COLING 2016

The 26th International Conference on Computational Linguistics (COLING 2016)

Proceedings of the 5th Workshop on Cognitive Aspects of the Lexicon (CogALex-V)

December 12, 2016
Osaka, Japan

ISBN978-4-87974-721-1

Preface

Whenever we read a book, write a letter or perform a (web or dictionary) search, we always use lexical items (words or more complex constructions), the expressive shorthand versions of more or less abstract thoughts. Yet, lexemes are not only expressive means, i.e. vehicles transporting thoughts, they are also means to conceive them. They are mediators between language and thought, allowing us to move quickly from one idea to another, summarizing, expanding or specifying possibly underspecified thoughts. Of course, lexemes can do a lot more, allowing us to organize, memorize and access knowledge, and even reveal hidden meanings via information contained in the target or its surrounding words (subliminal communication). No doubt the lexicon is a key component of language.

Lexical items are generally viewed as *objects*, yet when it comes to speaking, reading or writing, they are *processes*, that is, they carry meaning which they convey via an abstract (part of speech) and concrete form (phonemes, graphemes). Obviously, in order to access a lexeme (at least) its form must be stored. This is done *holistically* in the case of external resources (paper or electronic dictionary), which represent word forms as single tokens, and *distributed* in the case of the human brain, which decomposes the form into syllables and phonemes. While knowing the form is important, its storage is by no means sufficient. We may still fail to access it when needed. More importantly, when consulting a dictionary (off-line processing) other types of knowledge are used, most prominently meta-knowledge and cognitive states. *Meta-knowledge* is revealed by the fact that search is generally initiated via a close neighbor of the target lexeme, to be continued then via one of the links connecting the source and the target (synonym, hypernym, etc.). *Cognitive states* are revealed by the information given at the onset of the search. They express the information available at that very moment. As psychologists have shown, authors always know something concerning the eluding word. Alas, what information is available when being in this state varies from person to person and from moment to moment. It is unpredictable knowledge. Hence the importance of building a resource flexible enough to accommodate any of them, allowing the user to start from anywhere and to access the target via many diverse routes.

Ironically, although – once expressed – many of these observations sound obvious, most of them have been overlooked by the various communities dealing with the lexicon, its acquisition, usage or modeling (e.g. lexicographers and computational linguists). Nevertheless one must admit that things have actually changed quite a bit, and some of these changes are so deep, fast and wide-ranging that it is sometimes hard to keep track of them, and be aware of the new problems and possibilities. As this evolution is in full swing, and in order to contribute to its dynamism, we organize CogALex (Cognitive Aspects of the Lexicon), whose mission is to build a bridge between the different communities. More precisely, our goal is to provide a forum for computational lexicographers, researchers in NLP, psychologists and users of lexical resources to share their knowledge and needs concerning the construction, organization and use of a lexicon by people (lexical access) and machines (NLP, IR, data mining).

History, topics of interest and focus of the current workshop

Starting at Coling 2004 (Geneva) with the workshop *Enhancing and Using Electronic Dictionaries*, there have been four follow-up events so far (CogALex I–IV), each co-located with Coling (Manchester, UK, 2008; Beijing, China, 2010; Mumbai, India, 2012, and finally, Dublin, UK, 2014). Encouraged by the enthusiasm and interest expressed by the participants of these CogALex events, we decided to organize another edition of the workshop. Like in the past, we invited researchers to address various unsolved problems. This time we put stronger emphasis though on *relations* (lexical/conceptual) and on *distributional semantics*, to explore their relevance with respect to a cognitive model of the lexicon.

The interest in distributional approaches has grown considerably over the last few years, both in computational linguistics and cognitive sciences. An additional boost has come from the recent popularity of deep learning and neural embeddings. While all these approaches seem to have great potential, their added value in addressing cognitive and semantic aspects of the lexicon still needs to be demonstrated.

We were also interested in the organization of the mental lexicon and its potential to enhance lexical resources. Given recent advances in the neurosciences, it appears timely to seek inspiration from studies concerning the human brain. There is also a lot to be learned from other fields using graphs and networks, even if their object of study is something else than language, for example biology, economy or society.

As two years ago we proposed a shared task, or rather a "friendly competition". The goal of this year's edition was the automatic identification of semantic relations from corpus data (see below for details). Also, like in the past, we were interested in the enhancement of lexical resources and electronic dictionaries, so we invited contributions from researchers involved in the building of such tools. The idea is to discuss modifications of existing resources by taking the users' needs and knowledge states into account. Given the diversity of our goals we solicited papers including, but not limited to the following topics, each of which can be considered from various viewpoints: *linguistics* (theoretical or practical), *neuro-* or *psycholinguistics* (tip-of-the-tongue problem, associations), *network related sciences* (sociology, economy, biology), *mathematics* (vector-based approaches, graph theory, small-world problem), etc.

