techniques recently proposed within the Bidirectional Encoder Representations from Transformers (BERT) paradigm.

Italiano. *I recenti studi sull'applicazione di metodi di Deep Learning hanno portato a risultati importanti rispetto a diversi problemi di Natural Language Processing, come il Question Answering (QA) task. Sfortunatamente, i requisiti di tali sistemi di QA neurali sono molto stringenti per quanto riguarda le dimensioni dei dataset necessari per addestrare i modelli più complessi. Tuttavia, recenti lavori hanno dimostrato che è possibile applicare tecniche di traduzione automatica al fine di acquisire collezioni di esempi di larga scala e addestrare architetture neurali per il Question Answering nelle lingue in cui i dati di training sono scarsi, come l'italiano. In questo lavoro, mostriamo come queste risorse permettono l'addestramento di una architettura neurale molto efficace, basata sul*

paradigma noto come Bidirectional Encoder Representations from Transformers (BERT), con risultati che costituiscono lo stato dell'arte.

1 Introduction

Question Answering (QA) ((Hirschman and Gaizauskas, 2001)) tackles the problem of returning one or more answers to a question posed by a user in natural language, using as source a large knowledge base or, even more often, a large scale text collection: in this setting, the answers correspond to sentences (or their fragments) stored in the text collection. A typical QA process consists of three main steps: the question processing that aims at extracting requirements and objectives of the user's query, the retrieval phase where documents and sentences that include the answers are retrieved from the text collection and the answer extraction phase that locates the answer within the candidate sentences (Harabagiu et al., 2000; Kwok et al., 2001).

Various QA architectures have been proposed so far. Some of these rely on structured resources, such as Freebase, while others use unstructured information from sources such as Wikipedia (an example of such a system is the Microsoft's AskMSR (Brill et al., 2002)), or generic Web pages, e.g. the QuASE system (Sun et al., 2015). Hybrid models exist as well, that make use of both the structured and the unstructured information. These include IBM's DeepQA (Ferrucci et al., 2010) and YodaQA (Baudiš and Šedivý, 2015).

In order to initialize such systems, a manually constructed and annotated dataset is crucial, from which the mapping between questions and answers can be learned. Datasets designed for structured-knowledge based systems, such as WebQuestions (Berant et al., 2013), usually contain the questions, their logical forms and the answers. On the other side, datasets over unstructured information are usually composed of question-answer

*"Copyright © 2019 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0)."

pairs: WikiMovies (Miller et al., 2016) is an example of this class of systems and it is made of a collection of texts from the movie domain. Finally, some datasets contain the entire triplets made of the questions, the paragraphs and the answers, that are expressed as specific spans of the paragraph and thus located in the paragraph. This is the case of the recently proposed SQuAD dataset (Rajpurkar et al., 2016).

State-of-the-art approaches proposed in literature (Chen et al., 2017; Seo et al., 2017; Clark and Gardner, 2018; Peters et al., 2018) are based on neural paradigms and are often portable across different languages. Among them, the neural approach presented in (Devlin et al., 2019), beside achieving state-of-the-art results in several NLP tasks, is shown competitive in QA even with respect to human annotators.

Unfortunately, the limited availability of training data for languages different from English still remains an important problem. Even though multilingual data collections, such as Wikipedia, do exist for many languages, the portability of the corresponding annotated resources for supervised learning algorithms remains limited: large-scale annotated data mostly exist only for the English language.

Recent works show that the application of Automatic Machine Translation enables the acquisition of large corpora for QA in resource poor languages such as Italian (Croce et al., 2018; Croce et al., 2019). As a result, SQuAD-IT, i.e., a large scale dataset made of about 50,000 questions/answer pairs has been made available. It was not fully manually validated but still represents a valuable resource for training neural approaches.

In this work, we show how these resources enable the training of a recent and promising deep neural architecture, based on the effective techniques recently justified within the Bidirectional Encoder Representations from Transformers (BERT) paradigm (Vaswani et al., 2017; Devlin et al., 2019). The experimental evaluation carried out with respect to SQuAD-IT confirm the impressive results of BERT even in Italian QA, providing state-of-the-art results which are far higher with respect to previous methods.

In the rest of the paper, section 2 introduces the BERT architecture for QA. Section 3 reports the experimental evaluation, while Section 4 draws some conclusions.

2 Bidirectional Encoder Representations for QA

In the field of computer vision, researchers have repeatedly shown the beneficial contribution of transfer learning, i.e., the pre-training a neural network model on a known task, for instance image classification with respect to the ImageNet dataset, and then performing fine-tuning using the trained neural network as the basis of a new purpose-specific model, e.g., (Girshick et al., 2013).

The approach proposed in (Devlin et al., 2019), namely Bidirectional Encoder Representations from Transformers (BERT) provides a very effective model to pre-train a deep and complex neural network over very large scale of unannotated texts and to apply it to a large variety of NLP task by simply extending it to each new problem by fine-tuning the entire architecture.

The building block of BERT is the *Transformer* element, an attention-based mechanism that learns contextual relations between words (or sub-words, i.e. word pieces, (Schuster and Nakajima, 2012)) in a text. In its original form, proposed in (Vaswani et al., 2017), Transformer includes two separate mechanisms, an encoder that reads the text input and a decoder that produces a prediction for the targeted Machine Translation tasks.

In line with (Peters et al., 2018), BERT aims at providing a sentence embedding (as well as the contextualized embeddings of each word composing the sentence) where the pre-training stage aims at acquiring an expressive and robust language model, where only the encoder is used. As shown in Figure 1 (on the left) the Transformer encoder reads the entire sequence of words at once and acquire a language model by reconstructing the original sentence applying a MLM (*masked language model*) pre-training objective: the MLM randomly masks some of the tokens from the input, and the objective is to predict the original masked word based only on its context. In addition to the masked language model, BERT also uses a *next sentence prediction* task that jointly pre-trains text-pair representations. This last objective is crucial to improve the network capability of modeling relational information between text pairs, which is particularly important in tasks such as QA in order to relate an answer to a question.

After the language model is trained over a generic document collection, the BERT architecture allows encoding (i) specific words belong-

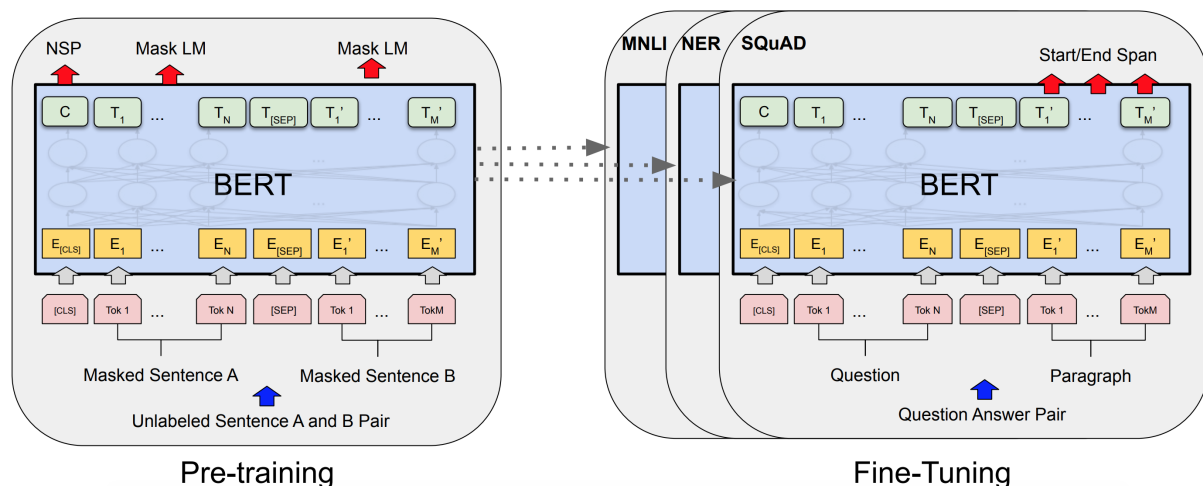


Figure 1: Overall pre-training and fine-tuning procedures for BERT. Apart from output layers, the same architectures are used in both pre-training and fine-tuning. The same pre-trained model parameters are used to initialize models for different down-stream tasks. During fine-tuning, all parameters are fine-tuned. [CLS] is a special symbol added in front of every input example, and [SEP] is a special separator token (e.g. separating questions/answers).

ing to a sentence, (ii) the entire sentence and (iii) sentence pairs with dedicated embeddings. These can be used in input to further deep architectures to solve sentence classification, sequence labeling or relational learning tasks by simply adding simple layers and fine-tuning the entire architecture. On top of such embeddings, *fine-tuning* is applied by adding task specific and simple layers on top of the architecture acquiring the language model. In a nutshell, this layer introduces minimal task-specific parameters, and is trained on the targeted tasks by simply fine-tuning all pre-trained parameters, optimizing the performance on the specific problem. The straightforward application of BERT has shown better results than previous state-of-the-art models on a wide spectrum of natural language processing tasks.

One of the most impressive results was achieved with respect to the Question Answering task proposed by (Rajpurkar et al., 2016): given a question and a passage from Wikipedia containing the answer, the task is to predict the answer text span in the passage. An example of paragraph, showing the Wikipedia answer to the question “*What was Marie Curie the first female recipient of?*” is reported in Figure 2. This specific task originated the Stanford Question Answering Dataset (SQuAD), a collection of 100k crowd-sourced question/answer pairs.

The fine-tuning process of BERT in the QA task

(shown on the right side of Figure 1) requires to encode the input question and passage as a generic text pair, such as the ones used for the next sentence prediction task used in the initial training stages.

In order to determine the correct span for the answer, (Devlin et al., 2019) introduces on top of embeddings encoding the words of the question/answer pairs a so-called *start vector* $S \in \mathcal{R}^H$ (with H the dimensionality of the embedding produced for each wordpiece T_i) and an *end vector* $S \in \mathcal{R}^H$. Then, the probability of word i being the start of the answer span is computed as a dot product between the associated embedding T_i and S followed by a softmax layer over all of the words in the paragraph: $P_i = \frac{e^{S \cdot T_i}}{\sum_j e^{S \cdot T_j}}$. The analogous formula is used for the end of the answer span. The score of a candidate span from position i to position j is defined as $S \cdot T_i + E \cdot T_j$, and the maximum scoring span where $j \geq i$ is used as a prediction. The training objective is the sum of the log-likelihoods of the correct start and end positions. The above fine-tuning of BERT achieved state-of-the-art results over the official benchmarking campaign related to SQuAD and, most noticeably, its accuracy is comparable to the one observed in human annotators¹.

It is worth noting that no bias over the input lan-

¹<https://rajpurkar.github.io/SQuAD-explorer/>

QUESTION: *What was Marie Curie the first female recipient of?*

WIKIPEDIA PARAGRAPH: One of the most famous people born in Warsaw was [Maria Skłodowska-Curie](#), who achieved international recognition for her research on radioactivity and was the first female recipient of the [Nobel Prize](#).^[198] Famous musicians include [Władysław Szpilman](#) and [Frédéric Chopin](#). Though Chopin was born in the village of [Żelazowa Wola](#), about 60 km (37 mi) from Warsaw, he moved to the city with his family when he was seven months old.^[199] [Casimir Pulaski](#), a Polish general and hero of the [American Revolutionary War](#), was born here in 1745.^[200]

GROUND TRUTH ANSWER: *Nobel Prize*

Figure 2: An example of the SQuAD dataset (Rajpurkar et al., 2016).

Element	Training set			Test set		
	English	Italian	Percent.	English	Italian	Percent.
Paragraphs	18,896	18,506	97.9%	2,067	2,010	97.2%
Questions	87,599	54,159	61.8%	10,570	7,609	72.0%
Answers	87,599	54,159	61.8%	34,726	21,489	61.9%

Table 1: The quantities of the elements of the final dataset obtained by translating the SQuAD dataset, with the percentage of material w.r.t the original dataset. The Italian test set was obtained from the English development set, being the English test set not available publicly.

	DrQA-IT	BERT-IT
EM	56.1	64.96
F1	65.9	75.95

Table 2: Results of BERT-IT over the SQuAD-IT dataset

guage exists, so that the language model underlying BERT can be acquired over any text collection independently from the input language. As a consequence a pre-trained model acquired over documents written in more than one hundred languages exists. It will be applied in the next section to train and evaluate such a QA model over a dataset of examples in Italian.

3 Experimental Evaluation

In order to assess the applicability of the BERT architecture against the targeted QA task, a multi-lingual pre-trained model has been downloaded²: in particular, this model has been acquired over documents written in one hundred languages, it is composed of 12 layers of Transformers and associates each token in input to a word embedding made of 768 dimensions. For consistency with (Devlin et al., 2019), 5 epochs have been considered to fine-tune the model.

We trained the architecture over SQuAD-IT³,

²https://storage.googleapis.com/bert_models/2018_11_23/multi_cased_L-12_H-768_A-12.zip

³<https://github.com/crux82/squad-it>

a dataset made available by (Croce et al., 2019). This dataset includes more than 50,000 question/paragraph pairs obtained by automatic translating the original SQuAD dataset. The details about the number of sentences is reported in Table 1 where a comparison with the original SQuAD in English is reported.

The parameters of the neural network were set equal to those of the original work, including the word embeddings resource. Two evaluation metrics are used: exact string match (EM) and the F1 score, which measures the weighted average of precision and recall at the token level. EM is a stricter measure evaluated as the percentage of answers perfectly retrieved by the systems, i.e. the text extracted by the span produced by the system is exactly the same as the gold-standard. The adopted token-based F1 score smooths this constraint by measuring the overlap (the number of shared tokens) between the provided answers and the gold standard.

Performances are reported in Table 2 together with the results achieved by a variant of the DrQA system (Chen et al., 2017), evaluated against the same SQuAD-IT dataset, as from (Croce et al., 2019). Improvements are impressive, as both EM and F1 are improved of more than 10%. Anyway, these results are in line with the impact of BERT over the original English dataset. In the final version of this paper we will provide an in depth comparison between DrQA and BERT.

4 Conclusions

This paper explores the application of Bidirectional Encoder Representations within the QA task in Italian, enabled by the recent availability of a large-scale annotated corpus, SQuAD-IT. The experimental results confirm the robustness of the adopted Transformer-based architecture, with a significant improvement with respect to earlier neural architectures. This result paves the way to the development of portable, robust and accurate neural models for QA in Italian, and future work will certainly consider other possible extensions of the adopted model.

References

- Petr Baudiš and Jan Šedivý. 2015. Modeling of the Question Answering Task in the YodaQA System. In Josanne Mothe, Jacques Savoy, Jaap Kamps, Karen Pinel-Sauvagnat, Gareth Jones, Eric San Juan, Linda Capellato, and Nicola Ferro, editors, *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, pages 222–228, Cham. Springer International Publishing.
- Jonathan Berant, Andrew Chou, Roy Frostig, and Percy Liang. 2013. Semantic parsing on freebase from question-answer pairs. In *EMNLP*, pages 1533–1544. ACL.
- E. Brill, S. Dumais, M. Banko, Eric Brill, Michele Banko, and Susan Dumais. 2002. An Analysis of the AskMSR Question-Answering System. In *Proceedings of EMNLP 2002*, January.
- Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading wikipedia to answer open-domain questions. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1870–1879.
- Christopher Clark and Matt Gardner. 2018. Simple and effective multi-paragraph reading comprehension. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 845–855, Melbourne, Australia, July. Association for Computational Linguistics.
- Danilo Croce, Alexandra Zelenanska, and Roberto Basili. 2018. Neural learning for question answering in italian. In Chiara Ghidini, Bernardo Magnini, Andrea Passerini, and Paolo Traverso, editors, *AI*IA 2018 – Advances in Artificial Intelligence*, pages 389–402, Cham. Springer International Publishing.
- Danilo Croce, Alexandra Zelenanska, and Roberto Basili. 2019. Enabling deep learning for large scale question answering in italian. *Intelligenza Artificiale*, 13(1):49–61.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June. Association for Computational Linguistics.
- David A. Ferrucci, Eric W. Brown, Jennifer Chu-Carroll, James Fan, David Gondek, Aditya Kalyanpur, Adam Lally, J. William Murdock, Eric Nyberg, John M. Prager, Nico Schlaefer, and Christopher A. Welty. 2010. Building Watson: An Overview of the DeepQA Project. *AI Magazine*, 31(3):59–79.
- Ross B. Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. 2013. Rich feature hierarchies for accurate object detection and semantic segmentation. *CoRR*, abs/1311.2524.
- Sanda M. Harabagiu, Dan I. Moldovan, Marius Pasca, Rada Mihalcea, Mihai Surdeanu, Razvan C. Bunescu, Roxana Girju, Vasile Rus, and Paul Morarescu. 2000. FALCON: boosting knowledge for answer engines. In *Proceedings of The Ninth Text REtrieval Conference, TREC 2000, Gaithersburg, Maryland, USA, November 13-16, 2000*.
- L. Hirschman and R. Gaizauskas. 2001. Natural language question answering: the view from here. *Natural Language Engineering*, 7(4):275–300.
- Cody C. T. Kwok, Oren Etzioni, and Daniel S. Weld. 2001. Scaling question answering to the web. In *WWW*, pages 150–161.
- Alexander H. Miller, Adam Fisch, Jesse Dodge, Amir-Hossein Karimi, Antoine Bordes, and Jason Weston. 2016. Key-value memory networks for directly reading documents. In *EMNLP*.
- Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana, June. Association for Computational Linguistics.
- Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ Questions for Machine Comprehension of Text. *CoRR*, abs/1606.05250.
- Mike Schuster and Kaisuke Nakajima. 2012. Japanese and korean voice search. In *International Conference on Acoustics, Speech and Signal Processing*, pages 5149–5152.
- Min Joon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi. 2017. Bidirectional attention flow for machine comprehension. In *5th International Conference on Learning Representations*,

ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings. OpenReview.net.

Huan Sun, Hao Ma, Wen tau Yih, Chen-Tse Tsai, Jingjing Liu, and Ming-Wei Chang. 2015. Open domain question answering via semantic enrichment. In *WWW*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.

Applying Psychology of Persuasion to Conversational Agents through Reinforcement Learning: an Exploratory Study

Francesca Di Massimo¹, Valentina Carfora², Patrizia Catellani² and Marco Piastra¹

¹Computer Vision and Multimedia Lab, Università degli Studi di Pavia, Italy

²Dipartimento di Psicologia, Università Cattolica di Milano, Italy

francesca.dimassimo01@universitadipavia.it

valentina.carfora@unicatt.it

patrizia.catellani@unicatt.it

marco.piastra@unipv.it

Abstract

This study is set in the framework of *task-oriented conversational agents* in which *dialogue management* is obtained via *Reinforcement Learning*. The aim is to explore the possibility to overcome the typical end-to-end training approach through the integration of a quantitative model developed in the field of persuasion psychology. Such integration is expected to accelerate the training phase and improve the quality of the dialogue obtained. In this way, the resulting agent would take advantage of some subtle psychological aspects of the interaction that would be difficult to elicit via end-to-end training. We propose a theoretical architecture in which the psychological model above is translated into a probabilistic predictor and then integrated in the reinforcement learning process, intended in its *partially observable* variant. The experimental validation of the architecture proposed is currently ongoing.

manager via machine learning techniques, *reinforcement learning* (RL) in particular, may seem attractive. At present, many RL-based approaches involve training an agent end-to-end from a dataset of recorded dialogues, see for instance Liu (2018). However, the chance of obtaining significant results in this way entails substantial efforts in both collecting sample data and performing experiments. Worse yet, such efforts ought to rely on the even stronger hypothesis that the RL agent would be able to elicit psychosocial aspects on its own. As an alternative, in this study we envisage the possibility to enhance the RL process by harnessing a model developed and accepted in the field of social psychology to provide a more reliable learning ground and a substantial accelerator for the process itself.

Our study relies on a quantitative, causal model of human behavior being studied in the field of social psychology (see Carfora et al., 2019) aimed at assessing the effectiveness of message *framing* to induce healthier nutritional habits. The goal of the model is to assess whether messages with different frames can be differentially persuasive according to the users' psychosocial characteristics.

1 Introduction

A typical conversational agent has a multi-stage architecture: spoken language, written language and dialogue management, see Allen et al. (2001). This study focuses on dialogue management for task-oriented conversational agents. In particular, we focus on the creation of a dialogue manager aimed at inducing healthier nutritional habits in the interactant.

Given that the task considered involves psychosocial aspects that are difficult to program directly, the idea of achieving an effective dialogue

2 Psychological model: Structural Equation Model

Three relevant psychosocial antecedents of behaviour change are the following: *Self-Efficacy* (the individual perception of being able to eat healthy), *Attitude* (the individual evaluation of the pros and cons) and *Intention Change* (the individual willingness of adhering to a healthy diet). These psychosocial dimensions cannot be directly observed and need to be measured as *latent* variables. To this purpose, questionnaires are used, each composed by a set of questions or *items* (i.e. *observed* variables). *Self-Efficacy* is measured with 8 items, each associated to a set of answers ranging from "not at all confident" (1)

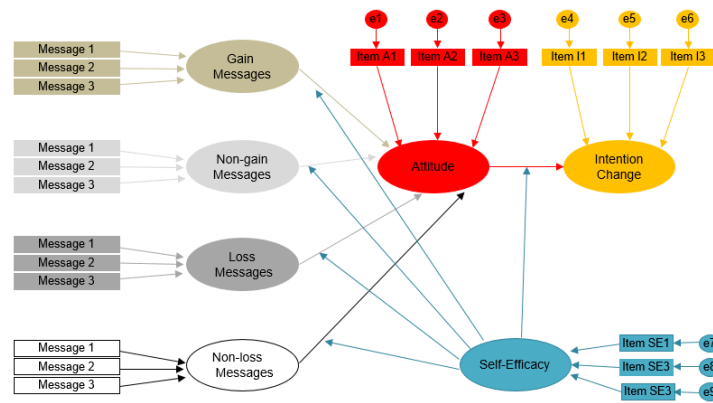


Figure 1: SEM simplified model for the case at hand.

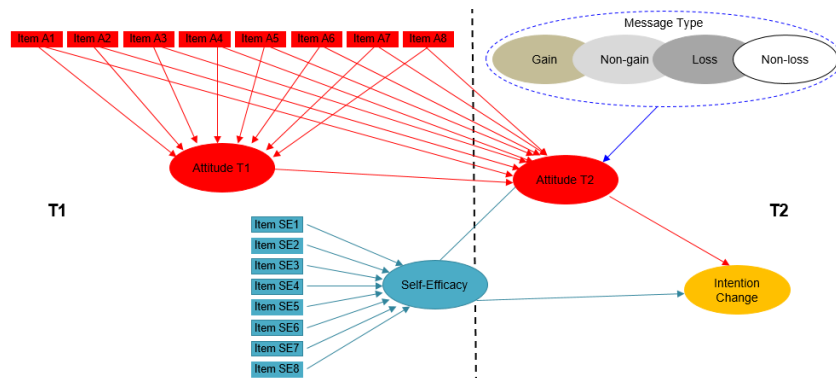


Figure 2: DBN translation of the SEM shown in Figure 1.

to "extremely confident" (7). *Attitude* is assessed through 8 items associated to a differential scale ranging from 1 to 7 (the higher the score, the more positive the attitude). *Intention Change* is measured with three items on a Likert scale, ranging from 1 ("definitely do not") to 7 ("definitely do"). See Carfora et al. (2019).

In our study, the psychosocial model was assessed experimentally on a group of volunteers. Each participant was first proposed a questionnaire (Time 1 – T1) for measuring *Self-Efficacy*, *Attitude* and *Intention Change*. In a subsequent phase (i.e. *message intervention*), participants were randomly assigned to one of four groups, each receiving a different type of persuasive message: *gain* (i.e. positive behavior leads to positive outcomes), *non-gain* (negative behavior prevents positive outcomes), *loss* (negative behavior leads to negative outcomes) and *non-loss* (positive behavior prevents negative outcomes) (Higgins, 1997; Cesario et al., 2013). In a last phase (Time 2 - T2), the effectiveness of the message intervention was then evaluated with a second questionnaire, to detect changes in participants' *Attitude* and *Intention Change* in relation to healthy eating.

Intention Change in relation to healthy eating.

The overall model is described by the *Structural Equation Model* (SEM, see Wright, 1921) in Figure 1. For simplicity, only three items are shown for each latent variable. Besides allowing the description of latent variables, SEMs are *causal* models in the sense that they allow a statistical analysis of the strength of causal relations among the latents themselves, as represented by the arrows in figure. SEMs are linear models, and thus all causal relations underpin linear equations.

Note that latent variables in a SEM have different roles: in this case *gain/non-gain/loss/non-loss* messages are *independent variables*, *Intention Change* is a *dependent variable*, *Attitude* is a *mediator* of the relationship between the independent and the dependent variables, and *Self-Efficacy* is a *moderator*, namely, it explains the intensity of the relation it points at. *Intention Change* was measured at both T1 and T2, *Attitude* was measured at both T1 and T2, and *Self-Efficacy* was measured at T1 only. Note that the time transversality (i.e. T1 → T2) is implicit in the SEM depiction above.

3 Probabilistic model: Bayesian Network

Once the SEM is defined, we aim to translate it into a probabilistic model, so as to obtain the probability distributions needed for the learning process. We resort to a graphical model, and in particular to a *Bayesian Network* (BN, see Ben Gal, 2007), namely a graph-based description of both the observable and latent random variables in the model and their conditional dependencies. In BNs, nodes represent the variables and edges represent dependencies between them, whereas the lack of edges implies their independence, hence a simplification in the model. As a general rule, the joint probability of a BN can be inferred as follows:

$$P(X_1, \dots, X_N) = \prod_{i=1}^N P(X_i \mid \text{parents}(X_i)),$$

where X_1, \dots, X_N are the random variables in the model and $\text{parents}(X_i)$ indicate all the nodes having an edge oriented towards X_i .

In the case at hand, a temporal description of the model, accounting for the time steps T1 and T2, is necessary as well. For this purpose, we use a *Dynamic Bayesian Network* (DBN, see Dagum et al., 1992). The DBN thus obtained is shown in Figure 2.

Notice that the messages are only significant at T2, as they have not been sent yet at T1. We gathered message in the one node *Message Type*, assuming it can take four, mutually exclusive values. The mediator *Attitude* is measured at both time steps while the moderator *Self-Efficacy* is constant over time, as suggested in Section 2. *Intention Change* has relevance at T2 only since, as we will mention in Section 5, it will be used to estimate a reward function once the final time step is reached.

4 Learning the BN

The collected data are as follows. The analysis was conducted on 442 interactants, divided in four groups, each one receiving a different type of messages¹. The answers to the items of the questionnaire always had a range of 7 values. However, this induces a combinatory explosion, making it impossible to cover all the subspaces ($7^8 = 5.764.801$ different combinations for *Attitude*, for instance). We thus decide to aggregate: *low* :=

¹The original study included also a control group, which we do not consider here.

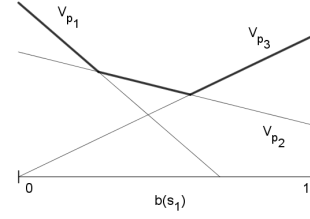


Figure 3: Basic example of computation of V_π in a case where $\mathcal{S} = \{s_1, s_2\}$. p_1, p_2, p_3 are three possible policies.

(1 to 2); *medium* := (3 to 5); *high* := (6 to 7).

Our aim is to learn the *Joint Probability Distribution* (JPD) of our model, as that would make us able to answer, through marginalizations and conditional probabilities, any query about the model itself. The conditional probability distributions to be learnt in the case in point are then the following:

- $P(\text{Item } Ai)$, for $i = 1, \dots, 8$;
- $P(\text{Item } SEi)$, for $i = 1, \dots, 8$;
- $P(\text{Message Type})$;
- $P(\text{Attitude } T1 \mid \text{Item } Ai, i = 1, \dots, 8)$;
- $P(\text{Self-Efficacy} \mid \text{Item } SEi, i = 1, \dots, 8)$;
- $P(\text{Attitude } T2 \mid \text{Item } Ai, i = 1, \dots, 8, \text{Message Type}, \text{Self-Efficacy})$;
- $P(\text{Intention Change} \mid \text{Attitude } T2, \text{Self-Efficacy})$.

The first three can be easily inferred from the raw data as relative frequencies. As for the following four, even aggregating the 7 values as mentioned, a huge amount of data would still be necessary ($3^8 \cdot 2^4 \cdot 3 = 314.928$ subspaces for *Attitude T2*, for instance). As conducting a psychological study on that amount of people would not be feasible, we address the issue with an appropriate choice of the method. To allow using *Maximum Likelihood Estimation* (MLE) to learn the BN, we resort to the *Noisy-OR* approximation (see Oniško, 2001). According to this, through a few appropriate changes (not shown) to the graphical model, the number of subspaces can be greatly reduced (e.g. $3 \cdot 2 \cdot 3 = 18$ for *Attitude T2*).

5 Reinforcement Learning: Markov Decision Problems

The translation into a tool to be used for reinforcement learning is obtained in the terms of *Markov*

Decision Processes (MDPs), see Fabiani et al. (2010).

Roughly speaking, in a MDP there is a finite number of situations or *states* of the environment, at each of which the agent is supposed to select an action to take, thus inducing a state transition and obtaining a *reward*. The objective is to find a *policy* determining the sequence of actions that generates the maximum possible cumulative reward, over time. However, due to the presence of latents, in our case the agent is not able to have complete knowledge about the state of the environment. In such a situation, the agent must build its own estimate about the current state based on the memory of past actions and observations. This entails using a variant of the MDPs, that is *Partially Observable Markov Decision Processes* (POMDPs, see Kaelbling 1998). We then define the following, with reference to the variables mentioned in Figure 2:

$$\mathcal{S} := \{\text{states}\} = \{\text{Attitude } T2, \text{Self-Efficacy}\};$$

$$\mathcal{A} := \{\text{actions}\} = \{\text{ask } A1, \dots, \text{ask } A8\} \cup \{\text{ask } SE1, \dots, \text{ask } SE8\} \cup \{G, NG, L, NL\},$$

where A_i denotes the question for *Item* A_i , SE_i denotes the question for *Item* SE_i and G, NG, L, NL denote the action of sending Gain, Non-gain, Loss and Non-loss messages respectively;

$$\Omega := \{\text{observations}\} = \{\text{Item } A1, \dots, \text{Item } A8, \text{Item } SE1, \dots, \text{Item } SE8\}.$$

Starting from an unknown initial state s_0 (often taken to be uniform over \mathcal{S} , as no information is available), the agent takes an action a_0 , that brings it, at time step 1, to state s_1 , unknown as well. There, an observation o_1 is made.

