COLING 2018

The 27th International Conference on Computational Linguistics

Proceedings of the First Workshop on Linguistic Resources for Natural Language Processing (LR4NLP-2018)

August 20, 2018
Santa Fe, New Mexico, USA

Order copies of this and other ACL proceedings from:

Association for Computational Linguistics (ACL)
209 N. Eighth Street
Stroudsburg, PA 18360
USA
Tel: +1-570-476-8006
Fax: +1-570-476-0860
acl@aclweb.org

ISBN 978-1-948087-54-4

ISBN 978-1-948087-54-4
Anabela Barreiro, Kristina Kocijan, Peter Machonis and Max Silberztein (eds.)

Proceedings of the First Workshop on Linguistic Resources for Natural Language Processing

Preface

Linguists and developers of NLP software have been working separately for many years. Since stochastic methods, such as statistical and neural-network based parsers, have shown to be overwhelmingly successful in the software industry, NLP researchers have typically turned their focus towards technical issues specific to stochastic methods, such as improving recall and precision, and developing larger and larger training corpora. At the same time, linguists kept focusing on problems related to the development of exhaustive and precise resources that are mainly "neutral" vis-a-vis any NLP application, such as parsing and generating sentences.

However, recent progress in both fields has been reducing many of these differences, with large-coverage linguistic resources being used more and more by robust NLP software. For instance, NLP researchers now use large dictionaries of multiword units and expressions, and several linguistic experiments have shown the feasibility of using large phrase-structure grammars (a priori used for text parsing) in "generation" mode to automatically produce paraphrases of sentences that are described by these grammars.

The First Workshop on Linguistic Resources for Natural Language Processing (LR4NLP) of the 27th International Conference on Computational Linguistics (COLING 2018) held at Santa Fe, New Mexico, August 20, 2018, brought together participants interested in developing large-coverage linguistic resources and researchers with an interest in developing real-world Natural Language Processing (NLP) software. The presentations at the LR4NLP Workshop were organized into four sessions, as follows:

- Clash of the Titans: Linguistics vs. Statistics vs. Neural Networks
- May the Force Be with NooJ
- One for the Road: Monolingual Resources
- Language Resources Without Borders

The first session, Clash of the Titans: Linguistics vs. Statistics vs. Neural Networks, focused on linguistic and stochastic approaches and results. Our invited speaker, Mark Liberman, showed how semi-automatic analysis of large digital speech collections is transforming the science of phonetics, and offered exciting opportunities to researchers in other fields, such as the possibility of improving parsing algorithms by incorporating features from speech as well as text. He was followed by Silberztein, who presented a series of experiments aimed at evaluating reference corpora, such as the Open American National Corpus, and proposed a series of tasks to enhance them. Zhang & Moldovan then made an assessment on the limitations and strengths of neural net systems to rule-based systems on Semantic Textual Similarity by comparing its performance with traditional rule-based systems against the SemEval 2012 benchmark.

Several workshop participants have been using the NooJ software to develop the large-coverage linguistic resources needed by their NLP applications. NooJ was particularly germane to this workshop, because it is not only being used by linguists to develop resources in the form of electronic dictionaries, and morphological and syntactic grammars, but by computational linguists to parse and annotate large corpora, as well as by software engineers to develop NLP applications. Thus, we allocated the entire second session, May the Force Be with NooJ, to researchers using this platform. Machonis showed how a lexicon grammar dictionary of English phrasal verbs can be transformed into a NooJ dictionary, in order to accurately identify these structures in large corpora. Phrasal verbs are located by means

of a grammar, and the results are then refined with a series of dictionaries, disambiguating grammars, and filters. Likewise, Kocijan et al. demonstrated how they use NooJ to detect and describe the major derivational processes used in the formation of perfective, imperfective, and bi-aspectual Croatian verbs. Annotated chains are exported into a format adequate for a web-based system and further used to enhance the aspectual and derivational information for each verb. Next, Boudhina & Fehri presented a rule-based system for disambiguating French locative verbs in order to accurately translate them into Arabic. They used the Dubois & Dubois Charlier French Verb dictionary, a set of French syntactic grammars, as well as a bilingual French-Arabic dictionary developed within the NooJ platform. Finally, Rodrigo et al. presented a NooJ application aimed at teaching Spanish as a foreign language to native speakers of Italian. Their presentation included an analysis of a journalistic corpus over a thirty-year time span focusing on adjectives used in the Argentine Rioplatense variety of Spanish.