Organization, i.e. structure of the lexicon

- Micro- and macrostructure of the lexicon;
- Indexical categories (taxonomies, thesaurus-like topical structures, etc.);
- Distribution of information and relations between words.

The meaning of words and techniques for revealing it

- Lexical representation (holistic, decomposed);
- Meaning representation (concept based, primitives);
- Distributional semantics (count models, neural embeddings, etc.).

Analysis of the input given by a dictionary user

- What information do language producers provide when looking for a word (terms, relations)?
- What kind of relational information do they give: typed or untyped relations?
- Which relations are typically used?

Methods for crafting dictionaries or additional functions like indexes

- Manual, automatic or collaborative building of dictionaries (crowd-sourcing, serious games, etc.);
- (Semi-)automatic induction of the link type (e.g. synonym, hypernym, association, etc.);
- Extraction of associations from corpora to build semantic networks supporting navigation.

Dictionary access (navigation and search strategies)

- Search based on sound (rhymes), meaning or functionally related words (associations);
- Determination of appropriate search space based on user's knowledge, etc.;
- Identification of typical word access strategies (navigational patterns) used by people.

Shared task on the corpus-based identification of semantic relations

The shared task was organized by Stefan Evert, Alessandro Lenci and Enrico Santus, with the precious help of Anna Gladkova. Its goal was the automatic identification of semantic relations from corpus data, which has great potential: it promises an efficient and scalable solution to NLP tasks, while (possibly) providing a cognitively plausible model for human acquisition and usage of such relations.

Semantic relations play a central role in human lexical retrieval (navigation) and the organization of words in the lexicon, the resource within which search takes place. Hence learning about them may shed some light on the mental lexicon and the knowledge people have when searching for a word. Discovering whether words are semantically related and how so (which kind of relation holds between them?) is also an important task in natural language processing (NLP) by and large, with a wide range of applications, such as automatic thesaurus creation, natural language generation (automatic creation of outlines), ontology learning, paraphrase generation, etc.

The aim of this "friendly competition" was not so much to find the team with the best-performing system, as to test different distributional models and other corpus-based approaches on a challenging semantic task, in order to gain a better understanding of their respective strengths and weaknesses.

The task was split into two subtasks:

1. Given a pair of words (e.g. *dog – fruit*), decide whether they are semantically related or not.
2. For each word pair (e.g. *cat – animal*), decide which of the following semantic relations holds between them: synonymy, antonymy, hypernymy, meronymy, or none (random combination).

The organizers provided a data set based on WordNet and ConceptNet, which was then cleaned by native speakers in a CrowdFlower task. The remaining word pairs were split into a training set and test set, and evaluation was carried out in terms of precision, recall and their harmonic mean (F_1). In subtask 2, the overall score was determined as the weighted average over all four semantic relations.

Outcome of the call and a word of thanks

We received 30 submissions, of which seven were accepted as full papers with oral presentation, eight as poster presentations, and seven as shared task papers.

We would like to take this opportunity to express our sincerest thanks to all the members of the Programme Committee. Their expertise was invaluable to ensure a good selection of papers despite the tight schedule. Their reviews were helpful not only for us to make the decisions, but also for the authors, helping them to improve their work.

Last, but not least, we would like to thank Chris Biemann for having accepted to be our invited speaker. His talk – *Vectors or Graphs? On Differences of Representations for Distributional Semantic Models* – fits perfectly well in this workshop.

We hope that the work presented here will inspire you, generate fruitful discussions, and possibly lead to new ideas, insights and collaborations.

The CogALex-V Organizers
Michael Zock, Alessandro Lenci, Stefan Evert

Organisers

Michael Zock (LIF, CNRS, Aix-Marseille University, Marseille, France)

Alessandro Lenci (Computational Linguistics Laboratory, University of Pisa, Pisa, Italy)

Stefan Evert (Friedrich-Alexander-Universität Erlangen-Nürnberg, Germany)