The process is then repeated over time, until a *goal* state of some kind has been reached. Hence, we can define the *history* as an ordered succession of actions and observations:

$$h_t := \{a_0, o_1, \dots, a_{t-1}, o_t\}, h_0 = \emptyset.$$

As at all steps there is uncertainty about the actual state, a crucial role is played by the agent's estimate about the state of the environment, i.e. by the *belief state*. The agent's belief at time step t , denoted as \mathbf{b}_t , is driven by its previous belief \mathbf{b}_{t-1} and by the new information acquired, i.e. the action taken a_{t-1} and observation made o_t . We then have:

$$b_{t+1}(s_{t+1}) = P(s_{t+1} \mid \mathbf{b}_t, a_t, o_{t+1}).$$

In the POMDP framework, the agent's choices about how to behave are influenced by its belief state and by the history. Thus, we define the agent's *policy*:

$$\pi = \pi(\mathbf{b}_t, h_t),$$

that we aim to optimize. To complete the picture, we define the following functions to describe the model evolution in time (the notation $'$ indicates a reference to the subsequent time step):

state-transition function:

$$T: (s, a) \mapsto P(s' \mid s, a) := T(s', s, a);$$

observation function:

$$O: (s, a) \mapsto P(o' \mid a, s') := O(o', a, s');$$

reward function:

$$R: (s, a) \mapsto \mathbb{E}[r' \mid s, a] := R(s, a).$$

These functions can be easily adapted to the specifics of the case at hand. It can be seen that, once the JPD derived from the DBN is completely specified, the reward is deterministic. In particular, it is computed by evaluating the changes in the values for the latent *Intention Change*.

As we are interested in finding an optimal policy, we now need to evaluate the goodness of each state when following a given policy. As there is no certainty about the states, we define the *value function* as a weighted average over the possible belief states:

$$V_\pi(\mathbf{b}_t, h_t) := \sum_{s_t} b_t(s_t) V_\pi(s_t, \mathbf{b}_t, h_t),$$

where $V_\pi(s_t, \mathbf{b}_t, h_t)$ is the *state value function*. The latter depends on the expected reward (and on a discount factor $\gamma \in [0, 1]$ stating the preference for fast solutions):

$$\begin{aligned} V_\pi(s_t, \mathbf{b}_t, h_t) := & R(s_t, \pi(\mathbf{b}_t, h_t)) + \\ & \gamma \sum_{s_{t+1}} T(s_{t+1}, s_t, \pi(\mathbf{b}_t, h_t)) * \\ & \sum_{o_{t+1}} O(o_{t+1}, \pi(\mathbf{b}_t, h_t), s_{t+1}) V_\pi(s_{t+1}, \mathbf{b}_{t+1}, h_{t+1}). \end{aligned}$$

Finally, we define the target of our seek, namely the *optimal value function* and the related *optimal policy*, as:

$$\begin{cases} V^*(\mathbf{b}_t, h_t) := \max_\pi V_\pi(\mathbf{b}_t, h_t), \\ \pi^*(\mathbf{b}_t, h_t) := \operatorname{argmax}_\pi V_\pi(\mathbf{b}_t, h_t). \end{cases}$$

It can be shown that the optimal value function in a POMDP is always piecewise linear and convex, as

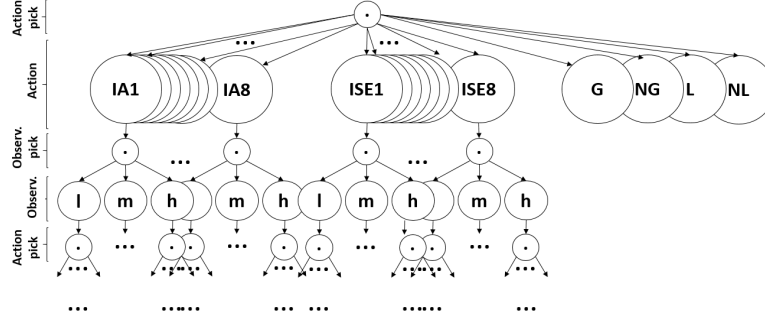


Figure 4: Expansion of the policy tree. l, m, h stand for *low, medium* and *high*.

exemplified in Figure 3. In other words, the optimal policy (in bold in Figure 3) combines different policies depending on their belief state values.

The next step is to use the POMDP to detect the optimal policy, that is the sequence of questions to ask to the interactant, in order to draw her/his profile, hence the message to send, which maximizes the effectiveness of the interaction. To this end, the contribution of the DBN is fundamental. From the JPD associated, in fact, we construct the probability distributions necessary to define the functions T, O, R that compose the value function.

6 Policy from Monte Carlo Tree Search

It is evident from Figure 4, describing the full expansion of the policy tree for the case in point, that the computational effort and power required for a brute-force exploration of all possible combinations is unaffordable.

Among all the policies that can be considered, we want to select the optimal ones, thus avoiding considering policies that are always underperforming. In other words, with reference to Figure 3, we want to find $V_{p_1}, V_{p_2}, V_{p_3}$ among those of all possible policies, and use them to identify the optimal policy V^* .

To accomplish this, we select the *Monte Carlo Tree Search* (MCTS) approach, see Chaslot et al. (2008), due to its reliability and its applicability to computationally complex practical problems. We adopt the variant including an *Upper Confidence Bound formula*, see Kocsis et al. (2006). This method combines *exploitation* of the previously computed results, allowing to select the game action leading to better results, with *exploration* of different choices, to cope with the uncertainty of the evaluation. Thus, using $V_\pi(s_t, \mathbf{b}_t, h_t)$ as defined before to guide the exploration, the MCTS method reliably converges (in probability) to op-

timal policies. These latter will be applied by the conversational agent in the interaction with each specific user, to adapt both the sequence and the amount of questions to her/his personality profile and selecting the message which is most likely to be effective.

7 Conclusions and future work

In this work we explored the possibility of harnessing a complete and experimentally assessed SEM, developed in the field of persuasion psychology, as the basis for the reinforcement learning of a dialogue manager that drives a conversational agent whose task is inducing healthier nutritional habits in the interactant. The fundamental component of the method proposed is a DBN, which is derived from the SEM above and acts like a predictor for the belief state value in a POMDP.

The main expected advantage is that, by doing so, the RL agent will not need a time-consuming period of training, possibly requiring the involvement of human interactants, but can be trained ‘in house’ – at least at the beginning – and be released in production at a later stage, once a first effective strategy has been achieved through the DBN. Such method still requires an experimental validation, which is the current objective of our working group.

Acknowledgments

The authors are grateful to Cristiano Chesi of IUSS Pavia for his revision of an earlier version of the paper and his precious remarks. We also acknowledge the fundamental help given by Rebecca Rastelli, during her collaboration to this research.

References

- Allen, J., Ferguson, G., & Stent, A. 2001. *An architecture for more realistic conversational systems*. In Proceedings of the 6th international conference on Intelligent user interfaces (pp. 1-8). ACM.
- Anderson, Ronald D. & Vastag, Gyula. 2004. *Causal modeling alternatives in operations research: Overview and application*. European Journal of Operational Research. 156. 92-109.
- Auer, Peter & Cesa-Bianchi, Nicolò & Fischer, Paul. 2002. *Kocsis, Levente & Szepesvári, Csaba. 2006. Bandit Based Monte-Carlo Planning. Finite-time Analysis of the Multiarmed Bandit Problem*. Machine Learning. 47. 235-256.
- Bandura, A. 1982. *Self-efficacy mechanism in human agency*. American Psychologist, 37, 122-147.
- Baron, Robert A. & Byrne, Donn Erwin & Suls, Jerry M. 1989. *Exploring social psychology, 3rd ed.* Boston, Mass.: Allyn and Bacon. 0205119085.
- Ben Gal I. 2007. *Bayesian Networks*. Encyclopedia of Statistics in Quality and Reliability. John Wiley & Sons.
- Bertolotti, M., Carfora, V., & Catellani, P. 2019. *Different frames to reduce red meat intake: The moderating role of self-efficacy*. Health Communication, in press.
- Carfora, V., Bertolotti, M., & Catellani, P. 2019. *Informational and emotional daily messages to reduce red and processed meat consumption*. Appetite, 141, 104331.
- Cesario, J., Corker, K. S., & Jelinek, S. 2013. *A self-regulatory framework for message framing*. Journal of Experimental Social Psychology, 49, 238-249.
- Chaslot, Guillaume & Bakkes, Sander & Szita, Istvan & Spronck, Pieter. 2008. *Monte-Carlo Tree Search: A New Framework for Game AI*. Bijdragen.
- Dagum, Paul and Galper, Adam and Horvitz, Eric. 1992. *Dynamic Network Models for Forecasting*. Proceedings of the Eighth Conference on Uncertainty in Artificial Intelligence.
- Dagum, Paul and Galper, Adam and Horvitz, Eric and Seiver, Adam. 1999. *Uncertain reasoning and forecasting*. International Journal of Forecasting.
- De Waal, Alta & Yoo, Keunyoung. 2018. *Latent Variable Bayesian Networks Constructed Using Structural Equation Modelling*. 2018 21st International Conference on Information Fusion. 688-695.
- Fabiani, Patrick & Teichteil-Königsbuch, Florent. 2010. *Markov Decision Processes in Artificial Intelligence*. Wiley-ISTE.
- Gupta, Sumeet & W. Kim, Hee. 2008. *Linking structural equation modeling to Bayesian networks: Decision support for customer retention in virtual communities*. European Journal of Operational Research. 190. 818-833.
- Heckerman, David. 1995. *A Bayesian Approach to Learning Causal Networks*.
- Higgins, E.T. 1997. *Beyond pleasure and pain*. American Psychologist, 52, 1280-1300.
- A. Howard, Ronald. 1972. *Dynamic Programming and Markov Process*. The Mathematical Gazette. 46.
- Pack Kaelbling, Leslie & Littman, Michael & R. Cassandra, Anthony. 1998. *Planning and Acting in Partially Observable Stochastic Domains*. Artificial Intelligence. 101. 99-134.
- Kocsis, Levente & Szepesvári, Csaba. 2006. *Bandit Based Monte-Carlo Planning*. Machine Learning: ECML 2006. Springer Berlin Heidelberg. 282-293.
- Lai, T.L & Robbins, Herbert. 1985. *Asymptotically Efficient Adaptive Allocation Rules*. Advances in Applied Mathematics. 6. 4-22.
- Liu, Bing. 2018. *Learning Task-Oriented Dialog with Neural Network Methods*. PhD thesis.
- Murphy, Kevin. 2012. *Machine Learning: A Probabilistic Perspective*. The MIT Press. 58.
- Pearl Judea. 1988. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Representation and Reasoning Series (2nd printing ed.). San Francisco, California: Morgan Kaufmann.
- Oniśko, Agnieszka & Druzdzal, Marek J. & Wasyluk, Hanna. 2001. *Learning Bayesian network parameters from small data sets: application of Noisy-OR gates*. International Journal of Approximate Reasoning. 27.
- Silver, David & Veness, Joel. 2010. *Monte-Carlo Planning in Large POMDPs*. Advances in Neural Information Processing Systems. 23. 2164-2172.
- Matthijs T. J. Spaan. 2012. *Partially Observable Markov Decision Processes*. In: Reinforcement Learning: State of the Art. Springer Verlag. 387-414.
- Sutton, Richard & G. Barto, Andrew. 1998. *Reinforcement Learning: An Introduction*. IEEE transactions on neural networks / a publication of the IEEE Neural Networks Council. 9. 1054.
- Wright, Sewall. 1921. *Correlation and causation*. Journal of Agricultural Research. 20. 557-585.
- Young, Steve & Gasic, Milica & Thomson, Blaise & Williams, Jason. 2013. *POMDP-based statistical spoken dialog systems: A review*. Proceedings of the IEEE, 101. 1160-1179.

WebIsAGraph: A Very Large Hypernymy Graph from a Web Corpus

Stefano Faralli¹, Irene Finocchi², Simone Paolo Ponzetto³, Paola Velardi²

¹University of Rome Unitelma Sapienza, Italy

stefano.faralli@unitelmasapienza.it

² University of Rome Sapienza, Italy

irene.finocchi@uniroma1.it velardi@di.uniroma1.it

³ University of Mannheim, Germany

simone@informatik.uni-mannheim.de

Abstract

In this paper, we present WebIsAGraph, a very large hypernymy graph compiled from a dataset of *is-a* relationships extracted from the CommonCrawl. We provide the resource together with a Neo4j plugin to enable efficient searching and querying over such large graph. We use WebIsAGraph to study the problem of detecting polysemous terms in a noisy terminological knowledge graph, thus quantifying the degree of polysemy of terms found in *is-a* extractions from Web text.

1 Introduction

Acquiring concept hierarchies, i.e., taxonomies from text, is a long-standing problem in Natural Language Processing (NLP). Much previous work leveraged lexico-syntactic patterns, which can be either manually defined (Hearst, 1992) or automatically learned (Shwartz et al., 2016). Pattern-based methods were shown by (Roller et al., 2018) to outperform distributional methods, and can be complemented with state-of-the-art meaning representations such as hyperbolic embeddings (Nickel and Kiela, 2017) to infer missing *is-a* relations and filter wrong extractions (Le et al., 2019). Complementary to these efforts, researchers looked at ways to scale hypernymy detection to very large, i.e., Web-scale corpora (Wu et al., 2012). Recently, (Seitner et al., 2016) applied Hearst patterns to the CommonCrawl¹ to produce the WebIsaDb. Using Web corpora makes it possible to produce hundreds of millions of *is-a* triples: the extractions, however, include many false positives and cycles (Ristoski et al., 2017).

Methods for hypernym detection like, e.g., pattern-based approaches, have a limitation in that they do not necessarily produce proper taxonomies (Camacho-Collados, 2017): automatically detected *is-a* relationships, on the other hand, can be used as input to taxonomy induction algorithms (Velardi et al., 2013; Faralli et al., 2017; Faralli et al., 2018, *inter alia*). These algorithms rely on the topology of the input graph, and, therefore, cannot be applied ‘as-is’ to Web-scale resources like WebIsaDb, since this resource merely consists of a set of triples. Moreover, WebIsaDb does not contain fully semantified triples, i.e., subjects and objects of the *is-a* relationships consist of potentially ambiguous terminological nodes. This is because, due to their large size, source input corpora like the CommonCrawl cannot be semantified upfront. Linking to the semantic vocabulary of a reference resource like DBpedia (Hertling and Paulheim, 2017) also barely mitigate this problem, since Wikipedia-centric knowledge bases have not, and cannot be expected to have, complete coverage over Web data (Lin et al., 2012).

In this paper, we present an initial solution to these problems by building the first very large hypernymy graph, dubbed WebIsAGraph, built from *is-a* relationships extracted from a Web-scale corpus. This is a relevant task: although WordNet (and other thesauri) already provides a catalog of ambiguous terms, many nodes of WebIsAGraph are not covered in available lexicographic resources, because they are proper names, technical terms, or polysemantic words. Our graph – which we make freely available to the research community to foster further work on Web-scale knowledge acquisition – is built from the WebIsaDb on top of state-of-the-art graph mining tools²: thanks to an accompanying plugin, it can be easily searched, queried, and explored. We-

Copyright ©2019 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

¹<http://commoncrawl.org>

²Neo4j: <https://neo4j.com/>

bIsAGraph may represent an opportunity to researchers for investigating approaches to a variety of tasks on large automatically acquired term tuples. As an example, we use our resource to investigate the problem of identifying ambiguous terminological nodes. To automatically detect whether a lexicographic node is ambiguous or not, we use information from both the graph (topological features) and textual labels (word embeddings) as features to train a model using supervised learning. Our results provide a first estimate of the degree of polysemy that can be found among *is-a* relationships from the Web.

2 Creating WebIsAGraph

We created a directed hypernymy graph from the WebIsADb (Seitner et al., 2016). WebIsADb is a Web-scale collection of noisy hypernymy relations harvested with 58 extraction patterns and consisting of 607,621,170 tuples. Since the aim of WebIsADb was to study the behaviour (on a large scale) of Hearst-like extraction patterns, rather than collecting relations with high precision, in order to reduce noise (false positives) we pre-selected the top-20 more precise extraction patterns in (2016) from the original 58 and identified 385,459,302 tuples.

After removing matches with a frequency lower than 3 and isolated nodes, i.e., nodes with degree equal to 0, we obtained a directed graph consisting of 33,030,457 nodes and 65,681,899 directed edges (see Table 1). The generation of such a large graph required several weeks of computation on a quad-core machine with 32 GB of RAM, using a state-of-the art graph-db system, like Neo4j. Note that the inherent sequential nature of the task of indexing tuples, nodes and edges does not benefit from the use of parallel computation. Next, we developed efficient tools for graph querying, which are released to the community, and described in <https://sites.google.com/unitelmasapienza.it/webisagraph/>, where we also include examples of queries.

3 Measuring the polysemy of WebIsAGraph

Let $p_{SI}(n)$ be the function that predicts if a terminological node n corresponds to a *monosemous* or a *polysemous* concept. We leverage a companion sense inventory as a ground truth, and we train different classifiers with a combination of topological

WebIsAGraph	
nodes	33,030,457
edges	65,681,899
weakly connected components	3,099,898
nodes of largest component	26,099,001
Avg. node Degree	3.97

Table 1: Structural statistics of WebIsAGraph

and textual features, described hereafter.

Topological features. Our conjecture is that in a taxonomy-like terminological graph (even a noisy one) there is a correlation between the mutual connectivity of a node neighborhoods and its polysemy. For example, consider the polysemous word *machine* – which, according to WordNet, has at least six heterogeneous meanings, ranging from the ‘any mechanical or electrical device’ to ‘a group that controls the activities of a political party’ – and the monosemous word *floppy disk*. We expect to observe a different degree of mutual connectivity across the corresponding incoming and outgoing nodes. In particular, for monosemous words, we expect a higher mutual connectivity. With reference to Figure 1, left side, the two hypernyms of “*floppy disk*”: “*memory*” and “*data storage*”, have also “*RAM*” as a common hyponym. In contrast, nodes in the direct neighborhood of “*machine*” (leftmost graph in Figure 1) do not have mutual connections.

Our aim is thus to identify topological features that may help quantifying the previously described connectivity properties. To cope with scalability, we consider topological features built on top of 1-hop/2-hop sub-graphs of a node n . Hence, we identify two induced sub-graphs $G^{-+}(n)$ and $G^{+-}(n)$, induced on $V^{-+}(n) = In(n) \cup_{v \in In(n)} Out(v)$ and $V^{+-}(n) = Out(n) \cup_{v \in Out(n)} In(v)$ respectively, where $In(x)$ and $Out(x)$ are the sets of incoming and outgoing nodes of x (including x). Next, we remove from these sub-graphs the node n , and compute the following features:

- $cc_{G^{-+}}(n)$ and $cc_{G^{+-}}(n)$: the resulting number of weakly connected components;
- $v_{G^{-+}}(n)$ and $v_{G^{+-}}(n)$: the resulting number of nodes;
- $e_{G^{-+}}(n)$ and $e_{G^{+-}}(n)$: the resulting number of edges.

With reference to the example of Figure 1, the light gray sub-graph (a) is $G^{-+}(n)$, the dark sub-

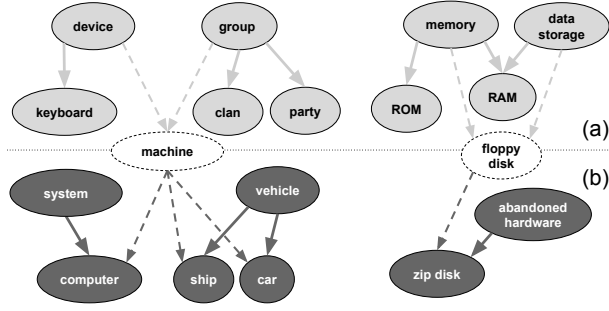


Figure 1: An example excerpt of the neighborhood induced sub-graphs for "machine" and "floppy disk", (a) $G^{+-}(n)$ in gray and (b) $G^{+-}(n)$ in dark gray. Dashed edges connect each n with its hypernyms and hyponyms.

graph (b) is $G^{+-}(n)$, and furthermore for $n = \text{"machine"}$: $cc_{G^{+-}}(n) = 2$, $cc_{G^{+-}}(n) = 2$, $v_{G^{+-}}(n) = 5$, $v_{G^{+-}}(n) = 5$, $e_{G^{+-}}(n) = 3$, and $e_{G^{+-}}(n) = 3$, while for the $n = \text{"floppy disk"}$: $cc_{G^{+-}}(n) = 1$, $cc_{G^{+-}}(n) = 1$, $v_{G^{+-}}(n) = 4$, $v_{G^{+-}}(n) = 2$, $e_{G^{+-}}(n) = 3$, and $e_{G^{+-}}(n) = 1$.

Textual features. Similarly to topological features, our hypothesis is that textual features of the neighborhood nodes should exhibit a lower average similarity when n is polysemous. We extract textual features on top of pre-trained word embeddings, widely adopted in many NLP-related tasks (Camacho-Collados and Pilehvar, 2018). Formally, given a node n :

- $\vec{W}(n)$ is the word embedding vector of n computed as follows:

$$\vec{W}(n) = \frac{\sum_{t \in \text{tokens}(n)} \vec{w}(t)}{|\text{tokens}(n)|} \quad (1)$$

where $\text{tokens}(n)$ is the function that retrieves the set of tokens composing the word n (e.g., if $n = \text{hot dog}$, $\text{tokens}(n) = \{\text{hot}, \text{dog}\}$), and $\vec{w}(t)$ is a pre-trained word embedding vector;

- $\Delta_{in}(n)$ and $\Delta_{out}(n)$: the cosine similarity between $\vec{W}(n)$ and the average word embeddings vector of incoming and outgoing nodes of n respectively;

$$\Delta_{in}(n) = \text{CosSim}(\vec{W}(n), \frac{\sum_{m \in In(n)} \vec{W}(m)}{|In(n)|}) \quad (2)$$

$$\Delta_{out}(n) = \text{CosSim}(\vec{W}(n), \frac{\sum_{m \in Out(n)} \vec{W}(m)}{|Out(n)|}) \quad (3)$$

Algo.	topological			Features textual			all		
	P	R	F ₁	P	R	F ₁	P	R	F ₁
WordNet	Rnd	0.47	0.47	0.47	0.47	0.47	0.47	0.47	0.47
	NN	±.05	±.05	±.05	±.05	±.05	±.05	±.05	±.05
	ABC	0.61	0.61	0.60	0.72	0.72	0.73	0.73	0.73
	GBC	±.02	±.02	±.02	±.03	±.03	±.04	±.04	±.04
DBpedia	Rnd	0.62	0.62	0.62	0.67	0.67	0.70	0.70	0.70
	NN	±.03	±.03	±.03	±.02	±.02	±.03	±.03	±.03
	ABC	0.62	0.62	0.62	0.69	0.68	0.72	0.71	0.71
	GBC	±.02	±.02	±.02	±.01	±.01	±.03	±.03	±.03
WordNet+DBpedia	Rnd	0.51	0.51	0.51	0.51	0.51	0.51	0.51	0.51
	NN	±.03	±.03	±.03	±.03	±.03	±.03	±.03	±.03
	ABC	0.60	0.60	0.59	0.73	0.73	0.74	0.74	0.74
	GBC	±.01	±.01	±.01	±.03	±.03	±.03	±.03	±.03
WordNet+DBpedia	Rnd	0.60	0.60	0.60	0.69	0.69	0.71	0.71	0.71
	NN	±.02	±.02	±.02	±.02	±.02	±.04	±.04	±.04
	ABC	0.61	0.61	0.61	0.70	0.70	0.73	0.73	0.73
	GBC	±.02	±.02	±.02	±.03	±.03	±.02	±.02	±.02
WordNet+DBpedia	Rnd	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50
	NN	±.01	±.01	±.01	±.01	±.01	±.01	±.01	±.01
	ABC	0.54	0.53	0.50	0.70	0.70	0.71	0.70	0.70
	GBC	±.03	±.05	±.12	±.02	±.02	±.02	±.01	±.01
WordNet+DBpedia	Rnd	0.55	0.55	0.55	0.64	0.64	0.65	0.65	0.65
	NN	±.02	±.02	±.02	±.01	±.01	±.02	±.02	±.02
	ABC	0.56	0.56	0.55	0.66	0.66	0.67	0.67	0.67
	GBC	±.02	±.01	±.01	±.02	±.02	±.02	±.02	±.02

Table 2: Performance of different algorithms to detect node ambiguity.

- $Gini(n)$: sparsity index (David, 1968) of $\vec{W}(n)$.

3.1 Evaluation

Computing features. Topological features are efficiently extracted using the query tool mentioned in Section 2. To compute *textual features* (see Section 3) we use the *Glove* pre-trained word embedding vector (Pennington et al., 2014) of length 300 from the CommonCrawl.³

By combining these two types of features (topological and textual) we obtained three different vector input representations consisting of 6 (only topological features), 303 (only textual features) and 309 (textual and topological) dimensions respectively.

Finally, we created three "ground truth" sets of nodes in the graph for which $ps_I(n)$ is known. We selected a balanced number of monosemous and polysemous nouns, using the following sense inventories: i) WordNet (14,659 examples); ii) DBpedia (17,041 examples); iii) WordNet and DBpedia (31,701 examples).

Algorithms. We compared four algorithms:

- Random (*Rnd*): a random baseline which randomly classifies the ambiguity of a node;
- Neural Network (*NN*): a neural network with Softmax activation function in the output layer and dropout (Srivastava et al., 2014);

³<https://nlp.stanford.edu/projects/glove/>.

	Features	WordNet			DBpedia			WordNet \cup DBpedia		
		$dCor$	ρ	PI weight \pm std	$dCor$	ρ	PI weight \pm std	$dCor$	ρ	PI weight \pm std
topological	cc_{G-+}	0.593	0.185	0.0039 \pm 0.0001	0.614	0.228	0.0628\pm0.0051	0.513	0.027	0.0038 \pm 0.0008
	v_{G-+}	0.602	0.203	0.0022 \pm 0.0003	0.597	0.194	0.0045 \pm 0.0010	0.513	0.025	0.0025 \pm 0.0003
	e_{G-+}	0.597	0.194	0.0100 \pm 0.0016	0.600	0.200	0.0048 \pm 0.0008	0.514	0.027	0.0024 \pm 0.0001
	cc_{G+-}	0.606	0.212	0.0131 \pm 0.0013	0.579	0.159	0.0092 \pm 0.0016	0.492	-0.014	0.0049 \pm 0.0003
	v_{G+-}	0.623	0.247	0.0383 \pm 0.0035	0.580	0.159	0.0029 \pm 0.0009	0.495	-0.010	0.0013 \pm 0.0008
	e_{G+-}	0.619	0.237	0.0074 \pm 0.0010	0.583	0.167	0.0034 \pm 0.0013	0.497	-0.006	0.0054 \pm 0.0004
textual	Δ_{in}	0.379	-0.242	0.0699\pm0.0036	0.399	-0.202	0.0231 \pm 0.0023	0.433	-0.134	0.0470\pm0.0027
	Δ_{out}	0.400	-0.199	0.0101 \pm 0.0004	0.415	-0.170	0.0037 \pm 0.0015	0.431	-0.138	0.0120 \pm 0.0007
	$Gini$	0.443	-0.114	0.0042 \pm 0.0004	0.460	-0.080	0.0035 \pm 0.0009	0.494	-0.013	0.0059 \pm 0.0006
	\vec{W} (300 dimensions)	Avg		0.0029 \pm 0.0004	Avg.		0.0030 \pm 0.0005	Avg		0.0028 \pm 0.0003
		Min		0.0016 \pm 0.0003	Min		0.0005 \pm 0.0005	Min		0.0014 \pm 0.0003
		Max		0.0077 \pm 0.0009	Max		0.0180 \pm 0.0013	Max		0.0123 \pm 0.0011

Table 3: Distance correlation $dCor$ and Pearson coefficient ρ between polysemy and features and Permutation Importance (PI) weights (NN estimator).

- Two ensemble-based learning algorithms, namely AdaBoost (*ABC*) (Zhu et al., 2009) and Gradient Boosting (*GBC*) (Friedman, 2001): both have been shown to have high predictive accuracy (Kotsiantis et al., 2006) and are good competitors of neural methods, especially with very large datasets.

Parameter selection. Based on the Area Under Curve ROC (AUC) analysis (Kim et al., 2017), *NN* parameters have been empirically set as follows: i) when testing only with topological features (6 dimensions), we use 2 hidden layers with 4 and 2 neurons respectively and a dropout of 0.2 and 0.15; ii) when using only textual (303 dimensions), or combined textual and topological features (309 dimensions), we use 4 hidden layers, with 128, 64, 32 and 8 neurons respectively and a dropout of 0.3, 0.25, 0.2 and 0.15.

Results. We show in Table 2 the resulting precision, recall and F_1 of the five systems across the ground truths datasets and for the combinations of features (see Section 3). The metrics are averaged on five classification experiments, with a random split (85% train, 10% validation and 5% test) of the ground truth sets. As shown in Table 2, *NN* outperforms the others ensemble methods, obtaining a F_1 score around 0.70. The comparison of performances across the three combinations of features reveals that topological features are not enough to build a model for polysemy classification but can slightly boost the overall already compelling performances of word embeddings-based features.

In Table 3 we show the Person coefficient ρ and

the distance correlation $dCor^4$, with the aim of analyzing how each feature correlates with the polysemy observed in the three ground truth dictionaries. We observed that the features with the highest correlation with polysemy are e_{G+-} , cc_{G-+} and v_{G-+} (see Section 3). Additionally we report the resulting weights of *Permutation Importance* (PI) applied to the *NN* system with the aim of measuring how the performance decreases when a feature is perturbed, by shuffling its values across training examples (Breiman, 2001). We observed that the features which most influenced the performances are $\Delta_{in}(n)$ (WordNet and WordNet \cup DBpedia) and cc_{G-+} (DBpedia). Furthermore, we found that although topological features affect the performance only by a 1% in the average, a number of topologically related features, such as cc_{G-+} , v_{G-+} and e_{G+-} are shown to be indeed related with polysemy. In our future work, we plan to create an ad-hoc ground-truth sense dictionary, since especially WordNet includes extremely fine-grained senses that do not help validating our conjecture about reduced mutual connectivity and contextual similarity of a node’s neighborhood in case of monosemy.