In the third session, One for the Road: Monolingual Resources, researchers examined a variety of large-coverage, monolingual linguistic resources for NLP applications. Dorr & Voss described the linguistic resource STYLUS (SysTematicallY Derived Language USe), which they produced through extraction of a set of argument realizations from lexical-semantic representations for a range of 500 English verb classes. Their Verb Database contains a total of 9,525 entries and includes information about components of meaning and collocations. STYLUS enables systematic derivation of regular patterns of language usage without requiring manual annotation. Then, Gezmu et al. presented a corpus of contemporary Amharic, automatically tagged for morpho-syntactic information. Texts were collected from 25,199 documents from different domains and about 24 million orthographic words were tokenized. Malireddy et al. discussed a new summarization technique, called Telegraphic Summarization, that, instead of selecting whole sentences, picks short segments of text spread across sentences in order to build the resulting summary. They proposed a set of guidelines to create such summaries and annotated a gold corpus of 200 English short stories. Finally, Abera et al. described the procedures that were used for the creation of the first speech corpus of Tigrinya a Semitic language spoken in the Horn of Africa for speech recognition purposes.

The closing session, Language Resources Without Borders, focused on the development of large-coverage, multilingual linguistic resources for Machine Translation (MT). Abate et al. described the development of parallel corpora for five Ethiopian Languages Amharic, Tigrigna, Afan-Oromo, Wolaytta and Geez. The authors conducted statistical machine translation experiments for seven language pairs that showed that the morphological complexity of these languages has a negative impact on the performance of the translation, especially for the target languages. Then, using the FrameNet and SALSA corpora, Sikos & Padó examined English and German, highlighting how inferences can be made about cross-lingual frame applicability using a vector space model. They showed how multilingual vector representations of frames learned from manually annotated corpora can address the need of accessing broad-coverage resources for any language pair. Next, Zhai et al. presented a parallel multilingual oral corpus the TED Talks in English, French, and Chinese. The authors categorized and annotated translation relations, to distinguish literal translation from other translation techniques. They developed a classifier to automatically detect these relations, with the long-term objective being to have better semantic control when dealing with paraphrases or translational equivalencies. Tomokiyo et al. aimed at improving the Cesselin, a well-known, open source Japanese-French dictionary. They hypothesized that the degree of lexical similarity between results of MT into a third language might provide insight on how to better annotate proverbs, idiomatic constructions, and phrases containing quantifiers. To test this, they used Google Translate to translate both the Cesselin Japanese expressions and their French translations into English. Their results showed much promise, in particular for distinguishing normal usage from idiomatic examples. Barreiro & Batista presented a detailed analysis on Portuguese contractions in an aligned bilingual Portuguese-English corpus and argued that the choice to decompose contractions or not depended on their context, for which the occurrence of multiword units is key. Finally, Dhar et al. presented a newly created parallel corpus of English and code-mixed English-Hindi. Using 6,088 code-mixed English-Hindi sentences previously available, they created a parallel English corpus using human translators. They then presented a technique to augment

run-of-the-mill MT approaches, which achieves superior translations without the need for specially designed translation systems, and which can be plugged into any existing MT system.

The common theme of all of the papers presented in this workshop was how to build large linguistic resources in the form of annotated corpora, dictionaries, and morphological and syntactic grammars that can be used by NLP applications. Linguists as well as Computational Linguists who work on NLP applications based on linguistic methods will find advanced, up-to-the-minute studies for these themes in this volume. We hope that readers will appreciate the importance of this volume, both for the intrinsic value of each linguistic formalization and the underlying methodology, as well as for the potential for developing automatic NLP applications.

Editors:
Anabela Barreiro, INESC-ID, Lisbon, Portugal
Kristina Kocijan, University of Zagreb, Zagreb, Croatia
Peter Machonis, Florida International University, Miami, USA
Max Silberztein, Université de Franche-Comté, Besançon, France

Organizers and Review Committee

Workshop Organizers

Anabela Barreiro, INESC-ID, Lisbon, Portugal
Kristina Kocijan, University of Zagreb, Croatia
Peter Machonis, Florida International University, USA
Max Silberztein, Université de Franche-Comté, France