4 Conclusion

The main contribution of this work is a new resource obtained by converting a large dataset of *is-a* (hypernymy) relations automatically extracted from the Web (such as WebIsADb) into a graph structure. This graph, along with its accompanying search tools, enables descriptive and predictive analytics of emerging properties of termino-

⁴ ρ and $dCor$ are indexes to estimate how two distributions are independent.

logical nodes. We used here our new resource to investigate whether a node *polysemy* can be predicted from its topological features (i.e., connectivity patterns) and textual features (meaning representations from word embeddings). The results of this preliminary study have shown that textual features are good predictors of polysemy, while topological features appear to be weaker predictors even if they have a significant correlation with the polysemy of the related node.

References

- Leo Breiman. 2001. Random forests. *Machine Learning*, 45(1):5–32, Oct.
- José Camacho-Collados and Mohammad Taher Pilehvar. 2018. From word to sense embeddings: A survey on vector representations of meaning. *J. Artif. Intell. Res.*, 63:743–788.
- Jose Camacho-Collados. 2017. Why we have switched from building full-fledged taxonomies to simply detecting hypernymy relations.
- H. A. David. 1968. Gini’s mean difference rediscovered. *Biometrika*, 55(3):573–575.
- Stefano Faralli, Alexander Panchenko, Chris Biemann, and Simone Paolo Ponzetto. 2017. The contrastmedium algorithm: Taxonomy induction from noisy knowledge graphs with just a few links. In *Proc. of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 590–600. Association for Computational Linguistics.
- Stefano Faralli, Irene Finocchi, Simone Paolo Ponzetto, and Paola Velardi. 2018. Efficient pruning of large knowledge graphs. In *Proc. of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden.*, pages 4055–4063.
- Jerome H. Friedman. 2001. Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5):1189–1232.
- Marti A. Hearst. 1992. Automatic acquisition of hyponyms from large text corpora. In *Proc. of COLING*, pages 539–545.
- Sven Hertling and Heiko Paulheim. 2017. Webisalod: Providing hypernymy relations extracted from the web as linked open data. In *The Semantic Web - ISWC 2017 - 16th International Semantic Web Conference, Vienna, Austria, October 21-25, 2017, Proc., Part II*, pages 111–119.
- Chulwoo Kim, Sung-Hyuk Cha, Yoo An, and Ned Wilson. 2017. On roc curve analysis of artificial neural network classifiers. In *Florida Artificial Intelligence Research Society Conference*.
- S. B. Kotsiantis, I. D. Zaharakis, and P. E. Pintelas. 2006. Machine learning: a review of classification and combining techniques. *Artificial Intelligence Review*, 26(3):159–190, Nov.
- Matt Le, Stephen Roller, Laetitia Papaxanthos, Douwe Kiela, and Maximilian Nickel. 2019. Inferring concept hierarchies from text corpora via hyperbolic embeddings.
- Thomas Lin, Mausam, and Oren Etzioni. 2012. Entity linking at web scale. In *Proc. of the Joint Workshop on Automatic Knowledge Base Construction and Web-scale Knowledge Extraction (AKBC-WEKEX)*, pages 84–88. Association for Computational Linguistics.
- Maximilian Nickel and Douwe Kiela. 2017. Poincaré embeddings for learning hierarchical representations. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 6341–6350. Curran Associates, Inc.
- Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *EMNLP*, volume 14, pages 1532–1543.
- Petar Ristoski, Stefano Faralli, Simone Paolo Ponzetto, and Heiko Paulheim. 2017. Large-scale taxonomy induction using entity and word embeddings. In *Proc. of the International Conference on Web Intelligence, WI ’17*, pages 81–87, New York, NY, USA. ACM.
- Stephen Roller, Douwe Kiela, and Maximilian Nickel. 2018. Hearst patterns revisited: Automatic hypernym detection from large text corpora. In *Proc. of the 56th ACL (Volume 2: Short Papers)*, pages 358–363. Association for Computational Linguistics.
- Julian Seitner, Christian Bizer, Kai Eckert, Stefano Faralli, Robert Meusel, Heiko Paulheim, and Simone Paolo Ponzetto. 2016. A large database of hypernymy relations extracted from the web. In *Proc. of the Tenth International Conference on Language Resources and Evaluation LREC 2016, Portorož, Slovenia, May 23-28, 2016*.
- Vered Shwartz, Yoav Goldberg, and Ido Dagan. 2016. Improving hypernymy detection with an integrated path-based and distributional method. In *Proc. of the 54th ACL (Volume 1: Long Papers)*, pages 2389–2398. Association for Computational Linguistics.
- Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: A simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.*, 15(1):1929–1958, January.
- Paola Velardi, Stefano Faralli, and Roberto Navigli. 2013. Ontolearn reloaded: A graph-based algorithm for taxonomy induction. *Computational Linguistics*, 39(3).

- Wentao Wu, Hongsong Li, Haixun Wang, and Kenny Q. Zhu. 2012. Probase: A probabilistic taxonomy for text understanding. In *Proc. of the 2012 ACM SIGMOD International Conference on Management of Data*, SIGMOD '12, pages 481–492, New York, NY, USA. ACM.
- Ji Zhu, Hui Zou, Saharon Rosset, and Trevor Hastie. 2009. Multi-class adaboost. *Statistics and Its Interface*, 2(3):349–360.

CULTURE as a 'Liquid' Modern Word. Evidence from Synchronic and Diachronic Language Resources

Maristella Gatto

Dipartimento di Lettere Lingue Arti.

Italianistica e Culture Compare

maristella.gatto@uniba.it

Abstract

The aim of this paper is to discuss the results of a corpus-based investigation of the process that has transformed the very specific material meaning of the word “culture” into the extremely elusive, liquid (Bauman 2011) concept we are familiar with today. The analysis starts from the lexicogrammar profile of the word “culture” in contemporary synchronic corpus resources (Gatto 2011; 2014) and attempts further exploration of these findings on the basis of diachronic language resources. In particular, data from Google Books, accessed both via Ngram Viewer and through the tools available at BYU Corpora, have been used to test hypotheses for the behaviour of the word “culture” in the 19th and early 20th century, whereas data from EEBO (Early English Books Online) have been used to explore patterns of usage in the period of time from 15th to 18th century.

The partial results of this research suggest that there is room for far reaching investigations into the (hi)story of this intriguing “complicated” word, as Williams (1985: 87) dubbed it, and that computational methods and language resources can well complement studies carried out in the context of the digital humanities, from the perspective of historical linguistics, sociolinguistics and cultural studies, when not providing the basis for fresh new insights and further explorations.

1 Introduction

Sketching the “historical peregrinations” of the concept of culture in his *Culture in a Liquid Modern World*, Bauman outlines the changing role of culture in society, from “an agent for

change”, to “a conservative force”, to an increasingly flexible and liquid concept “fashioned to fit individual freedom” (Bauman 2011: 1-17). It is against this background that this paper attempts an investigation of the multifaceted process that over the centuries has transformed the very specific material meaning of the word “culture” into the extremely elusive concept we are familiar with today. The basic assumption is that the process of semantic change which transformed a word originally referring to the concepts of tillage and husbandry (from the Latin *colere*) into a potentially polysemic word accommodating a far wider range of meanings is mirrored in changes in usage of the word, and in turn reflects changes in society. In the wake of a growing interest for the use of language resources for the investigation of cultural and social phenomena (e.g. Michel et al 2010) these changes can be observed through the quantitative and qualitative analysis of the lexicogrammar patterns the word “culture” has entered during its long history of existence.

The very choice for the word “culture” originates in Raymond Williams’ famous statement that culture is “one of the two or three most complicated words in the English language”. By identifying “culture” as one of the key words of our times, Williams reminds us (1985: 87–93) that culture used to be, in its early uses, the noun of a process: the tending of something, basically crops or animals. This meaning provided a basis for the important next stage of metaphorization, when the tending of natural growth was extended to a process of human development so that the word “culture” came to be taken in absolute terms as signifying a process of refinement. After tracing the key moments in the development of this word, Williams distinguishes three categories in modern usage:

- (i) the noun which describes a process of intellectual, spiritual and aesthetic refinement; e.g. a man of culture;

"Copyright © 2019 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0)."

(ii) the noun which describes the products of intellectual and especially artistic activity; e.g. Ministry of Culture;

(iii) the noun which indicates a particular way of life, whether of a people, a period, a group, or humanity in general; e.g. Jewish culture.

This paper takes Williams as a starting point to provide empirical evidence of the ways the word “culture” is used in the English language. Indeed Williams himself, in his introduction to *Culture and Society*, states that an enquiry into the development of this word should be carried out by examining “not a series of abstracted problems, but a series of statements by individuals” (1966: xvii), which bears striking similarities – though not intended – with the corpus-based approach adopted in the present research.

2 Sketches of CULTURE. Evidence from synchronic resources

The starting point of the investigation carried out in the paper are the preliminary results of investigation into the lexicogrammar profile of the word “culture” using Sketch Engine, a corpus query tool that provides a one-page summary of the lexico-grammar patterns of a word from a given corpus, as reported in Gatto (2011; 2014). In the first part of this paper data from three synchronic corpora of English (BNC UKWaC and EnTenTen) will be compared. For a start, the table below reports the number of occurrences and the normalized frequency of the word CULTURE in each corpus:

BNC	UKWAC	ENTENTEN
10,281	200,663	3,692,159
90,1	129,70	200,80
per million	per million	per million

It should be noted, that these occurrences obviously include both those in which “culture” is used with its general meaning in the humanities, which is the primary concern of the present analysis, and those in which it is used as a scientific term (e.g. *cell culture*). Since the tools and resources used for the present research do not allow for a disambiguation between the two meanings, an attempt has been made – heuristically – to estimate the number of occurrences of “culture” in its scientific sense, by computing the number of occurrences of “culture” with the

lemmas “cell” or “bacteria” in their co-text. This was done using the filter option and setting a broad co-text (15 words to the left and to the right of the node). The results seem to indicate that nearly 9,472 (2,46 per million) occurrences of “culture” can be related to its scientific meaning in the BNC, 2,499 (1,61 per million) in UKWaC, and 98,714 (3,47 per million) in EnTenTen. While the method used was not to be considered totally reliable, on the basis of the relative negligibility of the results this aspect has not been taken into account in the following commentary of the data (but this is certainly an aspect which needs to be handled with care when pursuing further research on this topic).

Given the limited scope of the paper only three lexico-grammar patterns will be focused on in the subsections below.

2.1 Culture as object

When considering the list of verbs having “culture” as their object, it seems that according to data computed by the Sketch Engine for all the three corpora, the word “culture” has a consistent tendency to occur as the object of such verbs as *foster*, *promote*, *create*, *reflect*, *understand*, *shape* *change*. *Respect* does not appear only in the list for EnTenTen, as it is n.16, just out of the first 15 positions chosen as a sample. While these data are encouraging in showing that corpora built in different ways yield consistent results for the collocates of “culture”, something interesting can be observed with reference to the changing position of some collocates. The collocates *foster*, *promote*, *create*, and *change* seem to gain prominence in the two more recent web corpora, but it is also interesting to consider how *experience*, which did not appear in the top 15 list from the BNC, is one of the most significant collocates for “culture” in the other two corpora. By contrast, such patterns as *assimilate*, *absorb* *transmit*, which indirectly refer to power relations, appear to be unique to the BNC.

Previous research has already discussed how concordance lines for such pattern as *create* + *culture* or *foster* + *culture* have a frequent co-occurrence with words relating to the socio-economic domain, such as *staff*, *enterprise*, *job*, *work*; this, in turn, reveals that CULTURE, in this context, has partially lost its original meaning of a process/product of refinement, as in the famous Arnoldian sense of “a pursuit of our total perfection, ...the best which has been thought and said in the world” (Arnold, 1869, p. viii) and

rather concerns a set of ideas/behaviours relating to a specific group in a specific context, like a workplace, company or organization – a new restricted meaning of “culture” (Gatto 2011; 2014).

This view is also supported by the prominence of the pattern *culture* + *within* computed by the Sketch Engine (542 occurrences, 4.2 per million) as an interesting pattern especially in UkWaC, which can well be interpreted in terms of the concept of “small culture” (Holliday 1999)

As to the more recent collocation of CULTURE with such verbs as *experience* and *explore*, we notice the emergence of the word *cultures* in the plural in their immediate co-text. These concordance lines clearly reflect the anthropological/ethnographic meaning of culture inaugurated between late nineteenth and early twentieth century by Franz Boas and other scholars. This pattern has a quite consistent collocation with adjectives like *new*, *different*, *origin*, *other*. It must be acknowledged, however, that – as far as UkWaC in particular is concerned, these concordance lines often originate in specialized sites dealing with the typically British experience of the gap year, a datum which relates to the choices made by the developers of ukWaC, who included academic sites (i.e. ac.uk sites) extensively in the crawl. Nonetheless, these occurrences testify to a radical shift in the meaning of culture, whereby culture is something to be experienced, rather than to be found in books (as Arnold would have argued), again something quite distant from its more traditional meaning.

2.2 Culture as modifier

As for the list of nouns modified by culture, this is opened, in the three corpora, by *shock*, a collocation which relates to a distinctively modern experience defined as *culture shock*, “the feelings of isolation, rejection, etc., experienced when one culture is brought into sudden contact with another, as when a primitive tribe is confronted by modern civilization” (*Collins Cobuild Dictionary*). Significantly, one of the most prominent collocates for *culture shock* in ukWaC is *reverse*, which originates in the phrase *reverse culture shock*, a form not yet attested in the BNC, probably because the experience itself had not yet been fully conceptualized. In this way the recent web corpora ukWaC and EnTenTen do not only provide evidence of a relatively new linguistic formation, but in doing so they point to

the emergence of a new social and psychological condition, resulting from a change in society itself. The very existence of a *reverse culture shock* is related to novel ways of experiencing mobility and migration, which entail continuous dislocations and relocations.

2.3 A culture of *

Finally a particularly significant pattern emerging from the word sketch for “**culture**” is the **pattern** *culture* + *of*, which seems to turn the word “culture” into an extraordinarily capacious and inclusive category that can be used for anything. And while some collocates might seem to confirm Stubbs’ intuition that the pattern has a relatively negative semantic prosody (1996, p. X), owing to collocations with words bearing negative connotations, there is ample evidence that the pattern can equally accommodate positive notions, like *secrecy/openness*, *blame/impunity*, *entrepreneurship/dependency*, etc. as the list of collocates reported below suggests:

pp. of -i	12095	1.6
secrecy	127	7.75
openness	153	7.62
blame	58	6.53
impunity	35	6.38
entrepreneurship	35	5.96
dependency	49	5.9
peace	217	5.72
consumerism	23	5.6
spin	36	5.41
complacency	20	5.33

Furthermore, the concordance lines for the pattern *a culture of* include many phrases in inverted commas, which seem to create a culture virtually ex-nihilo, as Barker (2003, xix) would argue, reducing culture to little more than an attitude, as in the examples reported below:

as them briefly. First there is often a **culture of** “we can do it ourselves”. Working in part : refusal to take up employment. Is a **culture of** “working poor” better than a culture of reflection, enquiry and dialogue, and a **culture of** “no blame” experimentation and challenge van of this hope, especially where a **culture of** “the here and now” leaves no room for down by and to everyone. Creating a **culture of** “allowing” seems to me to be the only opti insurance companies who criticise a **culture of** “bigger is best” when it comes to buying using social issues and engendering a **culture of** “collective responsibility” and the promoti assessment and the development of a **culture of** “continuous improvement”. The QSTG id “listeners are in danger of creating a **culture of** “disaster tourism” which actually makes

Here culture seems to have become a sort of neutral term that can keep the company of many different words: a culture of corruption, and a culture of accountability, a culture of violence, a

culture of peace, and even a culture of ‘buy now pay later’. As suggested by the various collocates for the pattern *a culture of* shown above, this lexico–grammar pattern really has the power of turning culture into a sort of *vox media*, a liquid modern word that can be used for anything.

3 Evidence from diachronic resources

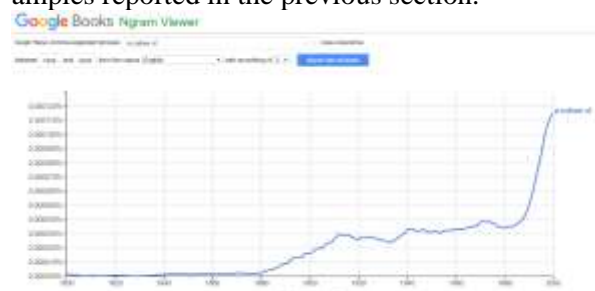
Starting from these preliminary observations on the lexico-grammar behaviour of the word “culture” in corpora representative of contemporary usage, a new research question emerged. To what extent can these patterns be considered as characteristic of contemporary usage? And if they are, how and when did they emerge? Is there any other information that could be gathered from the investigation of diachronic resources?

3.1. Google Books

A first attempt at answering these questions was to query the Google Books database, through an extremely limited and yet fascinating tool like Ngram Viewer, a tool which allows to read line charts representing n-grams i.e. continuous sequences from text, from the digitized books in the Google Book collections, in terms of frequency per year. The data can be accessed through a free web-based interface that enables relatively complex queries which support wildcards, POS-tagged search, case-sensitive queries, etc. For this reason, this is a tool commonly used in what has come to be known as “culturomics”, a research trend which aims “to observe cultural trends and subject them to quantitative investigation” on the basis of data obtained from Google books (Michel et al 2016). This approach is definitely controversial, especially from the perspective of corpus linguistics which is engaged in more theoretically sound and qualitatively reliable empirical research, and a very convincing overview of its limitations is found in McEnery and Baker (2016: 11-17). Nonetheless, in this specific case, information retrieved from such imperfect resources and limited tools could be still be used as indications, as fingers pointed to some interesting phenomena that might be worth being investigated in more detail with more appropriate resources. For instance, faced with the prominence and significance of the pattern “a culture of” as typical of the lexico-grammar profile of the word as described on the basis of data on contemporary usage, Ngram viewers was used to try and see whether the pattern had always

been there, or had it somehow emerged at a certain point in time.

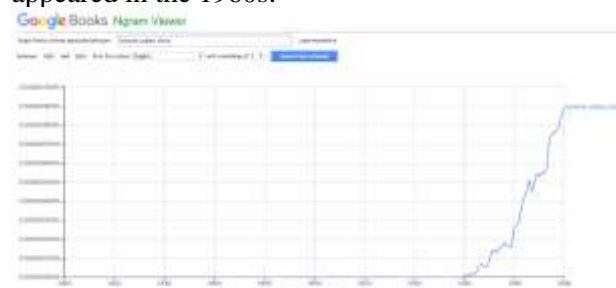
Indeed a search in Google Books using Ngram Viewer apparently suggests that the pattern emerged in the late 19th century, with most occurrences in the biological field, when it was referred to the recent discovery of bacteria. Anyway around the 1990s there was a dramatic surge in usage for this pattern, possibly connected with the growing tendency to use culture as a *vox media* devoid of any specific meaning as in the examples reported in the previous section.



Similarly, the emergence of the phrase “culture shock”, which seems to be prominent in contemporary corpora, can be located, with the help of Ngram Viewer, in a specific moment in history, in the late 50s:



Whereas “reverse culture shock” seems to have appeared in the 1960s:



Also interesting is the possibility to have a big picture in terms of changing behaviour of different lexico-grammar patterns. This is the case of the diverging fortunes of the two patterns experience + culture and understand + culture which seem to provide evidence of the fact that culture is more and more seen as something to be expe-

rienced than to be learnt or, as the verb suggests, cognitively appreciated and understood:



Besides using Ngram Viewer, the present research has also tried to profit from the interface for Google Books made available through Mark Davies well-known Corpora website to search the One Million Books and Fiction datasets. Based on the same Google Books data the interface was created by Mark Davies, Professor of Linguistics at Brigham Young University, and it is related to other large corpora made available through the same service. The system allows more refined queries than Google Books interface, and supports the comparison of the data in two different sections of the corpus. The interface available at Corpora BYU confirms at a glance that the collocation “culture shock” appeared between the 1950s and the 1970s and has dramatically grown in frequency since the 1980s. The same tools provide evidence of the emergence and decline of the collocation between *culture* and *refinement* around the publication of Arnold’s seminal *Culture and Anarchy* in the late 19th century.



3.2. Early English Books Online

A further attempt at casting a backward look to envisage the changing face of the world “culture” over time has been finally made by using data from the EEBO corpus available through Sketch Engine.

While limited in scope, these data provide clear evidence of the fact that the noun “culture” was not a particularly frequent in Written Early Modern English, as the EEBO corpus has only 2283 occurrences for this word (2.31 per million). Indeed, when CULTURE was used only in its original ‘agricultural’ meaning it was probably something which was not to be written about. Anyway data from EEBO makes us see firsthand the origins of its subsequent metaphorical meanings. Especially in the dataset for the period 1600-1699 the analysis of collocates for “culture” shows the emerging coexistence of the literal agricultural meaning and of a spiritual metaphorical meaning. It is at this stage that *cultivation* emerges as a meaningful collocate for “culture”, often in the such sentences as “cultivation of the minde”. However we have to wait until 1700-1799 for *civilization* to appear among the most salient collocates for “culture”. Which brings us back to the beginning of this story...

4 Conclusion

Using different resources to map such a complex research field, so as to obtain a general picture of significant patterns of usage in the evolution of language is certainly fascinating, but this is not enough. And it goes without saying that it is necessary to be extremely cautious before drawing conclusions, if any, from investigations like these. Anyway, the data analyzed confirm that there are resources and tools that can support the investigation of huge amount of data, pointing to interesting research areas to be analyzed with more refined *ad hoc* tools. In any case a rewarding exploration of these data from a cultural perspective can perhaps only come as the result of teamwork in the context of a multidisciplinary approach in the growing research field of the Digital Humanities.

Reference

- C. Barker (2003). *Cultural Studies: Theory and practice*. London: SAGE.
- Z. Bauman. (2011). *Culture in a Liquid Modern World*. Cambridge: Polity Press.
- M. Gatto (2014). *Web as Corpus. Theory and Practice*. London: Bloomsbury.
- M. Gatto (2011). “Sketches of CULTURE from the web. A preliminary study”. In G. Dimartino, L. Lombardo, S. Nuccorini (eds.). *Challenges for the 21st Century. Dilemmas, Ambiguities, Directions*. vol. II, Roma: Edizioni Q, pp. 275-283.

- A. Holliday (1999). "Small cultures". *Applied Linguistics* 20/2, pp.237-264.
- A. McEnery, H. Baker (2016). *Corpus Linguistics and 17th-Century Prostitution: Computational Linguistics and History*. London: Bloomsbury.
- J.B. Michel et al. (2011). "Quantitative analysis of culture using millions of digitized books". *Science*, 331(6014):176–182.
- R. Williams (1985). *Keywords. A Vocabulary of Culture and Society*. Oxford: Oxford University Press.

Resources and tools:

Early English Books Online,
<https://eebo.chadwyck.com/home>
Google Books Ngram Viewer,
<https://books.google.com/ngrams>
Google Books (British English),
<http://googlebooks.byu.edu/x.asp>
Sketch Engine, <https://www.sketchengine.co.uk/>
 [reference stub]

A Dataset of Real Dialogues for Conversational Recommender Systems

Andrea Iovine

Fedelucio Narducci

Marco de Gemmis

Department of Computer Science, University of Bari Aldo Moro, Italy

firstname.lastname@uniba.it

Abstract

Conversational Recommender Systems (CoRS) that use natural language to interact with users usually need to be trained on large quantities of text data. Since the utterances used during the interaction with a CoRS may be different depending on the domain of the items, the system should also be trained separately for each domain. So far, there are no publicly available datasets based on real dialogues for training the components of a CoRS. In this paper, we propose three datasets that are useful for training a CoRS in the movie, book, and music domains. These datasets have been collected during a user study for evaluating a CoRS. They can be used to train several components, such as the Intent Recognizer, Entity Recognizer, and Sentiment Recognizer.

1 Introduction

Recommender Systems (RS) are software systems that help people make better decisions (Jameson et al., 2015). They have become a fundamental tool for overcoming the *information overloading* problem, which is caused by the ever-increasing variety of information and products that people can access (Ricci et al., 2011). Choosing between such a large quantity of options is not easy, and this results in a decrease in the quality of the decisions. Recommender systems help alleviate the problem by providing personalized suggestions to users, based on their preferences.

Conversational Recommender Systems (CoRS) are a particular type of Recommender Systems,

that acquire the user's profile in an interactive manner (Mahmood and Ricci, 2009). This means that, in order to receive a recommendation, the system does not require that all the information is provided beforehand, but it guides the user in an interactive, human-like dialog (Jugovac and Jan-nach, 2017). Even though a CoRS can be implemented using several different interfaces, it is reasonable to think that an interaction based on natural language is suitable for the task. In particular, Digital Assistants (DA) such as Amazon Alexa, Google Assistant, or Apple's Siri are interesting platforms to deliver recommendations in a conversational manner. DAs, popularized with the diffusion of smartphones, are able to help users complete everyday tasks through a conversation in natural language. However, there is still a technological gap between CoRSs and DAs, as described in (Rafailidis and Manolopoulos, 2018). In particular, one of the main causes of that gap is the lack of labeled data. In fact, implementing a natural language-based interface for a CoRS is not easy, as it requires the use of several Natural Language Understanding (NLU) operations. For example, a basic conversational recommender needs at least three NLU components: an Intent Recognizer, an Entity Recognizer, and a Sentiment Analyzer. These components need to be trained on large quantities of real sentences, which may not always be available. The problem is worsened by the fact that each component may need to be trained separately for each different domain.

In this paper, we present three datasets that contain utterances used in real dialogues between users and a CoRS respectively in the movie, book, and music domains. These datasets can then be used to train the components of a new CoRS. To the best of our knowledge, this is the first time such a dataset of real dialogues is provided for the book and music domains, while there is already one example for the movie domain (Li et

Copyright 2019 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

al., 2018). The dataset is available at the following link¹.

Section 2 contains a literature review of datasets for training Question Answering and Conversational Recommender Systems. Section 3 illustrates the architecture of the CoRS that was used to collect the messages in the dataset. Section 4 describes in detail the three datasets, providing some statistics, and a small example of conversation.

2 Related Work

The problem of finding dialogues between humans and machines is not new, and in literature there are already some examples of conversational datasets that can be used to train a new conversational agent. Serban et al. (2015) published a literature survey of natural language datasets for CoRSs and Question Answering systems.

Dodge et al. (2015) presented a dataset for the evaluation of the performance of End-to-End Conversational Agents (CA), with a focus on the movie domain. End-to-End CAs use a single (usually deep learning-based) model to learn directly a response, given a user utterance. The objective of the dataset is to test the Question Answering and Recommendation abilities. The dataset is generated synthetically using data from MovieLens and Open Movie Database, and consists of 3.5 million training examples, covering 75,000 movie entities. This work differs from our contribution for several reasons. The most important difference is that our dataset is not used to learn what items to recommend, but rather, how to understand the user utterances. Thus, it is independent of the recommendation algorithm used. Furthermore, our dataset includes the book and music domains, and only uses real dialogues.

Braun et al. (2017) also developed two datasets for the evaluation of QA systems. The first dataset contains questions about public transport, and was collected through a Telegram chatbot. It consists of 206 manually annotated questions. The second dataset contains data collected from two StackExchange platforms, and consists of 290 questions and answers. The datasets were created to compare several NLP platforms in terms of their ability to recognize intents and entities for a QA system.

Asri et al. (2017) presented the Frames dataset, a corpus of 1369 dialogs generated through a Wizard-of-Oz setting. It was created to train

a goal-oriented information-retrieval Conversational Agent, that is able to find items in a database given a set of constraints. The main objective of the authors was to add memory capabilities to the CA. Each message is annotated using *frames*.

Suglia et al. (2017) propose an automatic procedure for generating plausible synthetic dialogues for movie-based CoRSs. This procedure takes in input a movie recommendation dataset (such as MovieLens), and turns each set of user preferences into a full conversation. The datasets created with this procedure can be used for training an End-to-End Conversational Recommender System. The purpose is then very similar to that of our contribution. However, we provide user-generated messages, rather than synthetic ones.

Kang et al. (2017) investigated how people interact with a natural language-based CoRS through voice or text. To do this, the authors developed a natural language interface, and integrated it in the MovieLens system. Then, they recorded the messages written (or spoken) by the users, i.e. what kinds of queries do they use. From the collected data, the authors classified three types of recommendation goals, and several types of follow-up queries. Data from 347 users was collected, and subsequently released. While interesting, this dataset does not specifically aim to train a new CoRS.

Li et al. (2018) developed ReDial, a dataset consisting of over 10,000 conversations, with the objective of providing movie recommendations. This dataset was conceived to train deep learning-based components, namely a sentiment analyzer and a recommendation algorithm. According to the authors, it is the only real-world, two-party conversational corpus for CoRSs. The dataset was used to train a movie-based CoRS that uses components based on deep learning, such as RNN for sentiment analysis, and an autoencoder for the recommendation. This dataset is probably the most similar to the one presented in this paper. However, it differs from it for two reasons: first, we provide datasets for three domains, rather than just the movie domain. Second, as stated earlier, our dataset is independent from the recommendation algorithm, and it only has the objective to understand how to maintain the conversation and acquire the user's preferences.

¹<https://github.com/aiovine/converse-dataset>

3 A Multi-Domain Conversational Recommender System

The dataset presented in this work is the result of the development and testing of a multi-domain Conversational Recommender System. The system is able to communicate with users via messages in natural language, both in acquiring their preferences, and providing suggestions. The recommendation process can be divided into two parts: a *preference acquisition phase* and a *recommendation phase*. In the first phase, the user is able to talk to the system freely. Preferences are expressed in the form of liked or disliked items. For example, a user can use a sentence like *"I love Stephen King, but I don't like The Shining"*. Multiple ratings can be given in the same sentence, and also can be given to different types of items (in this case, an author and his book). In case of ambiguity, the system may ask the user to clarify (*disambiguate*).

Once enough preferences are provided, the *recommendation phase* may start. This is done by asking for recommendations (e.g. *"What book can I read today?"*). During the recommendation phase, the system suggests a set of items, each of which can be rated positively or negatively by the user. A *critiquing* function also allows the user to criticize some aspects of the suggested item (e.g. *"I like this movie, but I don't like Mel Gibson"*). It is also possible to ask for more details about the recommended item, for a trailer/preview, or for an explanation (e.g. *"Why did you suggest this song?"*).

Our CoRS uses a *modular* architecture, that is made up of several components, each with a specific responsibility. It was deployed as a Telegram chatbot, but it can be easily ported to any other messaging platform, such as Facebook Messenger or any others. The components in question (as seen in Figure 1) are:

- **Dialog Manager:** This component is responsible for maintaining a conversation with the user in a persistent way. It decides what action should be performed given the user intent, invokes the other components, aggregates their outputs, and produces the final response.
- **Intent Recognizer:** This component is responsible for understanding the action that the user is requesting. For example, when the

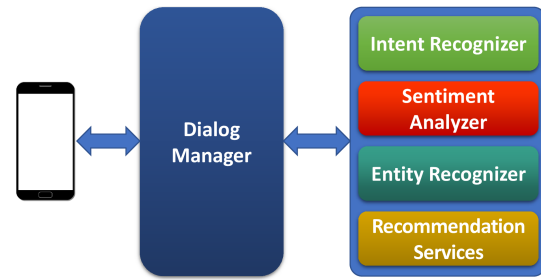


Figure 1: Architecture of the CoRS

user says *"I like Michael Jackson"*, the *preference* intent is recognized. The Intent Recognizer is powered by DialogFlow².

- **Entity Recognizer:** This component is responsible for recognizing entities mentioned by the user. Given the previous example, it is able to recognize *Michael Jackson* as an entity mention. It exploits Wikidata³, and does not require any training. This component was developed in-house.
- **Sentiment Analyzer:** This component is responsible for recognizing the user's sentiment on the recognized entities. Given the previous example, it recognizes a positive rating for *Michael Jackson*. This component is developed using Stanford CoreNLP⁴.
- **Recommendation Services:** This component is responsible for the recommendation algorithm. In particular, we use a Content-Based recommender based on the PageRank with priors.

4 ConverseSE Datasets

In this section, we describe the main features of the dataset and the process that we used to build it. The dialogues were recorded during an experimental session, in which participants were asked to interact with three CoRSs, each for a specific domain (movie, books, and music). During the preference acquisition phase, each participant wrote some positive/negative ratings. After that, participants were asked to request a recommendation, and then evaluated five recommended items. Finally, users asked the system to view their profiles. From this experiment, we collected

²<https://dialogflow.com/>

³<https://www.wikidata.org>

⁴<https://stanfordnlp.github.io/CoreNLP/>

	Movie	Book	Music
#Users	149	56	56
#Messages	5318	1862	2096
#Messages per user	35.7	33.3	37.4
#Preference messages	2172	734	1011
#Recomm. requests	456	369	144
%Liked (Preference)	89.8	91.6	93.5
%Disliked (Preference)	10.2	8.40	6.54
%Liked (Recomm.)	77.6	77.7	73.2
%Disliked (Recomm.)	22.4	22.3	26.8
%Critiquing	1.6	0.0	0.42
%Details requests	11.4	3.6	2.08
%Preview requests	6.98	1.7	0.625
%Explanation requests	10.5	1.49	2.5
%To check	39.6	28.8	26.0

Table 1: ConVerSE dataset statistics

5,318 messages for the movie domain, 1,862 for the book domain, and 2,096 for the music domain.

For each message, we collected the user’s utterance, the intent recognized by the system, unique IDs for the user and the message, a timestamp, a list of contexts, a list of recognized items, and a set of actions. We chose not to include the system’s responses in the dataset, since they are generated via a template. Instead, we report a set of *actions* that together map the reaction of the system to the user message, and the current *status* of the conversation. For example, the *recommendation* action means that the user is in the recommendation phase. The *question* action means that the system responded to the user by asking a question (i.e. requesting a disambiguation, or asking the user to rate a recommended item). Finally, the *finished_recommendation* actions signal that the message concludes a recommendation phase. An item is included in the list of recognized items only after it was correctly disambiguated (if a disambiguation was needed). For example, if the user writes “I like Tom Cruise”, the system responds “You said that you like Tom Cruise, can you be more specific? Possible values are: producer, actor”. Only when the user responds to this question the item will be recorded as recognized in the dataset. For each recognized item, we record its Wikidata ID, and a symbol that identifies the rating (‘+’ for positive, ‘-’ for negative).

We applied some heuristics for improving the quality of the data. In particular, the objective is to understand whether the recognized intents and

entities are correct. To do this, each conversation was split into *tasks*, where a task is defined as a sequence of messages with a specific goal. For each task, we observed whether it terminated successfully, or an anomaly occurred. Some examples of tasks that are completed correctly are:

- A preference message, followed by one or more disambiguations;
- A recommendation request, followed by one or more preferences to the recommended item, requests for details and explanations;
- A request for showing the profile.

Some examples of tasks that are not completed correctly are:

- Any task containing a *fallback* intent (means that the intent was not recognized)
- Tasks in which the user asks to skip a disambiguation request, or to stop the recommendation phase;
- Tasks in which an unexpected intent is found (e.g. preference to an unrelated item during the recommendation phase).

For each message, we added a field called *toCheck*. This field is set to *false* if the message is part of a completed task, *true* otherwise. In the latter case, it is advised to manually check the correctness of the intent.

Table 4 describes some statistics extracted from the dataset. More precisely, we collected the number of users and messages, the number of preference messages and recommendation requests, the average number of messages per user, the percentage of liked and disliked items (both in the preference acquisition and recommendation phases), the percentage of critiquing, details, preview and explanation requests (over all recommended items), and the percentage of messages for which *toCheck* is equal to *true*. For privacy reasons, we anonymized the dialogues by replacing the original Telegram user ID with a numerical index.

4.1 Example of conversation

In this section, we describe a small example of a conversation between a user and the movie-based instance of the CoRS. For each message in Table 2, we describe the utterance along with the main features, in order to make the underlying dialog model more understandable. The following paragraphs contain a short explanation for each message. For brevity reasons, the example contains

#	Message	Intent	Recognized objects	Status
1	I like the avengers	preference		question, disambiguation
2	The Avengers (2012)	preference - disambiguation	Q182218+	
3	Suggest some film	request_recommendation		recommendation, question
4	I like this movie	request_recommendation - preference	Q14171368+	recommendation, question
5	Why do you suggest this movie?	request_recommendation - why		recommendation, question
6	I love it, but I don't like director	request_recommendation - yes_but	Q220192+	recommendation, question
7	Can you show my preferences	show_profile		

Table 2: Short example of conversation in the movie dataset

messages from different conversations, in order to show more intents with fewer messages.

1. The user has provided a preference during the preference acquisition phase. The recognized intent is then *preference*. Since there are multiple movies matching with *The Avengers*, further disambiguation is required. This is indicated via the *question* and *disambiguation* actions.

2. The user has answered the disambiguation request, by specifying that he/she means the movie "The Avengers (2012)". This is associated with the *preference - disambiguation* intent. Note that only now the movie was included in the *recognized objects* field.

3. When the user sends this message, a new recommendation phase is started. The corresponding intent is *request_recommendation*. When this happens, the system proposes a movie that will be rated by the user. The actions *question* and *recommendation* are used to indicate that the CoRS is expecting a rating from the user.

4. When the user provides a rating to a recommended entity (in this case, *I like this movie*), the *request_recommendation - preference* intent is used. The rating of the recommended item is also registered in the *recognized objects* field. The *recommendation* and *question* actions in this case signify that the system responds by presenting another recommended movie to rate.

5. In this case, the user asks an explanation for the recommended item. The *request_recommendation - why* is used in this case. After the explanation was given, the system asks again to rate the movie, as evidenced by the recorded actions.

6. Here, the user provides the rating, but also criticizes the recommendation, by adding a negative rating to the director of the recommended movie (previously mentioned as *critiquing*). The *request_recommendation - yes_but* intent is used in this case. Our CoRS requests an additional confir-

mation when associating a property (i.e. *director*) to a recommended item, however it could be ignored when training a new CoRS.

7. In this case, the user is requesting to see his/her profile, as indicated by the *show_profile* intent. This can be optionally followed by requests for editing or deleting the profile.

5 Conclusions

In this paper, we presented three datasets that contain real user messages sent to Conversational Recommender Systems in the movie, book, and music domains. The datasets can be used to train a new CoRS to detect the intents, and with a few modifications, also to recognize entities and sentiments. The size of the data that we provide may not be sufficient to train deep learning-based End-to-End conversational recommendation models. However, this is outside the scope of our work: as stated in the previous sections, the aim of our datasets is to learn a conversational recommendation dialog model, independently from the actual recommendation algorithm. In any case, we believe that this is the first time that a dataset for training CoRSs in the book and music domain is released. Also, we believe that this is a good starting point for the release of further conversational datasets in multiple domains.

We propose, as future work, to expand the datasets, by collecting more messages, in more domains. We will also explore the possibility to use our datasets to evaluate new CoRSs.

References

Layla El Asri, Hannes Schulz, Shikhar Sharma, Jeremie Zumer, Justin Harris, Emery Fine, Rahul Mehrotra, and Kaheer Suleman. 2017. Frames: A Corpus for Adding Memory to Goal-Oriented Dialogue Systems. *arXiv:1704.00057 [cs]*, March. arXiv: 1704.00057.

Daniel Braun, Adrian Hernandez-Mendez, Florian

- Matthes, and Manfred Langen. 2017. Evaluating Natural Language Understanding Services for Conversational Question Answering Systems. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 174–185, Saarbrücken, Germany. Association for Computational Linguistics.
- Jesse Dodge, Andreea Gane, Xiang Zhang, Antoine Bordes, Sumit Chopra, Alexander Miller, Arthur Szlam, and Jason Weston. 2015. Evaluating Prerequisite Qualities for Learning End-to-End Dialog Systems. *arXiv:1511.06931 [cs]*, November. arXiv: 1511.06931.
- Anthony Jameson, Martijn C. Willemsen, Alexander Felfernig, Marco de Gemmis, Pasquale Lops, Giovanni Semeraro, and Li Chen, 2015. *Human Decision Making and Recommender Systems*, page 611648. Springer US.
- Michael Jugovac and Dietmar Jannach. 2017. Interacting with Recommenders Overview and Research Directions. *ACM Transactions on Interactive Intelligent Systems*, 7(3):1–46, September.
- Jie Kang, Kyle Condiff, Shuo Chang, Joseph A. Konstan, Loren Terveen, and F. Maxwell Harper. 2017. Understanding How People Use Natural Language to Ask for Recommendations. In *Proceedings of the Eleventh ACM Conference on Recommender Systems - RecSys '17*, pages 229–237, Como, Italy. ACM Press.
- Raymond Li, Samira Kahou, Hannes Schulz, Vincent Michalski, Laurent Charlin, and Chris Pal. 2018. Towards Deep Conversational Recommendations. page 17.
- Tariq Mahmood and Francesco Ricci. 2009. Improving recommender systems with adaptive conversational strategies. In *Proceedings of the 20th ACM conference on Hypertext and hypermedia - HT '09*, page 73, Torino, Italy. ACM Press.
- Dimitrios Rafailidis and Yannis Manolopoulos. 2018. The Technological Gap Between Virtual Assistants and Recommendation Systems. *arXiv:1901.00431 [cs]*, December. arXiv: 1901.00431.
- Francesco Ricci, Lior Rokach, and Bracha Shapira, 2011. *Introduction to recommender systems handbook*. Springer US.
- Iulian Vlad Serban, Ryan Lowe, Peter Henderson, Laurent Charlin, and Joelle Pineau. 2015. A Survey of Available Corpora for Building Data-Driven Dialogue Systems. *arXiv:1512.05742 [cs, stat]*, December. arXiv: 1512.05742.
- Alessandro Suglia, Claudio Greco, Pierpaolo Basile, Giovanni Semeraro, and Annalina Caputo. 2017. An Automatic Procedure for Generating Datasets for Conversational Recommender Systems. page 2.

Quanti anni hai? Age Identification for Italian

Aleksandra Maslennikova*, Paolo Labruna*, Andrea Cimino[◊], Felice Dell’Orletta[◊]

* Università di Pisa

a.maslennikova@studenti.unipi.it

pielleunipi@gmail.com

[◊]Istituto di Linguistica Computazionale “Antonio Zampolli” (ILC–CNR)

ItaliaNLP Lab - www.italianlp.it

{andrea.cimino, felice.dellorletta}@ilc.cnr.it

Abstract

English. We present the first work to our knowledge on automatic age identification for Italian texts. For this work we built a dataset consisting of more than 2.400.000 posts extracted from publicly available forums and containing authorship attribution metadata, such as age and gender. We developed an age classifier and performed a set of experiments with the aim of evaluating the possibility of assigning the correct age of an user and which information is useful to tackle this task: lexical or linguistic information spanning across different levels of linguistic descriptions. The performed experiments show the importance of lexical information in age classification, but also that exists writing style that relates to the age of an user.

Italiano. *In questo articolo presentiamo il primo lavoro a nostra conoscenza sul riconoscimento automatico dell’età per la lingua italiana. Per condurre il lavoro abbiamo costruito un dataset composto da più di 2.400.000 di post estratti da forum pubblici e associati a informazioni rispetto all’età e al genere degli autori. Abbiamo sviluppato un sistema di classificazione dell’età dello scrittore di un testo e condotto una serie di esperimenti per valutare se è possibile definire l’età e attraverso quali informazioni estratte dal testo: lessicali o di descrizione linguistica a diversi livelli. I risultati ottenuti dimostrano l’importanza del lessico nella classificazione, ma anche l’esistenza di uno stile di scrittura correlato all’età.*

1 Introduction

Social media platforms such as Facebook, Twitter and public forums allow users to communicate and share their opinions and to build social relations. The proliferation of such platforms allowed the scientific community to study many communication phenomena such as the analysis of the sentiment (Pak et al., 2010) or irony (Hernández Farías et al., 2016). Another related research field is the “author profiling” one, where the features that allow to discriminate age, gender, or native language of a person are analyzed. These studies are conducted both for forensic and marketing reasons, since the classification of these characteristics allow companies to better focus their marketing campaigns. In the author profiling scenario, many are the studies conducted by the scientific community, that were generally focused on English and Spanish language. The majority of these studies were performed in PAN ¹ (Rangel et al., 2016), a lab at CLEF ² that holds each year and in which many shared tasks related to the “authorship attribution” research topic are run. In these shared tasks participants were asked to identify the gender or the age using manually annotated training data from social media platforms. Among the most successful approaches proposed by participants the ones that achieved the best results (op Vollenbroek et al., 2016), (Modaresi et al., 2016) are based on SVM classifiers exploiting a wide variety of lexical and linguistic features, such as word n-grams, part-of-speech, and syntax. Only recently deep learning based approaches were proposed and have showed very good results especially when dealing with multi-modal data, i.e. text and images posted on Twitter (Takahashi et al., 2018).

In the present work we tackle a specific author-

Copyright ©2019 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

¹<https://pan.webis.de/>

²<http://www.clef-initiative.eu/association/steering-committee>

ship attribution task: the age detection for the Italian language. To our knowledge, this is the first time that such task is performed on Italian. For this reason, we built a multi-topic corpus, developed a classifier which exploits a wide range of linguistic features, and conducted several experiments to evaluate both the newly introduced corpus and the classifier.

The main contributions of this work are: *i*) an automatically built corpus for the age detection task for the Italian language; *ii*) the development of an age detection system; *iii*) the study of the impact of linguistic and lexical features.

2 Dataset construction

With the aim of building an automatic dataset from the web, we needed a set of Italian texts with the age of authors publicly available. Nowadays collecting this information is a challenging task, since the majority of the available platforms, for the sake of privacy, prefer not to make the user's age public. So, first-of-all, we had to find a website with such data. We choose the ForumFree platform³ which allows users to create their own forums without any coding skills, using an existing template. Having all the forums based on the same templates makes them perfect for automated crawling. We extracted all the posts of the users that decided to show publicly their age. We tried to collect the data from the top 200 most active forums. Not all the forums had users with all the user information filled and, in the end of the processes, we fetched messages from 162 different forums. Since our goal was to build a corpus with author profiling purposes, and such task is very difficult with very small comments, we selected only posts with a minimum length of 20 words.

Another problem we faced is that users are not age-balanced in the forums: for example, anime dedicated forum have mostly users aged under 35. Another example are cars dedicated forums, where usually users are more mature with respect to anime forums. Only a couple of forums have very balanced information, which usually is the best data for training machine learning based classifiers. For this reason, we decided to group the forums by their topics, because in this scenario it is more probable to gather enough textual data for each age gap. We manually looked the content of all forums and assigned the topic for each

one of them. We didn't have a preassigned settled list of possible topics. Instead, we were adding them in the process. For example, if we have an entire forum which discusses about only watches, we wouldn't assign some general "Hobby" tag, but we would create a special group "Watches" specifically for this forum.

At the end of the collection process, we collected 2.445.012 posts from 7.023 different users and 162 forums, that we divided in 30 different topic groups. All the information regarding the dataset are shown in Table 1.

3 The Age classifier

We implemented a document age classifier that operates on morpho-syntactically tagged and dependency parsed texts. The classifier exploits widely used lexical, morpho-syntactic and syntactic features that are used to build the final statistical model. This statistical model is finally used to predict the age range of unseen documents. We used linear SVM implemented in LIBLINEAR (Rong-En et al., 2008) as machine learning algorithm. The input documents were automatically POS tagged by the Part-Of-Speech tagger described in (Cimino and Dell'Orletta, 2016) and dependency-parsed by the DeSR parser (Attardi et al., 2009).

3.1 Features

Raw and Lexical Text Features

Word n-grams, calculated as presence or absence of a word n-gram in the text.

Lemma n-grams, calculated as the frequency of each lemma n-gram in the text and normalized with respect to the number of tokens in the text.

Morpho-syntactic Features

Coarse and fine grained Part-Of-Speech n-grams, calculated as the logarithm of the frequency of each coarse/fine grained PoS n-gram in the text and normalized with respect to the number of tokens of the text.

Syntactic Features

Linear dependency types n-grams, calculated as the frequency of each dependency n-gram in the text with respect to the surface linear ordering of words and normalized with respect to the number of tokens in the text.

Hierarchical dependency types n-grams calculated as the logarithm of the frequency of each hierarchy dependency n-gram in the text and nor-

³<https://www.forumfree.it/?wiki=About>

Topic		≤20	21-30	31-40	41-50	51-60	≥61
Cars	Users	36	158	187	209	158	45
	Posts	6056	50281	46746	62002	48939	15867
Bicycles	Users	10	11	12	35	25	1
	Posts	2056	2284	5532	13418	16959	6
Smoking	Users	3	52	78	69	46	18
	Posts	7	21399	41470	38149	17981	4742
Anime/Manga	Users	392	438	142	62	16	6
	Posts	60367	99165	39939	29086	3873	228
Role playing	Users	115	104	14	8	6	7
	Posts	22953	40652	3893	3945	534	2060
Gaming	Users	235	358	113	131	48	7
	Posts	54584	81535	20379	20055	4560	1323
Spirituality	Users	11	25	21	13	11	2
	Posts	336	1427	1342	1095	1517	965
Aesthetic medicine	Users	7	36	27	29	17	1
	Posts	1345	6135	11767	8208	3384	1
Sport	Users	215	338	192	136	52	24
	Posts	82495	310220	158382	103027	34627	16084
Culinary	Users	0	1	4	10	4	4
	Posts	0	52	10130	2414	747	438
Pets	Users	10	21	11	4	2	3
	Posts	4307	13222	7357	2592	5383	10353
Celebrities	Users	21	76	26	24	17	4
	Posts	548	21114	5820	6150	3139	1248
Politics	Users	0	2	4	10	6	0
	Posts	0	330	2801	3548	576	0
Different topics	Users	52	45	34	43	34	15
	Posts	9453	12000	21667	16316	4759	24418
Fishing	Users	11	57	79	62	30	5
	Posts	3040	14805	24306	17131	13155	8356
Institution community	Users	6	6	0	2	5	1
	Posts	13	12	0	18	11130	4364
Rail transport modelling	Users	0	6	7	5	5	1
	Posts	0	3597	2289	999	2470	751
Culture	Users	4	10	4	7	4	0
	Posts	1855	560	653	1174	219	0
Tourism	Users	0	2	2	4	1	2
	Posts	0	16	10	1378	2	14
Sexuality	Users	11	31	18	10	2	1
	Posts	185	2540	8201	1421	7	1179
Metal Detecting	Users	25	34	78	121	55	11
	Posts	7750	9830	19299	31288	16547	3529
Music	Users	12	25	15	0	0	0
	Posts	8731	15720	5276	0	0	0
Parenting	Users	1	4	1	1	0	0
	Posts	719	2250	626	420	0	0
Technologies	Users	37	47	12	4	8	5
	Posts	185	266	431	26	19	23
Nature	Users	5	9	10	6	6	2
	Posts	998	1304	3653	2171	292	10
Religion	Users	0	5	6	1	0	0
	Posts	0	2618	4125	896	0	0
Films	Users	25	26	10	5	1	2
	Posts	9476	6135	503	43	4	2477
Psychology	Users	12	14	2	0	1	2
	Posts	291	912	44	0	1	11
Gambling	Users	0	3	3	10	11	7
	Posts	0	458	134	364	715	274
Watches	Users	29	153	317	302	109	32
	Posts	5158	52623	114074	101869	50243	18085

Table 1: Distribution of number of users and posts per age gap in different topics in the corpus

malized with respect to the number of tokens in the text. In addition to the dependency relationship, the feature takes into account whether a node is a left or a right child with respect to its parent.

4 Experiments

In order to test the corpus and the classifier, we performed a set of experiments. The experiments were devised in order to test real-word scenarios where 1) we were interested to classify a set of posts written by a single user rather than a single post; 2) we always classified unseen users, i.e. no training data was available for such users. For these reasons, we merged all the posts of a single user in the original corpus in a single document. We then considered only the users that wrote a minimum of 200 tokens and limited the final merged document to a 'soft' limit of 1000 tokens for each user. When the soft limit was exceeded, we included the whole post that exceeded the soft limit. The described procedure allows training and test splits to never contain the same user. For the age detection tasks, similarly as in (Rangel et al., 2016), we considered age-splits as the classification classes. More precisely, we took into account two different age group splits: the first one, which we will refer with the name *5-class*, in which we split the documents in 5 different age groups: 20-29, 30-39, 40-49, 50-59, 60-69. The second age group split, which we will refer with the name *2-class*, is composed by the following age group splits: ≤ 29 , $\geq 50-69$ (excluding all the documents written by users that did not belong to these age groups). We conducted two different kind of experiments. In the first experiment (*in-domain*), we evaluated the performance of the classifier on in-domain texts, more precisely we selected three different topics starting from the main corpus and on each of the topics we trained the classifier on the 80% of the data, and evaluated the performance of the classifier on the remaining 20%. For this experiment we choose the the following domains: Sports, Watches and Cars. In the second experiment (*out-domain*) we trained the classifier on the all the 3 topics used for the *in-domain* experiments and evaluated the performance of the classifier on other 3 different topics (Smoking, Celebrities, Metal Detecting).

In addition, we devised 3 different machine learning models based on 3 different sets of features. The first one (*Lexicon*), which uses only

word and lemmas features, the second one (*Syntax*), which uses only the morpho-syntactic and syntactic features. Finally, the last model (*All*), which uses both the lexical, morpho-syntactic and syntactic features. We considered as baseline model a classifier which predicts always the most frequent class.

4.1 Results

Tables 2 and 3 report the results achieved by the classifier for the in-domain and out-domain experiments respectively. For what concerns all the experiments, we can notice that the results achieved by our classifier are higher than the baseline results, showing that there are features that are able to discriminate among the considered classes. The in-domain results show that the lexical features are the ones that have the most discriminative power with respect to the syntax ones. The f-score achieved by the *lexicon* model is 3-4 times better than the baseline in the 5-class setting, and 2 times better in average in the 2-class setting. The *syntax* model shown very good results but, as expected, lower than the results achieved by the lexicon model. This is an important result since it shows that syntax and morpho-syntax are relevant characteristics in each age-group, both in the 5-class and 2-class settings. Surprisingly, the *All* model didn't show in any experiment an increase in classification performance. The classification patterns revealed in the in-domain experiments are similarly shown also in the out-domain experiments. The results achieved in this setting as expected are lower than results achieved in the in-domain settings. The 5-class experiments show a drop in performance achieved by the considered learning models of 8-10% f-score points in average w.r.t. to the in-domain experiments. When we move to the 2-class experiments, no significant drop in performance is noticed. This shows that in case of domain shifting, the machine learning models are still able to well discriminate between young and aged people.

Figures 1 and 2 report the confusion matrices of the in-domain and out-domain experiments using the 5-class age-groups. More precisely, the in-domain confusion matrix is obtained by training the *All* model on all the three training in-domain topics and testing the model on the respective testset (f-score: 0.47). Similarly, the out-domain confusion matrix is obtained by training

	5-class				2-class			
Topic	Baseline	Lexicon	Syntax	All	Baseline	Lexicon	Syntax	All
Sport	0.27	0.45	0.42	0.48	0.74	0.74	0.75	0.75
Watches	0.19	0.43	0.35	0.42	0.44	0.85	0.75	0.83
Cars	0.12	0.54	0.34	0.45	0.47	0.87	0.77	0.84

Table 2: Results achieved in the in-domain experiments in terms of f-score

	5-class				2-class			
Topic	Baseline	Lexicon	Syntax	All	Baseline	Lexicon	Syntax	All
Smoking	0.14	0.30	0.25	0.32	0.42	0.79	0.68	0.79
Celebrities	0.33	0.45	0.39	0.47	0.62	0.83	0.73	0.81
Metal Detecting	0.21	0.36	0.27	0.34	0.52	0.80	0.66	0.78

Table 3: Results achieved in the out-domain experiments in terms of f-score

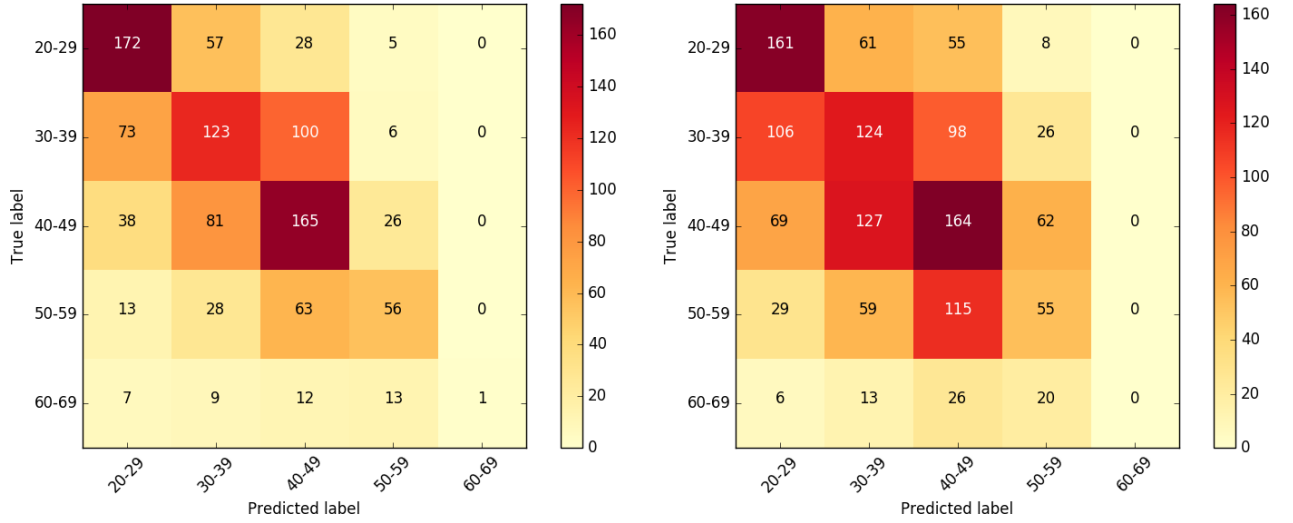


Figure 1: Confusion matrix calculated on the documents belonging to the in-domain topics

Figure 2: Confusion matrix calculated on the documents belonging to the out-domain topics

the *All* model on all the in-domain topics (including the test-sets), and testing the model on the out-domain documents of the selected 3 topics. As it can be seen, the errors both on the in-domain and out-domain experiments show very good performances of the classifier, i.e., in case of errors, usually it makes a mistake of a range of ± 10 years. Such results show also that the automatically built corpus is a very useful resource for the age classification task. Finally, it is interesting to notice that the most correct predicted classes are the ranges 20-29 and 40-49, both in the in-domain and out-domain settings, while the worst predicted class in both experiments is the 60-69 age range, most probably because is the most underrepresented class in the training set.

5 Conclusions

We presented the first automatically built corpus for the age detection task for the Italian language.

By exploiting the publicly available information on the FreeForum platform, we built a corpus consisting of more than 2.400.000 posts and 7.000 different users containing the user's age information. The first experiments performed through a machine learning based classifier that uses a wide range of linguistic features showed promising results in two different range classification tasks both in the in-domain and out-domain settings. The conducted experiments show that lexicon plays a fundamental role in the age classification task both in in-domain and out-domain scenarios. Lastly, the experiments shown that the corpus, even though if automatically generated, is suitable for real-world applications. We plan to release the full corpus as soon as privacy and legal issues will be fully investigated.

Acknowledgments

This work was partially supported by the 2-year project ARTILS, Augmented RealTime Learning for Secure workspace, funded by Regione Toscana (BANDO POR FESR 2014-2020).

Takumi Takahashi, Takuji Tahara, Koki Nagatani, Yasuhide Miura, Tomoki Taniguchi and Tomoko Ohkuma. 2016. *Text and image synergy with feature cross technique for gender identification*. In Working Notes of CLEF 2018 - Conference and Labs of the Evaluation forum, Avignon, France, 10 - 14 September, 2018.