Peer Review Committee

Program Committee Chair

Max Silberztein, Université de Franche-Comté, France

Jorge Baptista, University of Algarve, Portugal
Anabela Barreiro, INESC-ID Lisbon, Portugal
Xavier Blanco, Autonomous University of Barcelona, Spain
Nicoletta Calzolari, Istituto di Linguistica Computazionale, Italy
Christiane Fellbaum, Princeton University, USA
Héla Fehri, University of Sfax, Tunisia
Yuras Hetsevich, National Academy of Sciences, Belarus
Kristina Kocijan, University of Zagreb, Croatia
Mark Liberman, University of Pennsylvania, USA
Elena Lloret Pastor, Universidad de Alicante, Spain
Peter Machonis, Florida International University, USA
Slim Mesfar, Carthaga University, Tunisia
Simon Mille, Universitat Pompeu Fabra, Spain
Mario Monteleone, University of Salerno, Italy
Johanna Monti, University of Naples - L'Orientale, Italy
Bernard Scott, Logos Institute, USA

Invited Speaker

Mark Liberman, University of Pennsylvania, USA

Session Chairs

Anabela Barreiro, INESC-ID, Lisbon, Portugal
Kristina Kocijan, University of Zagreb, Croatia
Peter Machonis, Florida International University, USA
Max Silberztein, Université de Franche-Comté, France

Table of Contents

Conference Program

Monday, August 20, 2018

9:00–10:30 Session S1: Clash of the Titans: Linguistics vs. Statistics vs. Neural-nets

9:10–9:50 *Corpus Phonetics: Past, Present, and Future*
Mark Liberman

9:50–10:10 *Using Linguistic Resources to Evaluate the Quality of Annotated Corpora*
Max Silberztein

10:10–10:30 *Rule-based vs. Neural Net Approaches to Semantic Textual Similarity*
Linrui Zhang and Dan Moldovan

11:00–12:20 Session S2: May the Force Be with NooJ

11:00–11:20 *Linguistic Resources for Phrasal Verb Identification*
Peter Machonis

11:20–11:40 *Designing a Croatian Aspectual Derivatives Dictionary: Preliminary Stages*
Kristina Kocijan, Krešimir Šojat and Dario Poljak

11:40–12:00 *A Rule-Based System for Disambiguating French Locative Verbs and Their Translation into Arabic*
Safa Boudhina and Héla Fehri

12:00–12:20 *A Pedagogical Application of NooJ in Language Teaching: The Adjective in Spanish and Italian*
Andrea Rodrigo, Mario Monteleone and Silvia Reyes

Monday, August 20, 2018 (continued)

14:00–15:20 Session S3: One for the Road: Monolingual Resources

14:00–14:20 *STYLUS: A Resource for Systematically Derived Language Usage*
Bonnie Dorr and Clare Voss

14:20–14:40 *Contemporary Amharic Corpus: Automatically Morpho-Syntactically Tagged Amharic Corpus*
Andargachew Mekonnen Gezmu, Binyam Ephrem Seyoum, Michael Gasser and Andreas Nürnberger

14:40–15:00 *Gold Corpus for Telegraphic Summarization*
Chanakya Malireddy, Srivenkata N M Somisetty and Manish Shrivastava

15:00–15:20 *Design of a Tigrinya Language Speech Corpus for Speech Recognition*
Hafte Abera and Sebsibe H/Mariam

16:00–18:00 Session S4: Language Resources without Borders

16:00–16:20 *Parallel Corpora for bi-Directional Statistical Machine Translation for Seven Ethiopian Language Pairs*
Solomon Teferra Abate, Michael Melese, Martha Yifiru Tachbelie, Million Meshesha, Solomon Atinafu, Wondwossen Mulugeta, Yaregal Assabie, Hafte Abera, Binyam Ephrem, Tewodros Abebe, Wondimagegnhue Tsegaye, Amanuel Lemma, Tsegaye Andargie and Seifedin Shifaw

16:20–16:40 *Using Embeddings to Compare FrameNet Frames Across Languages*
Jennifer Sikos and Sebastian Padó

16:40–17:00 *Construction of a Multilingual Corpus Annotated with Translation Relations*
Yuming Zhai, Aurélien Max and Anne Vilnat

17:00–17:20 *Towards an Automatic Classification of Illustrative Examples in a Large Japanese-French Dictionary Obtained by OCR*
Christian Boitet, Mathieu Mangeot and Mutsuko Tomokiyo

17:20–17:40 *Contractions: To Align or Not to Align, That Is the Question*
Anabela Barreiro and Fernando Batista

17:40–18:00 *Enabling Code-Mixed Translation: Parallel Corpus Creation and MT Augmentation Approach*
Mrinal Dhar, Vaibhav Kumar and Manish Shrivastava