References

- Giuseppe Attardi, Felice Dell’Orletta, Maria Simi and Joseph Turian. 2009. *Accurate dependency parsing with a stacked multilayer perceptron*. In Proceedings of the 2nd Workshop of Evalita 2009. December, Reggio Emilia, Italy.
- Andrea Cimino and Felice Dell’Orletta. 2016. *Building the state-of-the-art in POS tagging of italian tweets*. In Proceedings of Third Italian Conference on Computational Linguistics (CLiC-it 2016) & Fifth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA), December 5-7.
- Delia Irazú Hernández Farías, Viviana Patti and Paolo Rosso. 2016. *Irony Detection in Twitter: The Role of Affective Content*. In ACM Transactions on Internet Technology (TOIT), Volume 15, number 3.
- Pashutan Modaresi, Matthias Liebeck and Stefan Conrad. 2016. *Exploring the Effects of Cross-Genre Machine Learning for Author Profiling in PAN 2016*. In Working Notes of CLEF 2016 - Conference and Labs of the Evaluation forum, Évora, Portugal, 5-8 September, 2016.
- Francisco Manuel Rangel Pardo, Paolo Rosso, Ben Verhoeven, Walter Daelemans, Martin Potthast and Benno Stein. 2016. *Overview of the 4th Author Profiling Task at PAN 2016: Cross-Genre Evaluations*. In Working Notes of CLEF 2016 - Conference and Labs of the Evaluation forum, Évora, Portugal, 5-8 September, 2016.
- Mart Busger op Vollenbroek, Talvany Carlotto, Tim Kreutz, Maria Medvedeva, Chris Pool, Johannes Bjerva, and Hessel Haagsma and Malvina Nissim. 2016. *Gronup: Groningen user profiling*. In Working Notes of CLEF 2016 - Conference and Labs of the Evaluation forum, Évora, Portugal, 5-8 September, 2016.
- Alexander Pak and Patrick Paroubek. 2010. *Twitter as a corpus for sentiment analysis and opinion mining*. In Proceedings of the International Conference on Language Resources and Evaluation, LREC 2010, 17-23 May 2010, Valletta, Malta
- Fan Rong-En, Chang Kai-Wei, Hsieh Cho-Jui, Wang Xiang-Rui and Lin Chih-Jen. 2008. *LIBLINEAR: A library for large linear classification*. Journal of Machine Learning Research, 9:1871–1874.

Multi-task Learning Applied to Biomedical Named Entity Recognition Task

Tahir Mehmood^{1,2}, Alfonso Gerevini², Alberto Lavelli¹, and Ivan Serina²

¹Fondazione Bruno Kessler, Via Sommarive, 18 - 38123 Trento, Italy
{t.mehmood, lavelli}@fbk.eu

²Department of Information Engineering, University of Brescia, Italy
{t.mehmood, alfonso.gerevini, ivan.serina}@unibs.it

Abstract

Recent deep learning techniques have shown significant improvements in biomedical named entity recognition task. However, such techniques are still facing challenges; one of them is related to the limited availability of annotated text data. In this perspective, with a multi-task approach, simultaneously training different related tasks enables multi-task models to learn common features among different tasks where they share some layers with each other. It is desirable to use stacked long-short term memories (LSTMs) in such models to deal with a large amount of training data and to learn the underlying hidden structure in the data. However, the stacked LSTMs approach also leads to the vanishing gradient problem. To alleviate this limitation, we propose a multi-task model based on convolution neural networks, stacked LSTMs, and conditional random fields and use embedding information at different layers. The model proposed shows results comparable to state-of-the-art approaches. Moreover, we performed an empirical analysis of the proposed model with different variations to see their impact on our model.

1 Introduction

Named entity recognition (NER) consists in recognizing chunks of text and labelling them with predefined categories (e.g., person name, organization, location, etc). NER is an information extraction task and has many applications for instance in co-reference resolution, question an-

swering systems, machine translation, information retrieval etc (Chieu and Ng, 2002). NER is also performed on biomedical data where it involves recognizing biomedical concepts (e.g., cell, chemical, drug, disease, etc) and classifying them into predetermined categories. This is referred as biomedical named entity recognition (BioNER). Large amounts of medical data are available as free, unstructured text and the quantity of annually generated biomedical data like books, scientific papers, and other publications makes it challenging for physicians to stay up to date.

Moreover, biomedical documents are more complex than normal texts and the names of the entities show peculiar characteristics. Long multi-word expressions (*10-ethyl-5-methyl-5,10-dideazaaminopterin*), ambiguous words (*TNF alpha* can be used for both DNA and Protein) (Gridach, 2017), spelling alternations (e.g., *10-Ethyl-5-methyl-5,10-dideazaaminopterin* vs. *10-EMDDA*) make the BioNER task even more challenging (Giorgi and Bader, 2018). BioNER is also an important preliminary task for other tasks like the extraction of relations between entities (e.g., chemical induced disease relation, drug-drug interaction, ...).

Recent applications of deep learning in BioNER minimize manual feature engineering process and at the same time produce promising results. Deep learning is now the state-of-the-art technique but, due to the complex structure of biomedical text data, deep learning models have difficulties in performing efficiently. Moreover, these systems require large amounts of input data while the available annotated biomedical data are not enough to train these systems effectively. Manually generating annotated biomedical text data is an expensive and time-consuming job. In order to address this limitation, one solution is to take advantage of a multi-task learning approach. Multi-task learning (MTL) involves training simultaneously different

Copyright 2019 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0)

but related tasks together. Such an approach has shown significant improvements in different fields.

In this paper, we propose a multi-task model (MTM-CW) using convolutional neural networks (CNN) (dos Santos and Guimarães, 2015), stacked layers of Bidirectional long-short term memories (BiLSTM), and conditional random fields (CRFs). Furthermore, we have conducted an empirical analysis of the impact of different word input representation to our model.

The rest of the paper is organized as follows; Section 2 gives a brief background of the multi-task learning followed by Section 3 where our multi-task model (MTM-CW) is discussed. Experimental setup is presented in Section 4 which is followed by the results and discussion (Section 5). Section 6 concludes and presents possible future research directions.

2 Multi-task Learning

In general, deep learning model performance highly depends on the amount of annotated data available. It performs better when large amount of data is available. Unfortunately, in different biomedical tasks only a limited quantity of annotated text data is available and in this case deep learning models have difficulties to generalize well. Moreover, manually annotating new data is a time consuming job and this issue can be reduced by using two methods: transfer learning and multi-task learning.

In transfer learning, the model is partially trained on an auxiliary task and is then reused on the main task. This enables the model to fine tune the weights of the layers which are learned during the training on the auxiliary task. This helps the model to generalize well on the main task, which implies learning generalized features between the auxiliary and the main task. This method learns and transfers shallow features from one domain to another domain (Luong et al., 2016).

On the other hand, multi-task learning (MTL) is an approach where different related tasks are trained simultaneously. Unlike transfer learning, multi-task learning optimizes the model under construction concurrently. In MTL approach, some of the layers in the model are shared among different tasks while keeping some layers task-specific. Training jointly on related tasks helps the multi-task model to learn common features among different tasks by using shared layers (Bansal et

al., 2016). The task-specific layers, usually the lower layers, learn features that are more related to the current task. MTL lowers the chances of overfitting as the model has to learn the common representation among all tasks. MTL has been widely adopted in many different domains (Luong et al., 2016).

Crichton et al. (2017) proposed a multi-task model (MTM) based on CNN to perform BioNER. However, they only focused on the word level features ignoring the character level ones. Although word level features give much information about the entities, character level features help to extract common sub-word structures among the same entities. Moreover, depending solely on the word level features can lead to out-of-vocabulary problems when a specific word is not found in the pre-trained word embedding. Wang et al. (2019) also performed BioNER using different multi-task models. They found that the MTM with the word level features and extraction of the character level features using BiLSTM enhances performance of the model. They concluded that the character level feature should be considered for the BioNER task. A similar model is proposed by Mehmood et al. (2019) where, apart from single shared BiLSTM, they introduce the task-specific BiLSTM as well to learn the features that are more specific to the task. Introduction of task-specific BiLSTM and use of CNN instead of BiLSTM at character level showed performance improvement.

3 Our Proposal

Neural networks work on a concept of hierarchical feature learning (Xiao et al., 2018). Hierarchical feature learning is done as sequences propagates through the network (LeCun et al., 2015). Deep learning can learn the complex hierarchical structure of the sequence with multiple layers. Moreover, it is always desirable to stack LSTMs when a large amounts of training data is available (Li et al., 2018). Such intuition can be noticed in the model proposed by Mehmood et al. (2019) where increasing the layer of BiLSTM leads to performance enhancement. However, moving towards deep LSTMs network can causes gradient vanishing problem as well (Li et al., 2018).

To tackle this issue we are proposing a model which induces the input information at different layers. Our proposed multi-task model with character and word input representations (MTM-CW)

propagates input embedding information along different shared layers as shown in Figure 1. This not only helps lower layers to learn the complex structure from encoded representation of the previous layer but also considers inputs embeddings as well to overcome the gradient vanishing problem in stacked LSTMs.

Furthermore, using stacked BiLSTMs will help hidden states of BiLSTM to learn hidden structure of the data presented at different level. This will help BiLSTM to learn features at a more abstract level. Apart from the shared stacked BiLSTMs, our model also uses task-specific BiLSTM as well to extract task-specific features. Furthermore, we use CNN to extract features at character level. Many of the previous approaches have used CNN at character level (dos Santos et al., 2015; Collobert et al., 2011) due to its finer ability of features extraction. CNN learns global level features from local level features. This enables CNN to extract more hidden features. More specifically, lower layers in our proposed MTM-CW model are task-specific. So for the specific task, both shared layers and layers belonging to that specific task are activated.

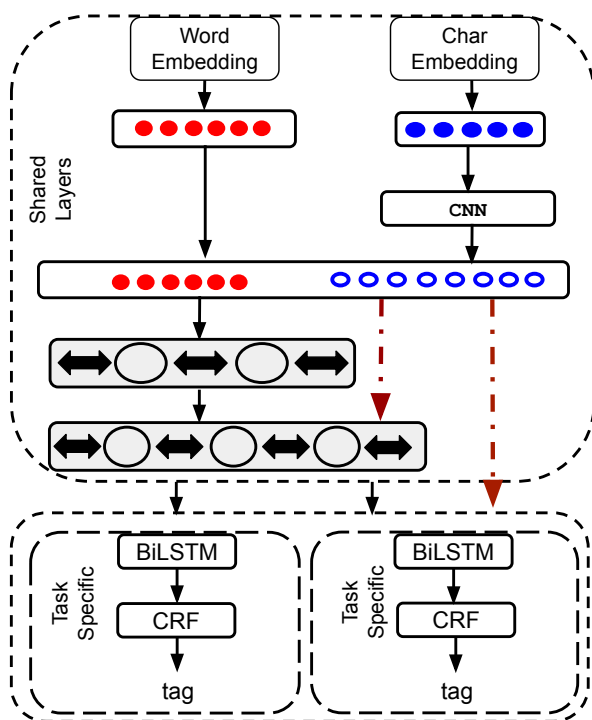


Figure 1: Proposed MTM-CW Model where dashed arrows show skip connections

Finally, we use CRFs for output labeling. CRFs have the ability to tag the current token by considering neighboring tags at sentence level (Huang et al., 2015). Yang et al. (2018) performed experiments comparing CRF and Softmax and found out that CRF produces better results compared to Softmax.

An alternative training approach was adopted for the training phase. Let suppose we have D_1, D_2, \dots, D_t training sets, related to the T_1, T_2, \dots, T_t tasks respectively. During training, a training set D_i is selected randomly and both shared layers and layers specific to the corresponding task T_i are activated. Every task has its own optimizer so during training only the optimizer specific to the task T_i is activated and the loss function related to that optimizer is optimized. It means that the parameters of the shared layers and of the task-specific layers are changed during the training of the specific task. Optimizing parameters of the shared layers for all the tasks helps the model to find the common features among different tasks.

4 Experiments

We performed experiments on the 15 datasets which were also used by Crichton et al. (2017), Wang et al. (2019), and Mehmood et al. (2019). The bio-entities in these datasets are Chemical, Species, Cell, Gene/Protein, Cell Component, and Disease¹. Descriptions of the datasets can be found in Crichton et al. (2017). Moreover, to represent words, we use domain-specific pre-trained word embeddings since generic word embeddings can cause a high rate of out-of-vocabulary words. In particular, we use WikiPubMed-PMC word embedding which is trained on a large set of the PubMedCentral(PMC) articles and PubMed abstracts as well as on English Wikipedia articles (Giorgi and Bader, 2018). On the other hand, character embedding is initialized randomly while orthographic (case) embedding is represented by the identity matrix where each diagonal 1 represents the presence of a word's orthographic feature. Moreover, we analyse the effect of different input representations (word level, character level, and case level) of a word on the performance of our proposed architecture. Furthermore, this paper reports the average F1-score where each experiment is run for 10 times. We use the Nadam

¹The datasets can be found at the following link <https://github.com/cambridgeltl/MTL-Bioinformatics-2016>

optimizer in our model and use CNN with a filter size of 30 while each LSTM in the model consists of 275 units and the experiment is run for 50 epochs and early stop is set to 10 epochs.

5 Results and Discussion

In Table 1 we compare the results produced by our model with state-of-the-art models (Wang et al., 2019; Mehmood et al., 2019). We can see a substantial improvement in the F1-score by MTM-CW compared to these models. However, to observe whether connecting embedding layers to the middle layers has truly contributed to the performance of the model, we made a variation in the model and dropped the skip connections coming from embedding layers (refer to Figure 1). Dropping these skip connections makes our model similar to the model by Mehmood et al. (2019) where we have introduced another layer of shared BiLSTM. The effect of such variation is reported in Table 2 where it can be noted that few datasets show moderate performance increase while for most of them performance degrades. This supports our intuition that passing embedding layer information to the lower layers has positive impact on the model. Moreover, it is interesting that, even after dropping those skip connections, our model is still able to perform better compared to state-of-the-art models. This suggests that, with increasing size of training examples, more layers of LSTM should be considered (Li et al., 2018). For this reason, the proposed model by Mehmood et al. (2019) performed better compared to model proposed by Wang et al. (2019) which used single layer of LSTM.

We then extended our experiments by introducing orthographic-level representation of a word in our model. Dugas and Nichols (2016) Segura-Bedmar et al. (2015) Huang et al. (2015) have shown that orthographic-level information can improve model’s performance. In addition, statistical models (e.g. CRF at the output layer) are also highly dependent on hand-crafted features (Limsopatham and Collier, 2016). In this work, the orthographic-level feature includes information on the structure of the word, i.e. either the word is starting with a capital letter followed by small letters or all the letters in the word are capital or contain digits, etc. Table 2 reports the comparison between MTM-CW and its variant with orthographic-level features (we name it

Datasets	Wang et al.	Mehmood et al.	MTM-CW
AnatEM	86.04	86.99	87.50
BC2GM	78.86	80.82	81.57
BC4CHEMD	88.83	87.39	89.24
BC5CDR	88.14	87.85	88.54
BioNLP09	88.08	88.74	88.52
BioNLP11EPI	83.18	84.75	85.36
BioNLP11ID	83.26	87.65	87.19
BioNLP13CG	82.48	84.25	84.94
BioNLP13GE	79.87	79.82	80.91
BioNLP13PC	88.46	88.84	89.16
CRAFT	82.89	83.15	85.23
Ex-PTM	80.19	80.95	81.72
JNLPBA	72.21	74.05	72.10
linnaeus	88.88	87.79	88.12
NCBI-disease	85.54	85.66	85.07

Table 1: Multi-task Models Comparison where CW represents character and word respectively

case, MTM-CW-Case). We observe that, for some datasets, orthographic-level features moderately improved the results. Thus, we can conclude that orthographic-level features might help the model to implicitly learn hidden features at an orthographic level which could be helpful for some entities. However, for simplicity we are limiting our work to explicitly representing the word-level features; thus we stick to the character-level representation and the word itself. We also replaced CRF with Softmax at the output layer to see the impact of both methods on predicting the output label of the entities. Table 2 also depicts the comparison of our proposed model with softmax (MTM-CW-Softmax) and CRF (proposed MTM-CW) at the output layer and model with CRF produce better results compared to the model with Softmax.

To statistically evaluate the results obtained by different variants of our model we perform the Friedman test (Zimmerman and Zumbo, 1993). We also analyse the pairwise comparison of different models to see which model is statistically better than the other. The graphical representation of the pairwise comparison is shown in Figure 2 as it can be seen in variant of the model proposed with softmax (MTM-CW-Softmax represented as just Softmax) which is statistically worse compared to the others and to other variants of the model. Figure 3 shows the post-hoc Conover Friedman test where it can be seen that the difference between results produced by all the models is significant with different p values.

Datasets	MTM-CW	MTM-CW (w/out skip connections)	MTM-CW Case	MTM-CW Softmax
AnatEM	87.50	86.94	87.37	86.36
BC2GM	81.57	81.29	81.66	80.04
BC4CHEMD	89.24	87.44	89.13	86.88
BC5CDR	88.54	88.11	88.64	87.39
BioNLP09	88.52	89.31	88.61	88.18
BioNLP11EPI	85.36	85.01	85.04	84.16
BioNLP11ID	87.19	88.16	87.76	87.28
BioNLP13CG	84.94	84.61	84.86	84.00
BioNLP13GE	80.91	82.28	80.16	80.49
BioNLP13PC	89.16	89.04	89.26	88.37
CRAFT	85.23	83.44	85.04	82.86
Ex-PTM	81.72	82.40	81.50	80.64
JNLPBA	72.10	72.02	72.21	70.31
linnaeus	88.12	88.69	88.74	88.33
NCBI-disease	85.07	85.12	85.56	84.36

Table 2: Comparison between the Results of Different Variants of the Model Proposed

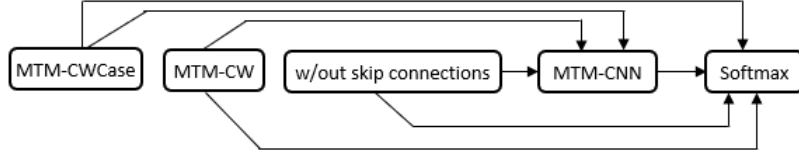


Figure 2: Pairwise Models Comparison w.r.t to Friedman Test

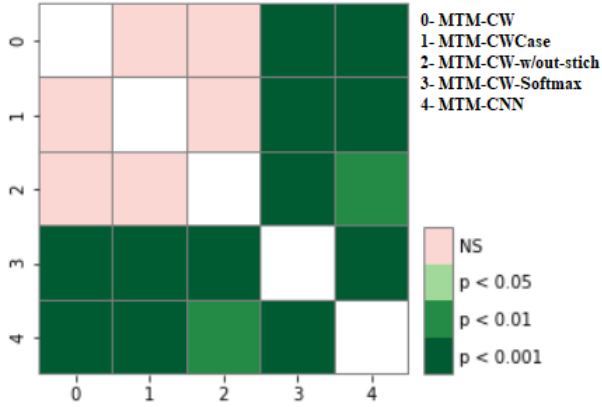


Figure 3: Post-hoc Conover Friedman Test (NS represents not significant)

6 Conclusion and Future Work

In this paper we showed that the BioNER performance can be drastically improved by using a multi-task approach. We showed that using stacked LSTMs in such models are effective to learn hidden structure of the data. Moreover, to overcome the vanishing gradient problem in using

stacked LSTMs is addressed by passing embedding information layers to layers. We showed that our model outperforms in F1-score compared to the state-of-the-art models.

For future work, we will extend the multi-task approach for relation extraction task. In such approach, BioNER can be used as an auxiliary task while keeping relation extraction task as the main task in the multi-task approach.

References

- Trapit Bansal, David Belanger, and Andrew McCallum. 2016. Ask the gru: Multi-task learning for deep text recommendations. In *Proceedings of the 10th ACM Conference on Recommender Systems*, pages 107–114. ACM.
- Hai Leong Chieu and Hwee Tou Ng. 2002. Named entity recognition: a maximum entropy approach using global information. In *Proceedings of the 19th international conference on Computational linguistics-Volume 1*, pages 1–7. Association for Computational Linguistics.
- Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa.

2011. Natural language processing (almost) from scratch. *Journal of machine learning research*, 12(Aug):2493–2537.
- Gamal Crichton, Sampo Pyysalo, Billy Chiu, and Anna Korhonen. 2017. A neural network multi-task learning approach to biomedical named entity recognition. *BMC bioinformatics*, 18(1):368.
- Cícero Nogueira dos Santos and Victor Guimarães. 2015. Boosting named entity recognition with neural character embeddings. In *Proceedings of the Fifth Named Entity Workshop, NEWS@ACL 2015, Beijing, China, July 31, 2015*, pages 25–33.
- Cícero dos Santos, Victor Guimaraes, RJ Niterói, and Rio de Janeiro. 2015. Boosting named entity recognition with neural character embeddings. In *Proceedings of NEWS 2015 The Fifth Named Entities Workshop*, page 25.
- Fabrice Dugas and Eric Nichols. 2016. DeepNNER: Applying BLSTM-CNNs and extended lexicons to named entity recognition in tweets. In *Proceedings of the 2nd Workshop on Noisy User-generated Text (WNUT)*, pages 178–187.
- John M Giorgi and Gary D Bader. 2018. Transfer learning for biomedical named entity recognition with neural networks. *Bioinformatics*, 34(23):4087–4094.
- Mourad Gridach. 2017. Character-level neural network for biomedical named entity recognition. *Journal of biomedical informatics*, 70:85–91.
- Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional LSTM-CRF models for sequence tagging. *CoRR*, abs/1508.01991.
- Yann LeCun, Yoshua Bengio, and Geoffrey E. Hinton. 2015. Deep learning. *Nature*, 521(7553):436–444.
- Jinyu Li, Changliang Liu, and Yifan Gong. 2018. Layer trajectory LSTM. In *Interspeech 2018, 19th Annual Conference of the International Speech Communication Association, Hyderabad, India, 2-6 September 2018*, pages 1768–1772.
- Nut Limsopatham and Nigel Collier. 2016. Learning orthographic features in bi-directional LSTM for biomedical named entity recognition. In *Proceedings of the Fifth Workshop on Building and Evaluating Resources for Biomedical Text Mining, BioTxtM@COLING 2016, Osaka, Japan, December 12, 2016*, pages 10–19.
- Minh-Thang Luong, Quoc V. Le, Ilya Sutskever, Oriol Vinyals, and Lukasz Kaiser. 2016. Multi-task sequence to sequence learning. In *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*.
- Tahir Mehmood, Alfonso Gerevini, Alberto Lavelli, and Ivan Serina. 2019. Leveraging multi-task learning for biomedical named entity recognition. In *International Conference of the Italian Association for Artificial Intelligence*. Springer.
- Isabel Segura-Bedmar, Víctor Suárez-Paniagua, and Paloma Martínez. 2015. Exploring word embedding for drug name recognition. In *Proceedings of the Sixth International Workshop on Health Text Mining and Information Analysis, Louhi@EMNLP 2015, Lisbon, Portugal, September 17, 2015*, pages 64–72.
- Xuan Wang, Yu Zhang, Xiang Ren, Yuhao Zhang, Marinka Zitnik, Jingbo Shang, Curtis Langlotz, and Jiawei Han. 2019. Cross-type biomedical named entity recognition with deep multi-task learning. *Bioinformatics*, 35(10):1745–1752.
- Cao Xiao, Edward Choi, and Jimeng Sun. 2018. Opportunities and challenges in developing deep learning models using electronic health records data: a systematic review. *JAMIA*, 25(10):1419–1428.
- Jie Yang, Shuailong Liang, and Yue Zhang. 2018. Design challenges and misconceptions in neural sequence labeling. In *Proceedings of the 27th International Conference on Computational Linguistics, COLING 2018, Santa Fe, New Mexico, USA, August 20-26, 2018*, pages 3879–3889.
- Donald W Zimmerman and Bruno D Zumbo. 1993. Relative power of the Wilcoxon test, the Friedman test, and repeated-measures ANOVA on ranks. *The Journal of Experimental Education*, 62(1):75–86.

Mining Italian Short Argumentative Texts

Ivan Namor

Università degli studi di Padova
ivan.namor@studenti.unipd.it

Samuele Garda

Potsdam Universität
garda@uni-potsdam.de

Pietro Totis

KU Leuven
pietro.totis@cs.kuleuven.be

Manfred Stede

Applied Computational Linguistics
University of Potsdam, Germany
stede@uni-potsdam.de

Abstract

We present the first model for argumentation mining for Italian short argumentative texts. We adapted to Italian the software developed by (Peldszus and Stede, 2015) and built a suitable corpus of Italian "microtexts" by semi-automatically translating the original English corpus. Our results are comparable to those of (Peldszus and Stede, 2015), which proves that their model is applicable successfully to languages other than English and German.¹

1. Introduction

In recent years, *argumentation mining* (Lippi and Torroni, 2016) has become an area of big interest in the field of natural language processing. Argumentation mining seeks to automatically recognize the structure of the argumentation in a text by identifying, classifying and connecting the central claim of a text, supporting premises, possible objections and counter-objection. Argumentation mining has many possible applications in very different fields. Recognizing automatically the argumentative structure of a text can be useful as an extension of opinion mining, in retrieval of court decisions from databases (Palau and Moens, 2011), in automatic document summarization (Teufel and Moens, 2002), in analysis of scientific papers as in biomedical text mining (Teufel, 2010; Liakata et al., 2012) in essay scoring, and more.

This task can be decomposed into several subtasks: segmentation of the text in elementary discourse units (EDUs), identification of argumentative discourse units (ADUs), classification of argumentative discourse units, identification of the relations between argumentative discourse units and classification of these relations. The argumentation structure of a text can be presented as a tree structure, with a node for each argumentative discourse unit and different edges between nodes representing the different types of relations. There are many simple models that recognize automatically the argumentation structure of a micro-text.

Our starting point is the model by (Peldszus and Stede, 2015), who developed a software to automatically mine the argumentation structure of short texts for English and German. In this paper we perform argumentation mining on a corpus of short Italian argumentative texts. To transfer the approach to Italian, we assembled a suitable corpus by semi-automatically translating the original German corpus and we adapted the features used by the software, by assembling a list of Italian connectives necessary to fulfill the task.

Our results are slightly lower than the ones for German and English, but they demonstrate that the model can be considered valid also for Italian. Besides, a major contribution of this paper is the free availability of the annotated Italian corpus.²

¹ Copyright © 2019 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

² <https://github.com/PietroTotis/evidencegraph>

2. Related works

(Peldszus and Stede, 2016) collected the *arg-microtext* corpus, a freely available parallel corpus of 112 texts with 576 argumentative ADUs (argumentative discourse units). It differs from other web-text corpora collected for argumentation mining purposes, such as the Internet Argument Corpus (Abbott et al., 2016) and the ABCD corpus (Rosenthal and McKeown, 2015), because the texts have been collected in a controlled text generation experiment.

(Peldszus and Stede, 2013) proposed an annotation scheme, which has been based on Freeman’s theory of argumentation structures (Freeman, 2011) and has been used to annotate the *arg-microtext corpus*. This annotation scheme has been proven to yield reliable structure in annotation and classification experiments (Peldszus and Stede, 2015; Potash et al., 2017).

One of a few similar approaches is that of (Stab and Gurevych, 2017), who introduced a corpus of persuasive essays annotated with argumentation structures related to the *arg-microtexts* and presented a similar approach for parsing argumentation structures.

An example of argumentation mining for Italian is presented in (Basile et al., 2016), where the researchers tested their method on a corpus of user comments to online newspaper articles.

3. Original Corpus

The interest in argumentation-oriented corpora of monologue text is rising, but most of the present data are not suitable for these operations. For this reason it is necessary to have well-formed and controlled corpora of short argumentative texts.

3.1 Data collection

In order to provide a corpus of Italian short argumentative texts we translated to Italian the *arg-microtexts* corpus, a freely available³ parallel

corpus of 113 short texts and a total of 576 ADUs (Peldszus and Stede 2015). The corpus is made by 90 short texts collected in a controlled text generation experiment and by 23 written directly by Andreas Peldszus, mainly in order to teach and test the probands of the experiment.

The texts are short but at the same time “complete” and the underlying argumentation structure is relatively clear. The probands were asked to first gather a list with the pros and cons of the trigger question, then take stance for one side and argue for it in a short argumentative text, which had to be at least five segments long with each segment argumentatively relevant, had to contain at least one objection and finally had to be understandable without having its trigger question as a headline. All of the microtexts were originally written in German and have been successively professionally translated in English.

3.2 Annotation scheme

The annotation scheme we used for our corpus is the same used for the original corpus, developed by Peldszus and Stede on the basis of different ideas from literature about argumentation structures (Peldszus and Stede, 2013). Two important steps in the development of a theory of argumentation are Toulmin’s influential analysis of argument (Toulmin, 1958) and Grewendorf’s dialog-oriented diagram method (Grewendorf, 1980).

The annotation scheme used for the *arg-microtexts* corpus is based mainly on Freeman’s theories, which integrate Toulmin’s ideas into the argument diagramming techniques of the informal logic tradition (Freeman, 1991, 2011). The central claim of Freeman’s theory is that the different ways in which premises and conclusions combine to form larger complexes, can be modeled as a hypothetical dialectical exchange between a proponent and an opponent. An argument is a non-empty set of premises supporting some conclusion. The argumentation structure of a text is defined as a graph with the text segments as nodes. Each node is associated with a specific argumentative role: the “proponent”, who presents and supports a central claim, and the “opponent”, who questions the proponent’s claims. Argumentative

3 <https://github.com/peldszus/arg-microtexts>

relations are represented by the edges between the nodes and have a specific argumentative function, which can be “support” or “attack”. Support relations can be of different types: basic, linked, multiple, serial and the example relation. Attack relations can target both premises or conclusions and can be of two different types: they are a “rebut” if they target another node or “undercut” if they target an edge between two nodes.

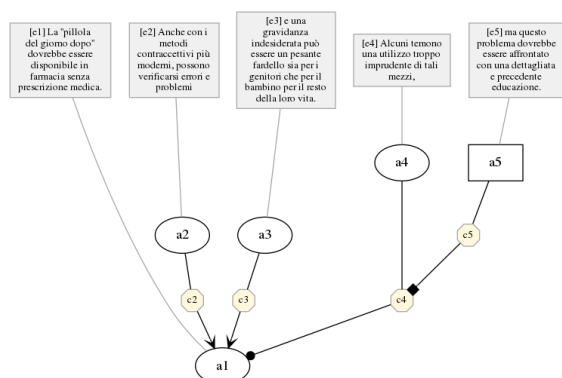


Figure 1: An example text (*micro_b037*) and its reduced argumentation structure: texts segments, proponent and opponent nodes (rounds and boxes), supporting, attacking and undercutting relations (arrow-head, circle-head and square-head).

4. Translation

The choice of translating into Italian the *arg-microtexts* corpus, likewise it was previously done for English, is motivated by the controlled setting of the experiment. The translation process had two phases. In the first phase we automatically translated the entire corpus using DeepL Translator⁴, a free and multilingual translation service. In the second phase, all the translations have been manually checked and, if needed, post-edited.

4.1 Post-editing

Some corrections were necessary in almost every microtext: from a syntactic point of view the translator respected most of the dependencies, losing however accuracy with increasingly complex syntactic structures. As

foreseeable, a lot of words were translated with the most common Italian translation, but not the most appropriate. All the microtexts have been thereby post edited in order to look as they were generated directly in Italian. Connectives have a fundamental role in the identification of function, role and attachments of a sentence. We therefore dedicated special attention to this aspect; in the automatic translation, many different original forms converged to the most common connective in the target language. For example, almost all the connectives expressing similarity were translated with “e” (“and”) and most of the connectives expressing contrast were translated with “ma” (“but”). In order to have a more realistic corpus we tried to use a more various set of connectives, comparable to the set used in the original corpus.

4.2 Projection annotations

The annotated graph structures are stored in XML format. The main advantage of translating the *arg-microtexts* corpus was that it was not necessary to make the annotations from scratch. As expected, there was a one by one correspondence between original sentences in German and the translations in Italian. In order to have Italian annotated graph structures it was only necessary to automatically substitute every German sentence in the XML file with the corresponding Italian sentence. In case a sentence contained more ADUs, it has been divided manually.

5. Software

The code for computing the tree predictions have been taken over from the work of Peldszus and Stede (Peldszus and Stede, 2015).

5.1 Original model

In order to recognize the argumentation structure, the model considers not only the probability of attachment of each segment pair, but also the probabilities of role, function and of being the central claim. In order to do so it is necessary to predict probabilities for each argumentative unit on different levels:

⁴ <https://www.deepl.com/translator>

attachment, central claim, role (proponent or opponent) and function (supporting or attacking).

The first step is to build a fully connected multigraph that connects every segment pair with as many edges as the function types. In order to get central claim, role, function and attachment probabilities, the model uses different classifiers and then jointly combines these probabilities in a single edge score, defined as the weighted sum of the level specific edge scores, on which it is possible to apply a MST (minimum spanning tree) algorithm (Chu and Liu, 1965; Edmonds, 1967).

The result represents the best global attachment structure for the text. This model outperformed other baseline and simpler models when tested on the German and English parallel corpus (Peldszus and Stede, 2015).

5.2 Adaptation to Italian

In order to run the original experiments on the Italian corpus, we adapted the sections of the code related to the corpus and the NLP tools. The latter represents the major divergence from the original setting, since it entailed upgrading the *spaCy* package, along with its language models. This also involved upgrading other packages and porting the whole project to *Python 3.x*, but these were minor modifications that should not have a meaningful impact on the performances.

A language-specific set of connectives is essential for the classification of the relations between ADUs. For this purpose, we used *LiCo*⁵, a lexicon of Italian connectives (Feltracco et al. 2016). The connectives are stored in XML format, each entry contains:

- Part type (phrasal or single).
- Syntactic type (preposition, adverb, coordinating conjunction, subordinating conjunction).
- Relation type (as cause, concession, contrast, purpose).
- An example of use in a sentence.

⁵ <http://connective-lex.info/>

6. Results

The metrics to evaluate our adaptation are Macro F1 and Micro F1 for each sub-task: central claim, role, function and attachment detection. The results are reported in Table 1.

Compared to the results obtained in the experiment with the English and the German corpus (Peldszus and Stede, 2015), the results for Italian are slightly lower. The results are almost the same for central claim and attachment detection and lower in function and role classification. The most significant drop of the F1 scoring regards the task of function classification. Nonetheless, the overall performances are sufficient to confirm the validity of the model for Italian. The smaller size of the Italian model provided by *spaCy* might explain the gap in performance with the other two languages.

	cc	ro	fu	at
Macro F1	0.813	0.724	0.413	0.690
Micro F1	0.883	0.811	0.593	0.792

Table 1: Results for Italian

	cc	ro	fu	at
Macro F1	0.825	0.765	0.431	0.706
Micro F1	0.888	0.841	0.618	0.796

Table 2: Results for English

	cc	ro	fu	at
Macro F1	0.817	0.750	0.671	0.663

Table 3: Results for English (Peldszus and Stede, 2015)

6.1 Error analysis

We investigated the reason for the lower performances in the task of function classification: Figure 2 and 3 show an example of misclassification. The prediction for the microtext mistakenly detects an attacking and an undercutting relation in place of two supporting relations. Wrong function classification of some argumentative unit can be found in most of the outputs of the corpus.

Another common error is the wrong attachment: Figure 3 and 4 present an interesting error for this task. In place of an “attach to first” structure, which is typical of the English style of

essay writing and can be used as baseline, our model has attached all the argumentative units to the preceding segment, which is also a typical baseline in discourse parsing (Muller et al., 2012).

We investigated the role of connectives in the attachment prediction and ran the same experiment on a less specific list of connectives, i.e. with more general relation types. With this simplified version of the connectives, the classifier achieved lower results in all the tasks. This suggests that specificity is not the reason behind these errors and at the same time proves the central role of the connectives in the recognition of an argumentation structure.

7. Conclusion

We presented, to our knowledge, the first model that transfers on an Italian microtexts corpus the approach developed by (Peldszus and Stede, 2015). We ran the experiment on an Italian corpus obtained by translating the original German one and by designing a suitable list of connectives. We adapted the code by changing the sections related to the corpus and the NLP tools. Our results are comparable to those of Peldszus and Stede, which proves that their model is applicable successfully to languages other than English and German.

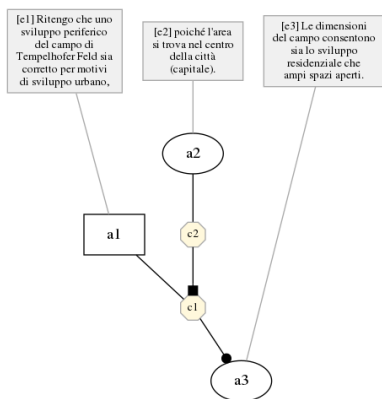


Figure 2: *micro_b033* wrong output

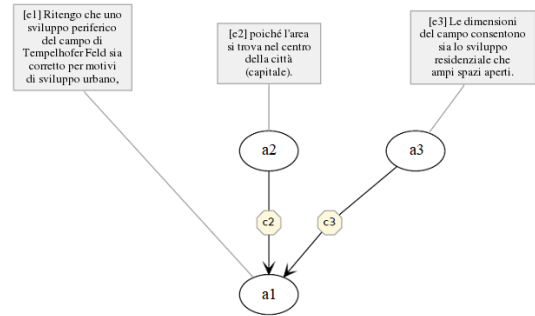


Figure 3: *micro_b033* expected output

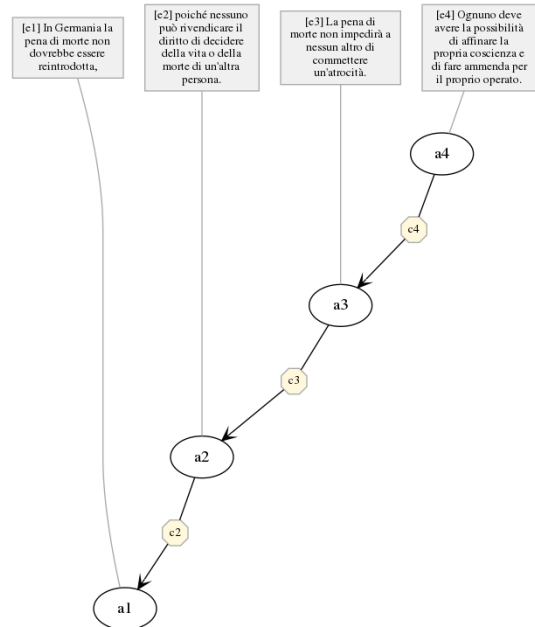


Figure 4: *micro_b031* wrong output

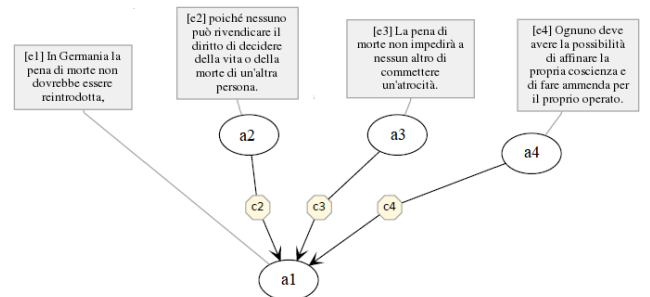


Figure 5: *micro_b031* expected output

References

- [Abbott et al.2016] Rob Abbott, Brian Ecker, Pranav Anand, and Marilyn Walker. 2016. Internet Argument Corpus 2.0: An SQL schema for dialogic social media and the corpora to go with it. In *Proc. Language Resources and Evaluation*, pages 4445–4452.
- [Basile et al.2016] Pierpaolo Basile, Valerio Basile, Elena Cabrio and Serena Villata. 2016. Argument Mining on Italian News Blogs.
- [Chu and Liu1965] Yoeng-Jin Chu and Tseng-Hong Liu. 1965. On the shortest arborescence of a directed graph. In *Science Sinica*, 14:1396–1400.
- [Edmonds1967] Jack Edmonds. 1967. Optimum Branchings. In *Journal of Research of the National Bureau of Standards*, 71B:233–240.
- [Feltracco, Jezek, Magnini and Stede2016] Anna Feltracco, Elisabetta Jezek, Bernardo Magnini and Manfred Stede. 2016. Lico: a lexicon of italian connectives. In *Proceedings of the 3rd Italian Conference on Computational Linguistics (CLiC-it)*. Napoli, Italy.
- [Freeman1991] James B. Freeman. 1991. Dialectics and the Macrostructure of Argument. Foris, Berlin.
- [Freeman2011] James B. Freeman. 2011. Argument Structure: Representation and Theory. *Argumentation Library (18)*. Springer.
- [Grewendorf1980] Günther Grewendorf. 1980. Argumentation in der Sprachwissenschaft. In *Zeitschrift für Literaturwissenschaft und Linguistik*, 10(38/39):129–150.
- [Liakata et al.2012] Maria Liakata, Shyamasree Saha, Simon Dobnik, Colin R. Batchelor and Dietrich Rebholz-Schuhmann. 2012. Automatic recognition of conceptualization zones in scientific articles and two life science applications. *Bioinformatics*, 28(7):991–1000.
- [Lippi and Torroni2016] Marco Lippi and Paolo Torroni. 2016. Argumentation Mining: State of the Art and Emerging Trends. In *ACM Transactions on Internet Technology*. 16. 1-25. 10.1145/2850417.
- [Muller et al.2012] Philippe Muller, Stergos Afantenos, Pascal Denis, and Nicholas Asher. 2012. Constrained decoding for text-level discourse parsing. In *Proceedings of COLING 2012*, pages 1883–1900, Mumbai, India, December. The COLING 2012 Organizing Committee.
- [Palau and Moens2011] Raquel Mochales Palau and Marie-Francine Moens. 2011. Argumentation mining. In *Artificial Intelligence and Law*, 19(1):1–22.
- [Peldszus and Stede2013] Andreas Peldszus and Manfred Stede. 2013. From argument diagrams to argumentation mining in texts: a survey. Universität Potsdam.
- [Peldszus and Stede2015] Andreas Peldszus and Manfred Stede. 2015. Joint prediction in MST-style discourse parsing for argumentation mining. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 938–948. Lisbon, Portugal.
- [Peldszus and Stede2016] Andreas Peldszus and Manfred Stede. 2016. An annotated corpus of argumentative microtexts. In *First European Conference on Argumentation: Argumentation and Reasoned Action*, Portugal, Lisbon.
- [Potash et al.2017] Peter Potash, Alexey Romanov and Anna Rumshisky. 2017. Here’s my point: Joint pointer architecture for argument mining. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1375–1384, Copenhagen, Denmark. Association for Computational Linguistics.
- [Rosenthal and McKeown2015] Sara Rosenthal and Kathy McKeown. 2015. I couldn’t agree more: The role of conversational structure in agreement and disagreement detection in online discussions. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 168–177, Prague, Czech Republic. Association for Computational Linguistics.
- [Stab and Gurevych2016] Christian Stab and Iryna Gurevych. 2016. Parsing Argumentation Structures in Persuasive Essays. In *Computational Linguistics*. 43. 10.1162/COLI_a_00295.
- [Teufel and Moens2002] Simone Teufel and Marc Moens. 2002. Summarizing scientific articles: Experiments with relevance and rhetorical status. *Computational Linguistics*, 28(4):409–445, December.
- [Teufel2010] Simone Teufel. 2010. The Structure of Scientific Articles: Applications to Citation Indexing and Summarization. In *CSLI Studies in Computational Linguistics*. CSLI Publications.
- [Toulmin1958] Stephen Toulmin. 1958. The Uses of Argument. Cambridge University Press, Cambridge.

Fixing Comma Splices in Italian with BERT

Daniele Puccinelli
DTI-ISIN
University of
Applied Sciences
of Southern Switzerland
Manno, Switzerland

Silvia Demartini
DFA-DILS
University of
Applied Sciences
of Southern Switzerland
Locarno, Switzerland

Renée E. D'Aoust
North Idaho College
Sandpoint, Idaho, USA
Casper College
Casper, Wyoming, USA

Abstract

We propose a fully unsupervised strategy to fix comma splices. Leveraging the pre-training of Bidirectional Encoder Representations from Transformers (BERT), our strategy is to mask out commas and let BERT guess what to replace them with. Our strategy achieves promising results on a challenging targeted corpus of awkwardly worded sentences from Italian-language college student essays.

1 Introduction

Comma splices can be defined as independent clauses joined by a comma without a coordinating conjunction (Hacker, 2009). Comma splices are frequent in both English and Italian and typically suggest a lack of basic understanding of sentence structure. As we will show, they come in various flavors, and there exist subtle differences between how they occur in English and Italian.

Comma splices are generally detected by commercial grammar and style checkers, but their automated correction has only been addressed by a few studies specific to English. Because the common denominator shared by such studies is the use of supervised machine learning techniques, the key research question that motivated the present study is whether we can use transfer learning to correct comma splices automatically in a completely unsupervised fashion and in languages other than English.

Thanks to contextualized word embeddings, and, in particular, thanks to BERT (Devlin et al., 2019), we show that it is possible to correct common cases of comma splices in Italian. We also discuss the limitations of our unsupervised approach.

2 Comma splices in Italian

Comma splices are widespread in contemporary written Italian language usage due to a tendency to over-extend the use of commas (Ferrari, 2017, 2018; Demartini and Ferrari, 2018). Several authors have studied this tendency in recent years. Some preserve the English language designation; this is the case in (Corno, 2019), where the expression *frasi fuse* (fused sentences) is also employed. Others employ alternate designations, such as *virgola passe-partout* (passe-partout comma) in (Tonani, 2010) and *virgola tuttofare* (factotum comma) in (Seriani and Benedetti, 2009).

Comma splices are one of the most frequent comma usage errors in Italian, especially among inexperienced L1 and L2 writers. Comma splices are also one of the principal and most common problems in the writing of university students, especially in science and engineering. Usually, these writers have failed to develop any linguistic awareness for text segmentation and organization, and they mistakenly assume that a comma can convey multiple functions, working both as a linker or as a strong stop.

There are some similarities and some differences compared to English usage, due to the fact that Italian punctuation is more communicative and less morphosyntactic. In gen-

eral, there are two main kinds of comma splices in Italian that are caused by the use of a comma where we would expect:

1. a logical connector to join two sentences that have a particular relationship;
2. a stronger punctuation mark to mark a logical-syntactic connection (colon) or break (semicolon or period).

According to (Ferrari, 2014), comma splices reflect a deep inability to handle both basic syntactic structures and text construction: if a text is characterized by coherence, cohesion, and topical organization, comma splices deconstruct these properties from the inside. For this reason, analyzing comma splices is extremely important in the context of improving language teaching.

Comma splices can be fixed in various ways, depending on the context and on the kinds of clauses involved. In the most straightforward cases, the comma can be replaced by a period or a semi-colon that explicitly separates the clauses on either side of the comma. In other cases, the comma can be replaced by an element that links the clauses, such as a colon, a conjunction, or a conjunctive adverb. Care must be exercised if sentences are more complex (i.e. with parenthetical elements) or syntactically inaccurate.

Due to the lack of an Italian-language corpus dedicated to comma splices, the authors have assembled a small corpus of 100 sentences containing a wide array of comma splices collected from college student writings (mostly in the field of engineering) at the Università del Piemonte Orientale (UPO) and the University of Applied Sciences of Southern Switzerland (SUPSI) in the mid-to-late 2010s. In the remainder of the paper, we will employ this UPO-SUPSI-SPLICE corpus (henceforth USS corpus) to evaluate the potential of our proposed method. Aside from containing comma splices, many USS sentences are poorly worded, syntactically inaccurate, and often unclear.

3 Related work

In the active research thread on automated grammar and style correction, the studies that are most closely related to ours are (Lee et al., 2014) on the automated detection of comma splices and, most recently, (Zheng et al., 2018) on the automated correction of run-on sentences. The techniques proposed in these studies, which are specific to English, rely on supervised learning techniques that require relatively extensive training sets. To the best of our knowledge, ours is the first investigation of the automated correction of Italian-language comma splices using unsupervised learning.

Our proposed unsupervised strategy leverages the rich research thread on word embeddings. Dense word embeddings went mainstream with Word2Vec (Mikolov et al., 2013) and gained traction in the mid-to-late 2010s in spite of their key limitation that a word type has the same word embedding regardless of context. Because words also have different aspects depending on semantics, syntactic behavior, and register/connotations, contextualized word embeddings have emerged as an elegant solution to capture word semantics across different contexts. TagLM (Peters et al., 2017) uses the hidden state of the bidirectional long-short term memory (LSTM) (Hochreiter and Schmidhuber, 1997) as a contextual word embedding. Instead of just using the output of the LSTM, ELMo (Peters et al., 2018) uses all the available hidden layers and combines them in a task-specific way with task-dependent trainable weights that can be learned for each task. ELMo embeddings have been shown to improve the state-of-the-art on a wide variety of challenging NLP tasks, but even more significant improvements have been shown with BERT (Devlin et al., 2019). Based on Transformer encoders (Vaswani et al., 2017), which are essentially a multi-headed attention stack where depth serves to compensate for the lack of recurrence, BERT pre-trains bidirectional representations by jointly conditioning on both the left and right context of individual tokens, and

allows for low-cost task-specific fine-tuning.

4 Fixing Comma Splices with BERT

While bidirectionality comes naturally to LSTM-based models, it is challenging to achieve it with Transformer-based models, because bidirectional conditioning with multiple layers inherently allows each word to see itself. BERT’s solution is to mask a relatively small portion of the tokens in the pre-training data and to train a bidirectional language model to guess them. If too few words are masked, training is too expensive, while if too many words are masked, BERT fails to learn about language; it was determined empirically that masking 15% of all tokens represents a reasonable compromise.

This specific aspect of BERT’s pre-training means that a pre-trained BERT model has the ability to predict missing tokens out of the box, i.e., with no task-specific fine-tuning and, therefore, no need for task-specific training data. For our purposes, this translates into a straightforward strategy to correct comma splices: **mask all commas** and **use BERT to guess what they should be**. In principle, if a masked comma is legitimate, we expect BERT to guess it is indeed a comma, while if it is not, as in a comma splice, we expect BERT to replace it with a more appropriate token.

BERT naturally lends itself to this task because it outputs an empirical probability distribution over a set of potential replacement tokens. Such tokens can be drawn out of the entire dictionary (including word pieces) or over a controlled subset. Jointly with the probabilistic nature of its output, BERT’s inherent bidirectionality may be directly harnessed by making predictions based on both the left and the right context of a masked comma and choosing the set of predictions associated with the highest probability.

If the array of potential replacement token is unrestricted, in complex sentences BERT may elect to replace commas with tokens belonging to inappropriate word classes, such as nouns or verbs. This can be avoided by re-

Strategy	Accuracy
Baseline	0.41
BERT - left context only	0.77
BERT - left & right context	0.81
BERT - PoS + left & right	0.87

Table 1: Sentence-level accuracy for the baseline strategy and the three different flavors of our BERT-based strategy described in this paper, measured on the USS corpus.

stricting the eligible potential replacement tokens to reasonable word classes.

5 Evaluation

As a proof of concept, we perform an empirical evaluation of our BERT-based strategy on the USS corpus, which contains sentences with at least one comma splice and a total number of commas ranging from one to seven. To the best of our knowledge, no directly comparable technique to fix Italian-language comma splices programmatically is freely available at the time of writing. To get a rough idea of the potential of our strategy, we use a simple baseline that replaces all commas with periods. While this baseline fails each time a sentence contains multiple commas, it fixes over 90% of the USS sentences that contain exactly one comma (41 out of 45). Aside from setting a performance floor, this baseline also offers a quick idea of the complexity of the sentences in the corpus.

As for our BERT-based strategy to fix comma splices, we make the following choices for the sake of simplicity:

- we employ `bert-multilingual-uncased` (and normalize all tokens to lower case);
- we draw potential replacement tokens out of the entire dictionary (aside from the PoS-based restrictions described below), but only consider potential replacement tokens with an estimated probability greater than 0.01 (arbitrary threshold);
- we make predictions based on both the

left and the right context of the masked tokens and choose the prediction associated with the highest probability, computed as the product of the probabilities of the most probable token replacement for each comma occurrence (we always mask out one comma at a time);

- we use PoS tags to exclude potential replacement tokens from word classes other than conjunctions and punctuation marks.

We employ TreeTagger to determine the PoS tags and use pre-trained BERT by way of `pytorch_pretrained_bert`.

We use sentence-level accuracy as our figure of merit and compute it as the fraction of error-free corrected sentences. A sentence is considered to be error-free by our strategy and/or by the baseline if the corrected version is acceptable according to two L1 human annotators. The corrected versions of sentences with multiple commas are only considered error-free if they contain no anomalies; while this is overly penalizing for our strategy in multi-comma sentences where a single mistake is made, it offers a conservative estimate of the performance of our BERT-based strategy.

As shown in Table 1, our BERT-based strategy is able to correct a total of 87 of the 100 sentences in the USS corpus to the satisfaction of the two L1 human annotators. An additional sentence is also corrected, but only if our strategy operates unidirectionally.

6 Discussion

Commas per sentence. The mean number of commas is 2.1 in the sentences where our strategy succeeds, while it is as high as 3.5 in the 12 sentences where our strategy fails. While multi-comma sentences are inherently more challenging, there doesn't seem to be a hard limit to the number of commas per sentence that our strategy can handle. Notably, our corpus contains a 7-comma excerpt:

Di solito, chi scrive senza conoscere le fasi della scrittura, scrive di

getto, seguendo i propri ragionamenti senza un ordine, così facendo, rischia di non scrivere un testo idoneo e fluente, dobbiamo essere attenti alle punteggiature, non scrivere le frasi molto lunghe e dividere in modo adeguato i capoversi.

which is fixed as

Di solito, chi scrive ... scrittura, scrive di getto, seguendo ... ordine. Così facendo, rischia ... fluente. Dobbiamo ... punteggiature, non ... capoversi.

Failures in single-comma sentences.

There are two single-comma sentences where our strategy fails: one contains a run-on sentence and also causes the baseline strategy to fail, while the other one has a mild form of comma splice:

Successivamente avviene la documentazione, si raccolgono e si scelgono le informazioni da fonti attendibili e si pianifica come esporle.

This sentence is the only instance in USS where our strategy fails and the baseline strategy succeeds. BERT chooses not to replace the comma, keeping the (borderline acceptable) comma splice unaltered. This happens due to the relative values of the probabilities assigned by BERT to a comma and a colon. Curiously, replacing *Successivamente* with the equivalent expression *Al passo successivo* is enough to nudge BERT in the right direction and assign a higher probability to a colon. This suggests that modifying individual tokens in a small corpus such as USS would be a meaningful dataset augmentation technique.

Left and right context. For 77 out of 100 sentences, a unidirectional pass based on the left context of the missing tokens is sufficient for our strategy to succeed. Only one of these 77 sentences can only be corrected unidirectionally; five other sentences can also be corrected by looking at the right context of the

missing tokens in a backward pass, which helps avoid blatantly erroneous replacements. Therefore, our strategy should be used with both left and right context. As an example, consider the sentence:

Essa consiste nel fatto che non c'è alcun legame naturalmente motivato, il significante cane non ha di per sé nulla che rimandi al suo nome, che faccia sì che quella cosa si possa chiamare così.

If BERT only relies on the left context of missing commas, the sentence is awkwardly split into three parts, with a striking error at the end:

Essa ...motivato. Il significante ...al suo nome. Che faccia sì che quella cosa si possa chiamare così.

With both left and right context, instead, our strategy offers an acceptable correction:

Essa ...motivato. Il significante ...al suo nome e che faccia ...così.

PoS filtering. A further six USS sentences can be fixed by combining left & right context and PoS filtering, which serves to avoid replacement tokens from implausible word classes and prevent awkward errors, such as the replacement of a comma with a preposition, a *che*, or a negation. Comma replacements with negations are particularly critical because they modify the meaning of the corrected sentence. PoS filtering is also helpful to prevent BERT from replacing commas with word pieces, which may occur with awkwardly worded sentences.

Unacceptable replacements. We have observed a limited number of unacceptable replacements of commas with colons, all of which occur in long-winded multi-comma sentences. Consider the six comma sentence:

Per la creazione della piattaforma web, il committente ha desiderato utilizzare una web application in Java, avendo la possibilità

di scegliere tra due framework, Spring e Struts, si è optato per l'utilizzo di Spring, siccome è uno strumento già utilizzato precedentemente, possiede un'ottima documentazione.

which becomes:

Per la creazione della piattaforma web. Il committente ...in Java, avendo la possibilità di scegliere tra due framework: Spring e Struts. Si è optato per l'utilizzo di Spring, siccome è uno strumento già utilizzato precedentemente e possiede ...

The first comma is erroneously replaced with a period, and the third one is questionably replaced with a colon. Replacing the fourth comma with a period is acceptable, as is preserving the second and fifth commas and turning the sixth comma into an *e*. Though our strategy makes four correct decisions out of six, this example is considered incorrect for our sentence-level quantitative analysis.

Interestingly, we have not observed any replacements with semicolons. We conjecture that BERT's strong preference for colons may be due to the relative frequency of colons versus semicolons in the pre-training text.

7 Conclusion

While the main limitation of the present study is the limited size of the USS corpus, we believe that the challenging nature of the writing excerpts in the USS corpus has enabled us to stress-test our strategy and to deliver a solid proof of concept that leverages the power of BERT-style contextualized word embeddings for automated style correction. Our future plans include using our BERT-based strategy to correct comma splices in English-language L1 and L2 student writing and to correct run-on sentences.

References

D. Corno. 2019. *Scrivere e comunicare. La scrittura italiana in teoria e in pratica*. Pearson.

- S. Demartini and P. Ferrari. 2018. La virgola splice nei testi di studenti universitari: un problema solo in apparenza superficiale. In *La Punteggiatura Italiana Contemporanea nella Varietà dei Testi Comunicativi*.
- J. Devlin, M. Chang, K. Lee, and K. Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL '19*.
- A. Ferrari. 2014. *Linguistica del testo. Principi, fenomeni, strutture..* Carocci.
- A. Ferrari. 2017. Usi "estesi" del punto e della virgola nella scrittura italiana contemporanea. *La lingua italiana. Storia, strutture, testi* XIII:137–153.
- A. Ferrari. 2018. *La punteggiatura italiana contemporanea. Unanalisi comunicativo-testuale*. Carocci.
- D. Hacker. 2009. *The Bedford Handbook for Writers*. Bedford Books.
- S. Hochreiter and J. Schmidhuber. 1997. Long short-term memory. *Neural Comput.* 9(8).
- J. Lee, C. Y. Yeung, and M. Chodorow. 2014. Automatic detection of comma splices. In *28th Pacific Asia Conference on Language, Information and Computation pages (PACLIC '14)*.
- T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean. 2013. Distributed representations of words and phrases and their compositionality. In *NIPS '13*.
- M. Peters, W. Ammar, C. Bhagavatula, and R. Power. 2017. Semi-supervised sequence tagging with bidirectional language models. In *ACL'17*. Vancouver, Canada.
- M. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer. 2018. Deep contextualized word representations. In *NAACL'18*. New Orleans, Louisiana.
- L. Serianni and G. Benedetti. 2009. *Scritti sui bianchi. L'italiano a scuola fra alunni e insegnanti*. Carocci.
- E. Tonani. 2010. *Il romanzo in bianco e nero. Ricerche sull'uso degli spazi bianchi e dell'interpunzione nella narrativa italiana dall'Ottocento a oggi*. Cesati.
- A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. Gomez, L. Kaiser, and I. Polosukhin. 2017. Attention is all you need. In *NIPS '17*.
- J. Zheng, C. Napoles, J. Tetreault, and K. Omelianchuk. 2018. How do you correct run-on sentences it's not as easy as it seems. In *The 4th Workshop on Noisy User-generated Text W-NUT*. Brussels, Belgium.

A Comparative Study of Models for Answer Sentence Selection

Alessio Gravina
Università di Pisa

gravina.alessio@gmail.com

Federico Rossetto
Università di Pisa

fedingo@gmail.com

Silvia Severini
Università di Pisa

sissisev@gmail.com

Giuseppe Attardi
Università di Pisa

attardi@di.unipi.it

Abstract

Answer Sentence Selection is one of the steps typically involved in Question Answering. Question Answering is considered a hard task for natural language processing systems, since full solutions would require both natural language understanding and inference abilities. In this paper, we explore how the state of the art in answer selection has improved recently, comparing two of the best proposed models for tackling the problem: the Cross-attentive Convolutional Network and the BERT model. The experiments are carried out on two datasets, WikiQA and SelQA, both created for and used in open-domain question answering challenges. We also report on cross domain experiments with the two datasets.

1 Introduction

Answer Sentence Selection is an important sub-task of Question Answering, that aims at selecting the sentence containing the correct answer to a given question among a set of candidate sentences. Table 1 shows an example of a question and a list of its candidate answers, taken from the *SelQA* dataset (Jurczyk et al., 2016). The last column contains a binary value, representing whether the sentence contains the answer or not.

Answer extraction involves natural language processing techniques for interpreting candidate sentences and establishing whether they relate to questions and contain an answer. More sophisticated methods of Answer Sentence Selection that

go beyond Information Retrieval approaches involve for example tree edit models (Heilman and Smith, 2010) and semantic distances based on word embeddings (Wang et al., 2016).

Recently, Deep Neural Networks have also been applied to this task (Rao et al., 2016), providing performance improvements with respect to previous techniques. The most common approaches exploit either *recurrent* or *convolutional* neural networks. These models are good at capturing contextual information from sentences, making them a nice fit for the problem of answer sentence selection.

Research on this problem has benefited in the last few years by the development of better datasets for training systems on this task. These datasets include *WikiQA* (Yang et al., 2015) and *SelQA* (Jurczyk et al., 2016). The latter is notable for its larger size, that reaches more than 60.000 sentence-question pairs. This allows for the creation of deeper and more complex models, with less risk of overfit.

The state of the art model on the *SelQA* dataset (Jurczyk et al., 2016), up to 2018, was *Cross-attentive Convolutional Network* (Gravina et al., 2018), with a score of 0.906 MRR (Craswell, 2009).

In this paper we present further experiments with the *Cross-attentive Convolutional Network* model as well as experiments that exploit the BERT language model by Devlin et al. (2018).

In the following sections we survey relevant literature on the topic, we describe the datasets used in our experiments and present the models tested in our experiments. Finally, we describe the experiments conducted with these models and report the results achieved.

2 Related work

We present a brief survey of the most recent approaches for answer selection in question answer-

All authors contributed equally to this manuscript.
Copyright © 2019 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

Table 1: Sample question/candidate answers.

How much cholesterol is there in an ounce of bacon?	
One rasher of cooked streaky bacon contains 5.4g of fat, and 4.4g of protein.	0
Four pieces of bacon can also contain up to 800mg of sodium.	0
The fat and protein content varies depending on the cut and cooking method.	0
Each ounce of bacon contains 30mg of cholesterol.	1

ing.

Tan et al. (2015) present four Deep Learning models for answer selection based on biLSTM (bidirectional LSTM) and CNN (Convolutional Neural Network), with different complexities and capabilities. The basic model, called QA-LSTM, implements two similar flows, one for the question and one for the answer. The biLSTM builds a representation of the question/answer pair that is passed by a max or average pooling layer. The two flows are then merged with a cosine similarity matching that expresses how close question and answer are.

A more complex solution, called QA-LSTM/CNN, uses a similar model, which replaces the pooling layer with a CNN. The output of biLSTM is sent to a convolution filter, in order to give a more complete representation of questions and answers. This filter is followed by 1-max pooling layer and a fully connected layer. Finally, the paper presents the most complex models, QA-LSTM with attention and QA-LSTM/CNN with attention, that extend the previous models with the addition of a simple attention mechanism between question and answer, which aims to better identify the best candidate answer to the question. The mechanism consists in multiplying the biLSTM hidden units of the answers with the output computed from the question pooling layer. These models are tested on the *InsuranceQA* (Feng et al., 2015) and *TREC-QA* (Yao et al., 2013) datasets, achieving quite good performances.

The HyperQA (Tay et al., 2017) model uses a pairwise ranking objective to represent the relationship between question and answer embeddings in a hyperbolic space instead of an euclidean space. This empowers the model with a self-organizing ability and enables automatic discovery of latent hierarchies while learning embeddings of questions and answers.

Wang et al. (2016) present a model that takes into account similarities and dissimilarities be-

tween sentences by decomposing and composing lexical semantics over sentences. In particular the model represents each word as a vector and calculates a semantic matching vector for each word based on all words in the other sentence. Then each word vector is decomposed into a similar and a dissimilar component, based on the semantic matching vector. Afterwards, a CNN model is used to capture features by composing these parts and a similarity score is estimated over the composed feature vectors to predict which sentence is the answer to the question.

3 Models

We describe here the models used in our experiments.

3.1 Simple Logistic Regression Classifier

Jurczyk et al. (2016) state that the SelQA dataset was created through a process that tried to reduce the number of co-occurrent words, so that simple word matching methods would be less effective. To evaluate whether this aim was indeed achieved, we built a simple linear regression classifier using as features the sentence and question length, the number of co-occurrent words and the *idf* coefficients of the word co-occurrences.

3.2 Cross-attentive Convolutional Network

The Cross-attentive Convolutional Network (CACN) is a model designed for the task of Answer Sentence Selection and in 2018 had achieved state of the art performance (Gravina et al., 2018). The model relies on a *Convolutional Neural Network* with a double mechanism of attention between questions and answers. The model is inspired by the light attentive mechanism proposed by Yin and Schütze (2017), which it improves by applying it in both directions to question and answer pairs.

The CACN model achieved top score in the "Fujitsu AI NLP Challenge 2018"¹, that used the

¹<https://openinnovationgateway.com/ai-nlp-challenge/>

SelQA dataset.

3.3 BERT language representation model

BERT (Bidirectional Encoder Representations from Transformers) (Devlin et al., 2018) is a language representation model. BERT usage involves two steps: *pre-training* and *fine-tuning*. During pre-training, the model is trained on a large collection of unlabeled text on a language modeling task. Fine-tuning BERT on a downstream task involves extending the model with additional layers tailored to the task, initializing the model with the pre-trained parameters, and then training the extended model with labeled data from the task. The extended model might consist just of a single output layer. Such models have been shown capable to achieve state-of-the-art accuracy for a wide range of tasks, such as question answering, machine translation, summarization and language inference.

Several pre-trained BERT models are publicly available, including the following ones that we used in our experiments:

- BERT-Base Uncased: with 12 layers, hidden size of 768 and a total number of 110M parameters;
- BERT-Large Uncased: with 24 layers, hidden size of 1024 and a total number of 340M parameters.

4 Datasets

We tested the models on two datasets: *SelQA* and *WikiQA*. The first one is the one used in the Fujitsu AI-NLP Challenge, while the second one is a commonly used dataset for open-domain Question Answering. A more detailed description follows.

4.1 SelQA

The *SelQA* dataset (Jurczyk et al., 2016) was specifically created to be challenging for question answering systems, in particular by explicitly reducing word co-occurrences between question and answers. Questions with associated long sentence answers were generated through crowd-sourcing from articles drawn from the ten most prevalent topics in the English Wikipedia.

The dataset consists of a total of 486 articles that were randomly sampled from the topics of: Arts, Country, Food, Historical Events, Movies, Music, Science, Sports, Travel, TV. The original data

was preprocessed into smaller chunks, resulting in 8,481 sections, 113,709 sentences and 2,810,228 tokens.

For each section, a question that can be answered in that same section by one or more sentences was generated by human annotators. The corresponding sentence or sentences that answer the question were selected. To add some noise, annotators were also asked to create another set of questions from the same selected sections excluding the original sentences previously selected as answers. Then all questions were paraphrased using different terms, in order to ensure the QA algorithm would be evaluated by their reading comprehension ability rather than from statistical measures like counting word co-occurrences. Lastly if ambiguous questions were found, they were rephrased again by a human annotator.

4.2 WikiQA

The *WikiQA* dataset (Yang et al., 2015) dataset consists of 3047 questions sampled from *Bing* query logs from the period of May 1st, 2010 to July 31st, 2011. Each question is associated to sentences taken from a *Wikipedia* page assumed to be the topic of the question based on the user clicks. In order to eliminate answer sentence biases caused by key-word matching, the sentences were taken from the summary of this selected page.

The *WikiQA* dataset contains also questions for which there are no correct sentences to enable researchers to work on *answer triggering*.

This dataset has the drawback to be smaller compared to *SelQA*. Because of this, a model is more likely to over-fit the training set. To avoid this problem we added some strong regularization to the models.

5 Experiments

5.0.1 GloVe, ELMo and FastText

We carried out some preliminary experiments on the *SelQA* dataset, in order to determine which embeddings would work best with the CACN.

We tested three types of embeddings: *GloVe* (size 300), *ELMo* (Che et al., 2018) (size 1024) and *FastText* (Joulin et al., 2016) (size 300). With *ELMo* the model achieved comparable results to *GloVe*, but the training time was almost twice.

Model	Dev MRR	Test MRR
ELMo	91.09%	90.00%
FastText	89.47%	88.43%
GloVe	91.37%	90.61%

Table 2: Results for CACN on SelQA with various embeddings.

5.1 SelQA results

The logistic regression classifier obtains a score of 83.36 %, which is 7 points lower than CACN, not bad considering the simplicity of the model. Nevertheless this confirms that a simple word matching method is not competitive with more sophisticated methods on SelQA.

CACN was the best performing model on the Fujitsu AI NLP Challenge 2018, with a MRR of 90.61 %.

After the introduction of BERT, we decided to compare CACN with several versions of BERT, both alone and in combination with CACN.

We tried a few variant approaches. First, we fine-tuned a fully connected layer on top of BERT, leaving his parameters frozen, on the SelQA training set. This model achieved 91.17, a marginal improvement over CACN.

We then explored adding different networks on top of the BERT architecture.

We added a full CACN, on top of either the BERT-Base and BERT-Large models, with no improvement and even a drop with BERT-Large. Also in this case we froze the parameters of the BERT model.

Since these experiments did not provide improvements, we didn't try to train the entire model.

The best results were achieved by fine-tuning the BERT model on the SelQA dataset with a simple feed-forward layer, that achieved an impressive improvement of about 5 points to a MRR score of 95.29 %. Fine-tuning required about 4 hours on a server with an Nvidia P100 GPU.

The results of all our experiments on SelQA are summarized in table 3.

5.2 WikiQA results

In the experiments with CACN on WikiQA, we removed from the training set questions with no correct answer, but left the test set unchanged, so that the results are comparable with those in the literature. This was done to preserve a similar structure to the SelQA dataset, which contains at least

Model	MRR
LR Classifier	83.36
CACN GloVe	90.61
BERT-Base + FCN	91.17
BERT-Base + CACN	91.11
BERT-Large + CACN	89.97
BERT-Base Fine-tuned	95.29

Table 3: Results on SelQA with various models.

one correct answer for each question. This significantly reduced the number of training examples but, despite this, the MRR score of the CACN model improved.

Also in this case we kept the word embeddings fixed during training the CACN. We also added a dropout and normalization to regularize the model, that helped the model to better learn from the training set.

We then fine-tuned BERT on the WikiQA training set, performing full updates to the model, achieving again a significant improvement to a top score of 87.53 % MRR.

From the current leaderboard on the WikiQA dataset ², we have extracted the top 5 entries and added the results with CACN and BERT-Base fine-tuned, as reported in Table 4.

Model	MRR	Year
BERT-Base Fine-tuned	87.53 %	2019
Comp-Clip + LM + LC	78.40 %	2019
RE2	76.18 %	2019
HyperQA (Tay et al., 2017)	72.70 %	2017
PWIM	72.34 %	2016
CACN (Gravina et al., 2018)	72.12 %	2018

Table 4: Experimental results on WikiQA.

5.3 Cross-domain experiments

In this section we report the results of our cross-domain experiments. The aim was to evaluate how well the CACN model performs in a context different from the one in which it was trained. In other words, we test the transfer learning ability of the model to a different domain.

The experiments consisted in training a model on one dataset and then testing it on the other one. We report in Table 5 the results of these experiments.

²<https://paperswithcode.com/sota/question-answering-on-wikiqa>

Trainset	Testset	MRR	Transfer score
SelQA	SelQA	90.61%	
SelQA	WikiQA	59.94%	82.95%
WikiQA	WikiQA	72.12%	
WikiQA	SelQA	69.45%	76.64%

Table 5: Cross domain experiments.

The drop in MRR score is small when training on WikiQA and testing on SelQA and larger in the other direction.

This is possibly due to the size of the datasets. In the second case in fact we are training on only 8000 pairs and testing on more than 80000 question/answer pairs.

However, the transfer score, computed as the ratio between the in-domain and out-domain MRR, is fairly good: about 83% in the SelQA to WikiQA case and over 76% in the other direction.

6 Conclusions

We compared the Cross-attentive Convolutional Network and several BERT based models on the task of Answer Sentence Selection on two datasets.

The experiments show that a BERT model, fine-tuned on an Answer Sentence Selection dataset, improves significantly the state of the art, with a gain of 5 to 9 points of MRR score on SelQA and WikiQA respectively. As a drawback, this approach takes a considerable amount of time to be trained even on GPUs.

The BERT-Base model without fine-tuning achieves almost the same accuracy as the CACN with GloVe embeddings, which uses a much smaller number of parameters in the model. The CACN also requires less data to train. On the other hand, BERT is quite effective at leveraging the knowledge collected from large amounts of unlabeled text, and at transferring it across tasks.

We also evaluated the abilities of CACN at transfer learning. BERT is a model that has been pre-trained on a large corpus, while CACN leverages the GloVe embeddings as a starting point for the training.

We also exploited the WikiQA and SelQA datasets in a cross-domain experiment using CACN. We found that the model maintains a good score across domains, with a transfer score of about 83% from SelQA to WikiQA.

We confirmed that the SelQA dataset is not eas-

ily solvable using simple word-occurrences methods like a logistic regression classifier on word count features.

BERT models confirmed their superiority to previous state of the art models for the task of Answer Sentence Selection. This was to be expected since they perform quite well also on the more complex task of Reading Comprehension, which requires not only to select a sentence but also to extract the answer from that sentence.

7 Acknowledgements

The experiments were carried on a Dell server with 4 Nvidia GPUs Tesla P100, partly funded by the University of Pisa under grant Grandi Attrezzature 2016.

References

- Wanxiang Che, Yijia Liu, Yuxuan Wang, Bo Zheng, and Ting Liu. 2018. Towards better UD parsing: Deep contextualized word embeddings, ensemble, and treebank concatenation. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 55–64, Brussels, Belgium, October. Association for Computational Linguistics.
- Nick Craswell. 2009. Mean Reciprocal Rank. In Ling Liu and M. Tamer Özsu, editors, *Encyclopedia of Database Systems*. Springer US, Boston, MA.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
- Minwei Feng, Bing Xiang, Michael R. Glass, Lidan Wang, and Bowen Zhou. 2015. Applying deep learning to answer selection: A study and an open task. *arXiv preprint arXiv:1508.01585*.
- Alessio Gravina, Federico Rossetto, Silvia Severini, and Giuseppe Attardi. 2018. Cross attention for selection-based question answering. In *NL4AI@ AI* IA*, pages 53–62.
- Michael Heilman and Noah A. Smith. 2010. Tree edit models for recognizing textual entailments, paraphrases, and answers to questions. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, HLT 10*, pages 1011–1019. Association for Computational Linguistics.
- Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Herve Jgou, and Tomas Mikolov. 2016. Fasttext.zip: Compressing text classification models. cite arxiv:1612.03651Comment: Submitted to ICLR 2017.

- Tomasz Jurczyk, Michael Zhai, and Jinho D. Choi. 2016. SelQA: A New Benchmark for Selection-based Question Answering. In *Proceedings of the 28th International Conference on Tools with Artificial Intelligence, of ICTAI'16*, pages 820–827.
- Jinfeng Rao, Hua He, and Jimmy Lin. 2016. Noise-contrastive estimation for answer selection with deep neural networks. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management (CIKM 16)*, pages 1913–1916. ACM.
- Ming Tan, Bing Xiang, and Bowen Zhou. 2015. Lstm-based deep learning models for non-factoid answer selection. *CoRR*, abs/1511.04108.
- Yi Tay, Anh Tuan Luu, and Siu Cheung Hui. 2017. Enabling efficient question answer retrieval via hyperbolic neural networks. *CoRR*, abs/1707.07847.
- Zhiguo Wang, Haitao Mi, and Abraham Ittycheriah. 2016. Sentence similarity learning by lexical decomposition and composition. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1340–1349. The COLING 2016 Organizing Committee.
- Yi Yang, Scott Wen tau Yih, and Chris Meek. 2015. WikiQA: A challenge dataset for open-domain question answering. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. ACL Association for Computational Linguistics, September.
- Xuchen Yao, Benjamin Van Durme, Chris Callison-Burch, and Peter Clark. 2013. Answer extraction as sequence tagging with tree edit distance. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 858–867.
- Wenpeng Yin and Hinrich Schütze. 2017. Attentive Convolution. *CoRR*.

An Italian Question Answering System for Structured Data based on Controlled Natural Languages

Lucia Siciliani and Pierpaolo Basile and Giovanni Semeraro
Department of Computer Science, University of Bari Aldo Moro, Italy
firstname.lastname@uniba.it

Matteo Mennitti
Sud Sistemi srl, Italy
mennittim@sudsistemi.it

Abstract

Question Answering over structured data represents one of the main challenges in the field of Natural Language Processing since it requires to render natural language, which is used by people every day, into a formal language, which can be processed by a machine. This task is particularly tricky due to the gap between the vocabularies adopted by users and the formalism that characterizes any query language. For this reason, although its birth as a discipline dates back to the late sixties, Question Answering over structured data is still accomplished to an unsatisfying degree. This result is even more critical if we take into account languages different from English, for which the amount of available resources is limited. In this paper we present MULIB, a Question Answering system capable of answering questions in Italian over both Knowledge Bases and databases.

1 Introduction and Motivation

Question Answering (QA) over structured data has the aim to interpret a natural language question issued by the user and retrieve an answer from a structured data source. Nowadays, the task of QA over structured data is usually performed over Knowledge Graphs (KGs), which encode an enormous amount of information and can thus provide a broad knowledge on many different domains.

However, QA over structured data has its roots in the late sixties as an attempt to make databases easily accessible even by non-expert users. For

this reason, QA systems were initially referred to as "Natural Language Interfaces".

Apart from the technical differences existing between KGs and databases, they still share the same properties hold by any structured resource: a Data Representation Language (DRL) allows describing the data in a data source, and a Data Query Language (DQL) is then used to retrieve the data. The standard DQL for databases is SQL, while its correspondent for KGs is SPARQL. The main goal of a QA system is to bridge the so-called *lexical gap* existing between the vocabulary adopted by the user and the labels used within the structured data source. In this way a QA system can allow users to have access to the information stored in the structured data source with no need for mastering a DQL: the system has to take over the management to this translation, hiding it to the user.

Due to its complexity, the majority of works available at the state of the art exploit a combination of several NLP techniques to process the question and transform it into its DQL equivalent. For this reason, the results available at the state of the art appear even more critical when looking for relevant solutions for non-English languages.

This problem is accentuated even more by the shortage of multilingual datasets. For example, the QALD evaluation campaign¹, starting from its third edition, has included a task for Multilingual Question Answering over DBpedia. The dataset created for this task provides each question in seven different languages (i.e. English, German, Spanish, Italian, French, Dutch, and Romanian) along with its SPARQL translation. Even if the dataset actually includes non-English languages, the SPARQL translation always makes use of the resources of the English version DBpedia since many properties and entities do not have a label for the aforementioned languages.

Copyright ©2019 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

¹<http://qald.aksw.org/>

Other datasets for Question Answering over Structured Data, like Simple Question (Bordes et al., 2015) and Web Question (Berant et al., 2013) are focused only on the English language and do not provide the translation for other languages. The same issue affects also the datasets available for the evaluation of Natural Language Interfaces for databases like the U.S. Geography database (Geoquery²) or IMDb³.

For all these reasons, there are only a few systems which propose an approach applicable for Italian. FuLL (Bombara et al., 2005) is a NLI for geographical data banks. FuLL exploits a fuzzy engine and a dialog manager to interpret the question inserted by the user and handle subjective elements (like the magnitude of adjectives) and ambiguous requests. However, in order to make the system more accurate, the authors have focused only over a specific domain.

QAnswer (Diefenbach et al., 2017) is one of the few QA systems with an architecture completely independent from the language thus it can process many different languages including Italian. The system splits the question in n-grams and tries to match them with the resources of the underlying knowledge graphs. Based on the retrieved resources, it generates all the possible queries that could satisfy the user's information need. Multilingualism is obtained by avoiding the usage of any NLP tool which could affect the performance of the system, especially for those languages where the accuracy of those tools is still very low. On the other hand, the main disadvantages of this approach are that the identification of relations is based just on the dictionary and the syntax of the question is ignored thus meaning that the lack of resources in a certain language can deeply affect the results.

Based on these observations, we decided to develop a QA system for the Italian language. Our approach is based on the one adopted in CANaLI (Mazzeo and Zaniolo, 2016) which obtain the best results within the QALD-6 evaluation campaign (Unger et al., 2016). CANaLI makes use of controlled natural languages and an auto-completion mechanism to guide the user toward the formulation of a natural language question which is then processed using a finite state automaton. By analyzing the advantages and the limitation of this ap-

proach, we developed a new system which is capable of reducing the lexical gap and extended it to cover the Italian language and to support queries over traditional databases.

The paper is organized as follows: in Section 2 we will introduce and describe our system MULIB a QA system capable of answering natural language questions written in Italian over an underlying structured data source, in Section 3 is described the evaluation we performed to assess MULIB's effectiveness, finally in Section 4 we will discuss the results obtained by MULIB and outline the future directions for our work.

2 Methodology

2.1 Bridging the lexical gap

As stated in Section 1, QA systems like CANaLI can achieve good results if the syntactic structure of the question is compliant with the controlled natural language.

The main drawback of this approach lies in the vocabulary that can be accepted by the finite state automaton. In fact, it is created by collecting the labels of the resources in the KG and a match exists only if there is a complete string matching, hence only those labels can be employed in the question. A simple example is represented by the question *Who is the writer of the Divine Comedy?*. Since there is no string matching between the words "writer" and the label of the property "author", CANaLI is not able to retrieve the right answer.

This method appears to be in contrast with what discussed in Section 1 regarding the lexical gap since it requires the user to know in advance how data is stored in the data source. In order to cope with this problem, we extended the vocabulary using an approach based on distributional semantics methods, i.e. Word2Vec (W2V) (Mikolov et al., 2013). The vector space was built upon Wikipedia abstracts in order to obtain representation which could be suitable with an open domain scenario. In this way, if the data source is changed, there is no need to re-train the model to adapt it to a specific topic. During the phrase mapping step, the system not only checks if there is a match with one of the labels of the KG like in its vanilla version, but it also computes a ranked list of phrases which are semantically similar to the original one. Therefore, the system substitutes in an iterative fashion the phrase in the question with the ones retrieved

²<http://www.cs.utexas.edu/users/ml/geo.html>

³<https://www.imdb.com/interfaces/>

A second problem occurs when the automaton enters a deadlock state. This happens when a token is misinterpreted, i.e. the automaton applies a wrong transition rule and shifts into a state where no other rules can be fired. For example, let us consider the question *Which are the prizes of Albert Einstein?*. After recognizing the starting phrase "Which are the", the automaton shifts in a state where it can accept an entity or a class. The word "prize" is erroneously matched by the system to the class `dbr:Prize` and so the automaton proceeds in the following state where, however, it can not accept an entity such as "Albert Einstein". For this reason, the procedure is forced to stop, returning as overall output an empty result set. To prevent this behavior, we introduced a backtracking algorithm that, in combination with the semantic matching mechanism described above, allows the automaton to reconsider the previous choices thus leading to the correct resource which is `dbr:award`.

The main problem to deal with in order to adapt this kind of solution for a different language, like Italian, is to modify the automaton since it is designed specifically keeping in mind the English grammar. For example, the English automaton was not able to recognize a question not beginning with a "question start" token, e.g.: *Give me the, Who is the, Is, Are*, while this syntactic structure is relatively common in Italian. To overcome this problem, we modified the transition rules related to the state S0 so that there is a transition to the state S1 either if a question token is recognized or if the first token represents an entity. In this way, we are capable to answer to question like *Matrix è un film?* (*Is Matrix a film?*) or *L'ordine #1123 è in stato concluso?* (*Order #1123 is in Finished state?*).

sition which allows it to shift from the state S_2 to the state S_1 if the incoming token is a property. This allowed a correct recognition of requests like: *Dammi tutti i film in lingua inglese.* (Give me all the English films). In Figure 1 is shown the updated version of the automaton i.e. capable to process sentences written in Italian.

One of the main features of MULIB is its capability to query not only Knowledge Graphs but also relational databases. In order to make a database compliant with the structure of the finite state automaton, we employ a particular framework called D2RQ (Bizer and Seaborne, 2004) which is developed under the Apache License⁴. This tool is essential for our system since the database, once converted using RDF can be queried both in English and Italian.

To express a join using D2RQ, it is necessary to create an object of type *d2rq:PropertyBridge* which allows creating a mapping between one or more database columns and a custom RDF property.

⁵<http://d2rq.org/d2rq-language>

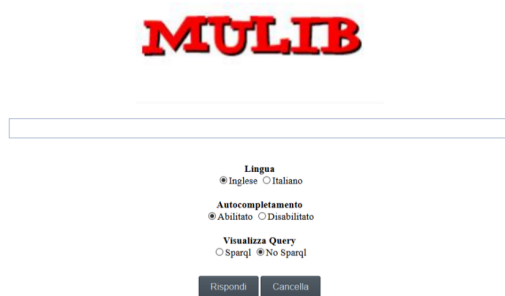


Figure 2: MULIB web interface.

2.4 Web Interface

We developed a Web Interface in order to allow users to interact with it and test the system in a real-world scenario (details about this experiment will be discussed in Section 3). A screenshot of the actual interface is shown in Figure 2.

We decided to design an interface as simple as possible in order to not insert elements which could confuse the users and make the interaction with the system unnecessarily difficult. The interface is composed of a text box, where the user can insert her questions and a list of options.

Since MULIB is a multilingual system, one of the options allows the user to switch from English to Italian. The system leaves to the user also the possibility to disable the auto-completion mechanism and freely insert a question without any suggestion. In this case, the system will bridge the lexical gap existing between the question inserted by the user and the database using W2V and the backtracking mechanism. Finally, the last option can enable the visualization of the SPARQL query which translates the question along with the final answer.

3 Evaluation

As stated in Section 1, in the literature there is a lack of resources for non-English languages which makes the creation and evaluation of novel approaches troublesome. It is very hard to create a solution completely language independent which allows achieving good results and NLP tools for English usually perform better than the others.

For the evaluation of our approach, we conducted an in-vivo experiment involving Sud Sistemi srl, a company that has expressed its willingness to participate in the experiment. The company made available one of its databases to be in-

tegrated and queried by MULIB. In this way, we could actually test the effectiveness of MULIB in a real-world scenario. Only the tables useful for the purposes of the experiment were used in the mapping, namely: Personal data, Articles, Agents. In the conversion, some fields were omitted, due to the sensitive data contained or to their limited significance with the purposes of the experiment.

The in-vivo experiment involved a total of 25 subjects. Participants were selected accordingly to their degree of knowledge with SQL so that the ratio between expert and non-expert user would be balanced. The experiment was composed of the following four phases:

- Phase 1: gathering personal information, i.e.: age and gender;
- Phase 2: gathering information about the participant's skills in IT and SQL;
- Phase 3: participants are asked to interact with the system and complete some simple tasks;
- Phase 4: survey about the system, to collect feedback coming from the participants.

From the second phase of the experiment emerged that the 52% of the participants declared that they had low-mid IT skills and the 48% of them declared having none or little knowledge of SQL.

During phase 4, we asked the participants to express their overall opinion about the system using a 10 point Likert scale, which ranged from a minimum of 1, that expressed the lowest liking, to a maximum of 10. The 80% of the participant assigned a score greater than five, thus corroborating the effectiveness of MULIB as a Natural Language Interface.

The usage of MULIB's web interface has been considered easy to use by the 76% of the participants, while the remaining 20% of them judged it of mid/high difficulty. This result underlines how the simplicity of the User Interface that we designed for MULIB has been appreciated by the participants. In particular, what has been judged positively by the users is the auto-completion interface, which can guide them through the interaction with the system and allows to reduce the number of mistakes.

We asked the users to select a preference between SQL and Natural Language when querying

the database after the interaction with the system. The majority of users expressed their preference for the natural language. This result is surely influenced by the presence among the participants of several users that have never used SQL, thus feeling more confident in using natural language rather than a DQL.

Another question asked if it was easy to perform the SQL *join* operation using natural language. The answer was affirmative in 89.5% of cases. In fact, thanks to D2RQ, a join is mapped to a simple property and make a question over a table which represents a join does not represent a problem. Of course, this flexibility can be obtained only by means of a careful mapping of the database structure to the final ontology.

The last set of questions was used to estimate to which extent MULIB could be useful within the context of a company. The 84% of participants think that a system like MULIB could actually be helpful and beneficial in such contexts, allowing to non-expert people to query the database without the need of knowing its underlying structure.

Finally, we asked the people involved in the experiment if MULIB managed to satisfy their information need and their expectations. In the case of a negative answer, we also proposed them to give us suggestions to improve the system. The 80% of participants declared that on average the system was able to satisfy their information need, while the remaining 20% was not completely satisfied and the main causes were the following: absence of data due to the General Data Protection Regulation, lack of aggregate data in the database, and failures caused by too complex queries.

Regarding the suggestions, they can be summarized in three main points: enhance the answer to the query with other details, make the system more flexible (i.e. extending the range of questions that the system can answer), and finally improve the User Interface of the system.

4 Results and Conclusions

From the answers to the questionnaires, it is clear that MULIB has been perceived positively by the users, which think that it would represent a powerful tool to support their interaction with a DBMS.

As future work, we could improve the graphical interface of our system, making it more appealing for the users and integrating some visualization tools which could help to provide a more complete

answer by integrating complementary information coming from the database.

In conclusion, in this paper, we have presented MULIB, a QA system for Structured Data which is capable to answer questions formulated in English and Italian. We decided to adopt an approach based on Controlled Natural Languages, i.e. the one adopted in systems like CANaLI. By the analysis of the shortcomings of this approach, we designed a specific solution aimed at overcoming them.

First of all we adopted distributional semantics principles in order to cope with the lexical gap and we modified the algorithm to cover the issue represented by ambiguous words. Next we extended the approach to cover also the Italian language and allow to query databases as well as Knowledge Graphs.

By performing an in-vivo experiment along with 25 participants, we could actually evaluate how helpful user perceive our system.

Acknowledgment

This work is partially funded by the “DECISION Data-drivEn Customer Service Innovation” project, POR Puglia FESR 2014-2020 - INNONET-WORK program - “Sostegno alle attività di R&S per lo sviluppo di nuove tecnologie sostenibili, di nuovi prodotti e servizi”.

References

- Jonathan Berant, Andrew Chou, Roy Frostig, and Percy Liang. 2013. Semantic parsing on freebase from question-answer pairs. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1533–1544.
- Christian Bizer and Andy Seaborne. 2004. D2rq-treating non-rdf databases as virtual rdf graphs. In *Proceedings of the 3rd international semantic web conference (ISWC2004)*, volume 2004. Proceedings of ISWC2004.
- Maurizio Bombara, Davide Calì, Ivana Calì, and Giuseppe Tropea. 2005. Servizi innovativi web gis: impiego di full (fuzzy logic and language) per l’accesso in linguaggio naturale ai db geografici. In *Proceedings of the 9th national conference of the Italian Federation of Scientific Associations for Territorial and Environmental Information (ASITA2005)*. Proceedings of ASITA2005.
- Antoine Bordes, Nicolas Usunier, Sumit Chopra, and Jason Weston. 2015. Large-scale simple question answering with memory networks. *arXiv preprint arXiv:1506.02075*.

- Dennis Diefenbach, Kamal Singh, and Pierre Maret. 2017. Wdaqua-core0: A question answering component for the research community. In *Semantic Web Evaluation Challenge*, pages 84–89. Springer.
- Giuseppe M Mazzeo and Carlo Zaniolo. 2016. Answering controlled natural language questions on rdf knowledge bases. In *EDBT*, pages 608–611.
- Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.
- Christina Unger, Axel-Cyrille Ngonga Ngomo, and Elena Cabrio. 2016. 6th open challenge on question answering over linked data (qald-6). In *Semantic Web Evaluation Challenge*, pages 171–177. Springer.

The Tenuousness of Lemmatization in Lexicon-based Sentiment Analysis

Marco Vassallo, Giuliano Gabrieli

CREA Research Centre
for Agricultural Policies and Bio-economy
{marco.vassallo,
giuliano.gabrieli}@crea.gov.it

Valerio Basile, Cristina Bosco

Department of Computer Science
University of Turin
{basile,bosco}@di.unito.it

Abstract

English. Sentiment Analysis (SA) based on an affective lexicon is popular because straightforward to implement and robust against data in specific, narrow domains. However, the morpho-syntactic pre-processing needed to match words in the affective lexicon (lemmatization in particular) may be prone to errors. In this paper, we show how such errors have a substantial and statistical significant impact on the performance of a simple dictionary-based SA model on data from Twitter in Italian. We test three pre-trained statistical models for lemmatization of Italian based on Universal Dependencies, and we propose a simple alternative to lemmatizing the tweets that achieves better polarity classification results.¹

1 Introduction

In the last few years a very large variety of approaches has been proposed for addressing Sentiment Analysis (SA) related tasks. In several approaches, lexical resources play a crucial role: they allow systems to move from strings of characters to the semantic knowledge found, e.g., in an affective lexicon². For achieving this result and calculating the polarity of sentiment, or of some related categories, some shallow morphological analysis has to be applied, which mostly consists in lemmatization.

When we refer to standard text, available resources and robust lemmatizers make lemmatization a practically solved issue, but the presence

of misspellings, lingo and irregularities makes the application of lemmatization on user-generated content drawn from social media and micro-blogs not equally easy.

A possible solution consists in applying supervised machine learning techniques in order to create robust lemmatization models. However, the large manually curated datasets necessary for this task are currently very rare, in particular for languages other than English. For what concerns Italian, a good quality gold standard resource in Universal Dependency has been released which includes texts drawn from micro-blogs, namely PoSTWITA-UD (Sanguinetti et al., 2018). Unfortunately it is not nearly large enough to be of practical use in a supervised machine learning setting.

In this paper, we focus on the lemmatization of social media texts, observing and evaluating its impact on SA. The goal of this work is to address the following research questions: *what is the impact of lemmatization in SA tasks? Can we classify lemmatization errors and automatically adjust (a relevant portion of) them?*

We start from the empirical evidence found in a corpus of tweets from the agriculture domain that has initially raised our attention on this problem. After that, we present further experiments on a manually annotated dataset. We further propose some hints about a solution based on an affecting lexicon of inflected forms.

2 Datasets

We collected two datasets of microblogs in Italian language, in order to experiment on realistic data.

AGRITREND is a corpus of Italian posts collected from the Twitter accounts of the main institutional and media actors related to the agricultural sector during the period of January-April 2019. The data related to the first two months of the year have been used for the publication of the

¹Copyright © 2019 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

²For an informal definition of *affective lexicon* see: <http://www.ai-lc.it/lessici-affettivi-per-litaliano/>

first issue of the Institutional bulletin of the CREA Research Centre for Agricultural Policies and Bio-economy (Monda et al., 2019). Institutional motivations drove the initiative of setting up this corpus: exploring the sentiment in agriculture and thus providing insights about current and emerging trends of the agricultural sector. The dataset is composed of 8,883 tweets, including 2,554 retweets (28.75% of the total).

SENTIPOLC is the corpus distributed for the SENTiment POLarity Classification task (Barbieri et al., 2016) within the context of the evaluation campaign EVALITA 2016³. The corpus, consisting of 9,392 tweets, was created partly by querying Twitter for specific keywords and hashtags marking political topics, and partly with random tweets on any topic. Experts and crowdsourcing contributors annotated the dataset with subjectivity (binary classification: objective/subjective), polarity (4-fold multiclass classification: positive/negative/neutral/mixed) and irony (binary classification: ironic/not-ironic).

3 Processing the *AGRITREND* corpus

In this section, we describe the processing applied on the *AGRITREND* with the goal of SA, after the pre-processing which consisted in filtering out hashtags, @mentions, URLs and tokenization.

3.1 Lexicon-based Sentiment Analysis

While most modern SA approaches are supervised⁴, our SA approach is unsupervised and based on an affective lexicon. However, given the narrow topic scope of our data of interest and the unavailability of annotated data for agriculture, the application of an unsupervised classifier allowed us to avoid domain adaptation issues. Moreover, the dictionary-based approach is more transparent, allowing us to evaluate its errors at a finer-grained lexical level.

The method is straightforward. Given a pre-processed tweet and an affective lexicon with lemmas paired to their polarity scores, we match the tokens in the tweet to their respective entries in the lexicon, and compute the sum of their values. We use *Sentix* (Basile and Nissim, 2013), an affective lexicon for Italian, created by the align-

ment of SentiWordNet (Baccianella et al., 2010) and the Italian section of MultiWordNet (Pianta et al., 2002). In particular, we adopt Sentix version 2.0⁵.

3.2 Lemmatization

In order to match the tweets' words with a Sentix entry, we need to transform them into their base forms, i.e., lemmatize the tweets. For this purpose *UDPipe* R package with the function *udpipe_annotate* was used, applying all the three available models for Italian language: *ISDT* (Italian-isdt-ud-2.3-181115), *POSTWITA* (Italian-postwita-ud-2.3-181115), and *PARTUT* (Italian-partut-ud-2.3-181115). *UDPipe* (Straka and Straková, 2017) is an end-to-end NLP pipeline including part-of-speech tagging and syntactic parsing with Universal Dependencies.

We ran the models on *AGRITREND*. In order to automatically estimate the quality of the lemmatization, the produced lemmas were checked against the Hoepli dictionary, a large, general-purpose online Italian dictionary comprising over 500,000 lemmas⁶. The results, in Table 2, show how the *UDPipe* models generated a substantial amount of improper Italian lemmas. Moreover, for each of the three models, a number between 20% and 30% of incorrect lemmas were generated correctly by at least one of the two other models.

In Table 1 an example is shown of the lemmatization according to the three models: among other errors, the named entity *Adige* was incorrectly lemmatized by all models.

3.3 Polarity detection

We compute the polarity of the lemmatized tweets, including wrong lemmatizations, by matching the produced lemmas in Sentix. Incorrect lemmatization, even for a single word, may cause serious distortions of the polarized scores. For instance, comparing the overall polarity scores calculated for the three models in Table 1, we can see that when *PARTUT* has been used, a wrong lemma (which is a non-existing verbal form of the noun *acqua* (water)) has been associated to the word *acqua* determining the attribution of negative rather than positive score. This phenomenon often occurs in *AGRITREND* regardless of the lemmatiza-

³<http://www.evalita.it/2016>

⁴Already in 2016, only one team out of 13 participated to the SENTIPOLC shared task on Italian SA with an unsupervised system.

⁵<https://github.com/valeriobasile/sentixR>

⁶<https://dizionari.repubblica.it/italiano.html>

Table 1: A tweet from *AGRITREND* with the output of the three UDpipe lemmatization models where the lemmas are alphabetically ordered and the errors marked in bold.

Original	@ANBI.Nazionale Allarme idrico. Dopo il Po anche l'Adige è in crisi d'acqua https://t.co/GLTlMNqzEv di @AgricolturaIT
ISDT	acqua adigire allarme crisi d dopo idrico po - Sentix score: 0.080
POSTWITA	acqua adigere allarme crisi di dopo idrico po - Sentix score: 0.080
PARTUT	acquare adigere allarme crisi d dopo idrico po - Sentix score: -0.078

Table 2: Number and rate of lemmas produced by the UDpipe lemmatization models and not found in the Hoepli dictionary.

Model	Incorrect lemmas	%
ISDT	19,707	44.5
POSTWITA	21,444	48.4
PARTUT	22,440	50.7

tion model applied. Table 3 shows the percentages of negative, neutral and positive tweets based on the assigned polarity for each model. Here we consider positive a tweet whose Sentix score is greater than zero, negative when lower than zero, and neutral if it is exactly zero.

Table 3: Polarity classification on *AGRITREND* lemmatized with different UDpipe models.

Model	Negative	Neutral	Positive
ISDT	32.6%	9.5%	57.9%
POSTWITA	32.3%	10.2%	57.5%
PARTUT	33.8%	11.1%	55.1%

At the first glance, from percentages only, we might argue that the lemmatization models, each one with its own bias, classified the tweets in a similar manner. However, at this step of analysis, we cannot say anything about statistical differences in the size and in the signs of the polarity scores between each model.

3.4 Statistical significance

If the differences between the scores were not statistically significant, the incorrect lemmatization should not impact on the polarity scores. Conversely, if significant differences exist, the lemmatization models will generate different polarity scores, severely affected by the incorrect lemmatization. In order to verify this hypothesis, we applied the non-parametric statistical signed rank test of Wilcoxon (1945) for paired samples to the polarity scores for each pair of models. This test is commonly used to verify if the difference between two scores from the same respondents (i.e., samples) is significantly different without the need for the data to follow a known probability distribution or high precision in the measures to be tested for.

In our case the samples are coupled, since they are composed of the same tweets with potential different lemmas and the scores are the polarity of the tweets after lemmatization. As a consequence, the test is able to simply evaluate if the difference between the polarity of the tweets is due to the sign and the magnitude of the score simultaneously. The results of the Wilcoxon test, computed with the statistical package SPSS, are presented in Table 4.

The results of the Wilcoxon test are not statistically significant between ISDT and POSTWITA. The polarity obtained with the PARTUT lemmatization is significantly different from the other two, in line with the observation of a higher number of incorrect lemmas (51%, see Table 2). The result of this test indicates that an incorrect lemmatization produces *statistically significant* differences between the subsequent polarity scores and confirms our hypothesis.

4 Experiments on *SENTIPOLC*

In the previous section, we analyzed the lemmatization errors produced by three UDpipe models on *AGRITREND* and we observed how statistically significant is the failure in lemmatization on the result of dictionary-based SA. Nevertheless, being the *AGRITREND* corpus not annotated for sentiment polarity, we could not say anything about the *accuracy* of the prediction. To bridge this gap, we repeated the experiment on *SENTIPOLC*, where ground truth labels (also called *gold standard labels*) were manually annotated, starting by running the same processing pipeline as for *AGRITREND*. Table 5 shows an example tweet with the corresponding polarity scores. In this dataset, the percentages of incorrect lemmas, according to the Hoepli dictionary, is generally smaller than in the *AGRITREND* data, but still substantial: 35% for ISDT, 41% for POSTWITA, 44% for PARTUT (see Table 2 for a comparison with the other dataset).

Comparing the predictions obtained with Sentix with the labels annotated in *SENTIPOLC*, we eval-

Table 4: Wilcoxon signed rank test results between pairs of UDPipe models.

	ISDT vs. POSTWITA	ISDT vs. PARTUT	POSTWITA vs. PARTUT
Standardized test statistic	-1.317	-6.996	6.208
Asymptotic Sign. (2-sided test)	0.188 ($p > 0.05$)	0.000 ($p < 0.05$)	0.000 ($p < 0.05$)
Positive differences	2,190	2,250	2,913
Negative differences	2,281	2,824	2,404
Number of Ties	4,412	3,809	3,566

Table 5: Example tweet from *SENTIPOLC* with the output of three UDPipe lemmatization models. The lemmas are ordered alphabetically, since they are further processed as a bag of words.

Original text	Capitale Europea della Cultura che combacia con la fine consultazioni de #labuonascuola: gran bel segnale :)
Bag of words	bel Capitale combacia consultazioni Cultura della Europea fine gran segnale
ISDT	bello capitale combaciare consultazione cultura di europeo fine grande segnale - Sentix score: 0,8449
POSTWITA	bello capitale combaciare consultazione cultura da europeo fine grande segnale - Sentix score: 1,0739
PARTUT	bel capitale combacia consultazione cultura dere europeo fine grande segnale - Sentix score: -0,2715

Model	F1 (pos.)	F1 (neg.)	F1 (avg.)
ISDT	0.404	0.535	0.470
POSTWITA	0.414	0.540	0.477
PARTUT	0.409	0.540	0.474

Table 6: Performance of the dictionary-based SA, with different lemmatization models.

uate the performance of the dictionary-based approach in terms of precision, recall, F1-measure, and thus simultaneously measuring the impact of the different lemmatization models on the prediction accuracy. The results are shown in Table 6, in terms of F1-score for the positive polarity, negative polarity, and their average, following the official evaluation metrics of the *SENTIPOLC* task. The Wilcoxon test applied on *SENTIPOLC* gave very similar results to those achieved on *AGRITREND*, confirming the similarity of the classification obtained with ISDT and POSTWITA, while PARTUT tends to stand apart. Moreover, errors in lemmatization have a statistically significant impact on the SA on the *SENTIPOLC* dataset to the same extent as *AGRITREND*.

5 Morphologically-inflected Affective Lexicon

The analyses presented in the previous sections highlight how low coverage and errors in lemmatization have a negative impact on the performance of downstream tasks such as SA. In an attempt to mitigate this issue, we propose an alternative approach to link the lexical items found in tweets with the entries of an affective lexicon such as Sentix without an explicit lemmatization step.

We expand the lexicon by considering all the acceptable forms of its lemmas. Each form takes the

same polarity score of the original lemma. When different lemmas can assume the same form, we assign it the arithmetic mean of the lemmas' polarity scores. We use the *morph-it* morphological resource for Italian (Zanchetta and Baroni, 2005) to extract all possible forms from the lemmas of Sentix 2.0, and create a Morphologically-inflected Affective Lexicon (MAL) of Italian. The MAL comprises 148,867 forms, more than three times the size of Sentix 2.0 (41,800 lemmas).

The classification performance obtained using the MAL instead of a lemmatization model is in line with the results of the experiment in Table 6: 0.408 F1 (positive), 0.542 F1 (negative), and 0.475 F1 (average). However, so far we have employed a heuristic to map the Sentix score to polarity classes which is highly polarizing, that is, only tweets with an exact score of zero are classified as neutral. We therefore investigated a more conservative approach, where a parametric threshold T is introduced. After computing the polarity score of a message by summing up the polarity of its constituent words (or lemmas), we assign it a positive polarity label if the score is greater than T and negative if the score is lower than $-T$. The results of this experiment are shown in Figure 1. Several observations can be drawn from these results. First, using a threshold to assign polarity classes is indeed beneficial, with the right threshold empirically estimated around 5. Second, using the MAL instead of a lemmatization step improves the SA performance overall, in particular due to a better prediction of the negative polarity. Finally, the variation in threshold has opposite impact on the prediction of negative and positive tweets. We speculate that this may be due to asymmetries in

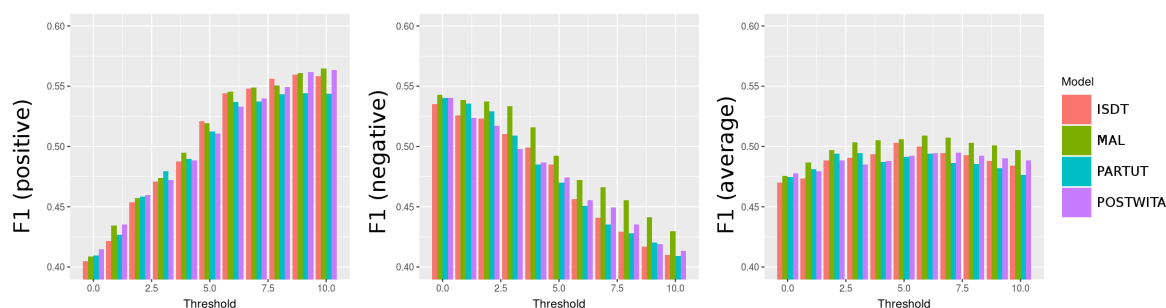


Figure 1: F1-score for the positive polarity (right), negative polarity (center) and average F1 (left) of the prediction of the dictionary-based SA approach on the *SENTIPOLC* test set.

the data, in the lexicon, or both, and intend to carry out future studies to understand this result.

6 Discussion

Our empirical study highlights important issues arising from language analysis errors (in lemmatization, in particular) propagating down the pipeline of a simple dictionary-based SA model. Without double-checking the outcome of the lemmatization step against a dictionary, a significant amount of noise is introduced in the system, leading to unstable results. The problem is even more substantial when dealing with data in a specific domain, such as the *AGRITREND* dataset of tweets about the agricultural domain, which indeed raised our attention on this problem.

We confronted the POS distribution of the parsed Agritrend and SENTIPOLC corpora with the set of UD-parsed corpora in Italian. In the Twitter data, content words are slightly more prominent, while function words are less present, although the general POS distributions have similar shapes. We report however an inverse correlation between the correctness of the lemmatization and the frequency of the POS, that is, words with infrequent POS are more likely to be wrongly lemmatized.

We tested the performance in a setting with no lemmatization at all, and measured a relatively good performance on the *SENTIPOLC* benchmark with some of the parameter configurations. This is unsurprising, following our observations on the significant impact of incorrect lemmatization on the SA performance. However, such a setting is linguistically questionable (matching only an arbitrary subset of words in a lemma-based resources) and its results are highly variable.

It is also important to notice that an incorrect

lemmatization is likely hurtful not only to SA. The high reported number of non-existent lemmas created by the UDpipe models may severely alter the results of large-scale statistical studies on social media data, such as the ones planned by the creators of the *AGRITREND* data. Moreover, evaluating the correctness of a word by checking an external dictionary (in our case, Hoepli), is sensible to potential drawbacks of that resource, e.g., leading to overestimating lemmatization errors.

In sum, when choosing a pre-processing strategy for dictionary-based SA, the need arises to strike a balance between two extremes: 1) potentially incorrect lemmatization provided by a statistical model, that possibly *underestimates* the polarity; 2) an inclusive approach like MAL, that possibly *overestimates* the polarity.

7 Conclusion and Future Work

In this paper, we presented an empirical and statistical study on the impact of lemmatization on a NLP pipeline for SA based on an affective lexicon. We found that lemmatization tools need to be used carefully, in order to not introduce too much noise, deteriorating the performance downstream. Then we propose an alternative approach that skips the lemmatization step in favor of a morphologically rich affective resource, in order to alleviate some of the observed issues.⁷ We plan on integrating the proposed solutions, including the MAL and an automatic check of the lemma produced by UDpipe, in a pre-processing pipeline based on UDpipe.

⁷The MAL is available for download at <https://github.com/valeriobasile/sentixR/blob/master/sentix/inst/extdata/MAL.tsv>

Acknowledgments

The work of Marco Vassallo and Giuliano Gabrieli is funded by the Statistical Office of CREA. The work of Valerio Basile and Cristina Bosco is partially funded by Progetto di Ateneo/CSP 2016 (*Immigrants, Hate and Prejudice in Social Media*, S1618.L2_BOSC_01).

References

- Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. 2010. SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta, May. European Languages Resources Association (ELRA).
- Francesco Barbieri, Valerio Basile, Danilo Croce, Malvina Nissim, Nicole Novielli, and Viviana Patti. 2016. Overview of the Evalita 2016 SENTiment POLarity Classification Task. In *Proceedings of Third Italian Conference on Computational Linguistics (CLiC-it 2016) & Fifth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2016)*, Naples, Italy, December.
- Valerio Basile and Malvina Nissim. 2013. Sentiment analysis on Italian tweets. In *Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 100–107.
- Mafalda Monda, Giuliano Gabrieli, and Marco Vassallo. 2019. Sentiment in agricoltura: Il termometro dell'agricoltura - i principali temi discussi su Twitter e gli umori degli addetti. In *I numeri dell'Agricoltura Italiana*. CREA, Centro Politiche e Bio-economia, June.
- Emanuele Pianta, Luisa Bentivogli, and Christian Giardi. 2002. Multiwordnet: developing an aligned multilingual database. In *Proceedings of the First International Conference on Global WordNet*, January.
- Manuela Sanguinetti, Cristina Bosco, Alberto Lavelli, Alessandro Mazzei, Oronzo Antonelli, and Fabio Tamburini. 2018. PoSTWITA-UD: an Italian Twitter treebank in Universal Dependencies. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*, Miyazaki, Japan, May. European Languages Resources Association (ELRA).
- Milan Straka and Jana Straková. 2017. Tokenizing, pos tagging, lemmatizing and parsing ud 2.0 with UDPipe. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 88–99, Vancouver, Canada, August. Association for Computational Linguistics.
- Frank Wilcoxon. 1945. Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6):80–83.
- Eros Zanchetta and Marco Baroni. 2005. Morph-it! a free corpus-based morphological resource for the Italian language. *Corpus Linguistics 2005*, 1(1).

Author Index

- Abbruzzese, Roberto, 417
Adorni, Giovanni, 14
Agirre, Eneko, 345
Agnello, Patrizia, 436
Ahmadi, Sina, 367
Alfano, Domenico, 417
Alloatti, Francesca, 429
Alzetta, Chiara, 14
Ansaldi, Silvia Maria, 436
Attardi, Giuseppe, 508
- Bacciu, Davide, 70
Ballarè, Silvia, 243
Basile, Pierpaolo, 312, 423, 514
Basile, Valerio, 304, 312, 442, 520
Basili, Roberto, 454
Battisti, Alessia, 22
Bellomaria, Valentina, 29, 187
Bernardi, Raffaella, 1, 361
Bernareggi, Cristian, 250
Bianchi, Federico, 448
Bolioli, Andrea, 429
Boschetti, Federico, 35
Bosco, Cristina, 98, 151, 172, 304, 442, 520
Brambilla, Silvia, 41
Brandi, Giorgio, 454
Brunato, Dominique, 49, 92, 166
Bruno, Simone, 436
Buongiovanni, Chiara, 49
Busetto, Nicolò, 55, 143
- Cabrio, Elena, 129
Cafagna, Michele, 63, 70
Cammisa, Marco, 298
Capozzi, Arthur T. E., 442
Cappetta, Donato, 417
Caputo, Annalina, 423
Carfora, Valentina, 460
Carlino, Carola, 367
Caselli, Tommaso, 77
Castellucci, Giuseppe, 29
Catellani, Patrizia, 460
Cenceschi, Sonia, 85
Cerruti, Massimo, 243
Chesi, Cristiano, 207, 396
Chiriatti, Giulia, 92
Cignarella, Alessandra Teresa, 98
Cimino, Andrea, 484
- Cirillo, Nicola, 106
Colla, Davide, 113
Combei, Claudia Roberta, 121
Corazza, Michele, 129
Coro, Gianpaolo, 136
Cremaschi, Marco, 448
Croce, Danilo, 454
- D'Aoust, Renée E., 502
Damiano, Emanuele, 194
De Gasperis, Giovanni, 257
de Gemmis, Marco, 312, 478
De Mattei, Lorenzo, 63, 70
Declerck, Thierry, 339
Dell'Orletta, Felice, 14, 49, 92, 159, 166, 484
Delmonte, Rodolfo, 55, 143
Demartini, Silvia, 502
Di Iorio, Ernesto, 298
Di Massimo, Francesca, 460
Di Nuovo, Elisa, 151
Dominutti, Elisa, 159
Ducceschi, Luca, 353
- Ebling, Sarah, 22
Elia, Annibale, 403
Esposito, Massimo, 194
- Faralli, Stefano, 466
Favalli, Andrea, 29, 187
Fernández, Raquel, 11, 361
Fieromonte, Martina, 166
Finocchi, Irene, 466
Fiorentini, Antonio, 298
Francesconi, Chiara, 172
Franzini, Greta, 179
- Gabrieli, Giuliano, 520
Garda, Samuele, 496
Gatto, Maristella, 472
Gerevini, Alfonso, 490
Gerevini, Alfonso Emilio, 332
Giannone, Cristina, 187
Globo, Achille, 298
Goria, Eugenio, 243
Gracci, Francesco, 49
Gravina, Alessio, 508
Greco, Matteo, 207
Guadalupi, Mariafrancesca, 429

Guarasci, Raffaele, 194
 Herbelot, Aurélie, 325, 353
 Hovy, Dirk, 13
 Iovine, Andrea, 478
 Jezek, Elisabetta, 228
 Karakanta, Alina, 201
 Koceva, Frosina, 14
 Labruna, Paolo, 484
 Lai, Mirko, 442
 Lanzi, Roberta Iolanda, 429
 Laudanna, Alessandro, 215
 Lavelli, Alberto, 332, 490
 Lenci, Alessandro, 270
 Leontino, Marco, 113
 Lopez de Lacalle, Oier, 345
 Lorusso, Paolo, 207
 Losio, Davide, 85
 Luchetti, Mauro, 85
 Macchiarulo, Nicola, 298
 Mancuso, Azzurra, 215
 Manna, Raffaele, 221
 Marini, Costanza, 228
 Maslennikova, Aleksandra, 484
 Mastronardo, Claudio, 236
 Mauri, Caterina, 243
 Maurino, Andrea, 448
 Mazzei, Alessandro, 151, 250, 318
 Mehmood, Tahir, 490
 Menini, Stefano, 41, 129, 257
 Mennitti, Matteo, 514
 Mensa, Enrico, 113
 Meo, Rosa, 318
 Miaschi, Alessio, 14
 Mihaljević, Josip, 264
 Miliani, Martina, 270
 Minutolo, Aniello, 194
 Moneglia, Massimo, 278
 Montemagni, Simonetta, 159
 Monti, Johanna, 221
 Monticone, Michele, 250
 Moretti, Giovanni, 374, 388
 Moro, Andrea, 12, 207
 Musto, Cataldo, 285, 442
 Namor, Ivan, 496
 Narducci, Fedelucio, 478
 Navigli, Roberto, 1
 Negri, Matteo, 201
 Nieuwenhuis, Moniek, 291
 Nissim, Malvina, 63, 70, 291
 Pagano, Pasquale, 136
 Palmero Aprosio, Alessio, 41, 403
 Panichi, Giancarlo, 136
 Panunzi, Alessandro, 278
 Pardelli, Gabriella, 35
 Pascucci, Antonio, 221
 Passalacqua, Samuele, 14
 Passaro, Lucia C., 270
 Passarotti, Marco, 179, 374
 Patti, Viviana, 304, 442
 Paziienza, Andrea, 298
 Pericolo, Chiara, 106
 Peverelli, Andrea, 179
 Piastra, Marco, 460
 Pierotti, Andrea Primo, 448
 Pifferi, Lucia, 159
 Poletto, Fabio, 172, 304, 442
 Polignano, Marco, 285, 312, 442
 Ponzetto, Simone Paolo, 466
 Porporato, Aureliano, 318
 Pregnolato, Giorgia, 429
 Preissner, Simon, 325
 Puccinelli, Daniele, 502
 Putelli, Luca, 332
 Quochi, Valeria, 159
 Racioppa, Stefania, 339
 Radicioni, Daniele P., 113, 318
 Ravelli, Andrea Amelio, 345
 Rigutini, Leonardo, 298
 Romagnoli, Raniero, 29, 187
 Rossetto, Federico, 508
 Rosso, Paolo, 98
 Ruffo, Giancarlo, 442
 Ruffolo, Paolo, 179
 Sanguinetti, Manuela, 98, 151, 172, 442
 Sanna, Helena, 179
 Sansonetti, Angelo, 285
 Sbattella, Licia, 85
 Seltsmann, Johann, 353
 Semeraro, Giovanni, 1, 285, 312, 423, 442, 514
 Serina, Ivan, 332, 490
 Severini, Silvia, 508

Shekhar, Ravi, 361
Siciliani, Lucia, 514
Signoroni, Edoardo, 179
Speranza, Giulia, 367
Sprugnoli, Rachele, 374, 388
Stede, Manfred, 496
Stranisci, Marco, 285, 304, 442
Suriano, Francesco, 243

Tamburini, Fabio, 236
Tavosanis, Mirko, 381
Tedesco, Roberto, 85
Testoni, Alberto, 361
Tonelli, Sara, 129, 257, 388, 403
Topciu, Kledia, 396
Torre, Ilaria, 14
Totis, Pietro, 496
Trevisi, Antonio, 298
Trotta, Daniela, 403
Tufano, Pasquale, 106
Turchi, Marco, 201

Turolla, Andrea, 429

Üstün, Ahmet, 77

Varvara, Rossella, 278
Vassallo, Marco, 520
Velardi, Paola, 466
Ventura, Viviana, 179
Venturi, Giulia, 35, 92, 166
Villata, Serena, 129
Vittorini, Pierpaolo, 257
Vitulano, Felice, 298
Volk, Martin, 22

Yavuz, Mehmet Can, 410

Zampedri, Federica, 179
Zanzotto, Fabio Massimo, 436