Programme Committee

Biemann, Chris (Universität Hamburg, Germany)
Babych, Bogdan (University of Leeds, UK)
Brysbaert, Marc (Experimental Psychology, Ghent University, Belgium)
Cristea, Dan (“Al. I. Cuza” University, Iasi, Romania)
De Deyne, Simon (University of Adelaide, Australia)
de Melo, Gerard (IIIS, Tsinghua University, Beijing, China)
Evert, Stefan (FAU Erlangen-Nürnberg, Germany)
Ferret, Olivier (CEA LIST, France)
Fontenelle, Thierry (CDT, Luxemburg)
Grefenstette, Gregory (Inria, Paris, France)
Hirst, Graeme (University of Toronto, Canada)
Hovy, Eduard (CMU, Pittsburgh, USA)
Hsieh, Shu-Kai (National Taiwan University, Taipei, Taiwan)
Lafourcade, Matthieu (LIRMM, Université de Montepellier, France)
Lapalme, Guy (RALI, University of Montreal, Canada)
Lebani, Gianluca (University of Pisa, Italy)
Lenci, Alessandro (University of Pisa, Italy)
L’Homme, Marie Claude (University of Montreal, Canada)
Mititelu, Verginica (RACAI, Bucharest, Romania)
Paradis, Carita (Centre for Languages and Literature Lund University, Sweden)
Pihlevar, Taher (University of Cambridge, UK)
Pirrelli, Vito (ILC, Pisa, Italy)
Polguère, Alain (ATILF-CNRS, Nancy, France)
Purver, Matthew (King’s College, London, UK)
Ramisch, Carlos (AMU, Marseille, France)
Rayson, Paul (UCREL, University of Lancaster, UK
Rosso, Paol (NLEL, Universitat Politècnica de València, Spain)
Sahlgren, Magnus (Gavagai Inc. & SICS, Sweden)
Schulte im Walde, Sabine (University of Stuttgart, Germany)
Schwab, Didier (LIG, Grenoble, France)
Sharoff, Serge (University of Leeds, UK)
Stella, Massimo (Institute for Complex Systems Simulation, University of Southhampton, UK)
Tokunaga, Takenobu (TITECH, Tokyo, Japan)
Tufiş, Dan (RACAI, Bucharest, Romania)
Zarcone, Alessandra (Saarland University, Saarbrücken, Germany)
Zock, Michael (LIF-CNRS, Marseille, France)

Table of Contents

Workshop Program

09:00–09:05 ***Introduction***

09:05–10:00 **Invited talk**

09:05–10:00 *Vectors or Graphs? On Differences of Representations for Distributional Semantic Models*
Chris Biemann

10:00–11:20 **Papers I: Distributional semantics**

10:00–10:30 *"Beware the Jabberwock, dear reader!" Testing the distributional reality of construction semantics*
Gianluca Lebani and Alessandro Lenci

10:30–10:50 ***coffee break***

10:50–11:20 *Regular polysemy: from sense vectors to sense patterns*
Anastasiya Lopukhina and Konstantin Lopukhin

11:20–12:20 **Papers II: Semantic relations**

11:20–11:50 *Path-based vs. Distributional Information in Recognizing Lexical Semantic Relations*
Vered Shwartz and Ido Dagan

11:50–12:20 *Semantic Relation Classification: Task Formalisation and Refinement*
Vivian Santos, Manuela Huerliman, Brian Davis, Siegfried Handschuh and André Freitas

12:20–14:00 ***lunch break***

14:00–15:30 Papers III: Cognitive aspects of language

14:00–14:30 *The Power of Language Music: Arabic Lemmatization through Patterns*
Mohammed Attia, Ayah Zirikly and Mona Diab

14:30–15:00 *Word Sense Disambiguation using a Bidirectional LSTM*
Mikael Kågebäck and Hans Salomonsson

15:00–15:30 *Towards a resource based on users' knowledge to overcome the Tip of the Tongue problem.*
Michael Zock and Chris Biemann

15:30–16:00 *coffee break*

16:00–16:40 Shared task

16:00–16:20 *The CogALex-V Shared Task on the Corpus-Based Identification of Semantic Relations*
Enrico Santus, Anna Gladkova, Stefan Evert and Alessandro Lenci

16:20–16:40 *CogALex-V Shared Task: LexNET - Integrated Path-based and Distributional Method for the Identification of Semantic Relations*
Vered Shwartz and Ido Dagan

16:40–16:50 Shared task poster boosters

CogALex-V Shared Task: GHHH - Detecting Semantic Relations via Word Embeddings
Mohammed Attia, Suraj Maharjan, Younes Samih, Laura Kallmeyer and Thamar Solorio

CogALex-V Shared Task: Mach5 – A traditional DSM approach to semantic relatedness
Stefan Evert

CogALex-V Shared Task: ROOT18
Emmanuele Chersoni, Giulia Rambelli and Enrico Santus

CogALex-V Shared Task: CGSRC - Classifying Semantic Relations using Convolutional Neural Networks
Chinnappa Guggilla

CogALex-V Shared Task: LOPE
Kanan Luce, Jiaxing Yu and Shu-Kai HSIEH

CogALex-V Shared Task: HsH-Supervised – Supervised similarity learning using entry wise product of context vectors
Christian Wartena and Rosa Tsegaye Aga

16:50–17:15 **Poster boosters**

17:15–17:50 **Poster session**

A Study of the Bump Alternation in Japanese from the Perspective of Extended/Onset Causation
Natsuno Aoki and Kentaro Nakatani

GhoSt-PV: A Representative Gold Standard of German Particle Verbs
Stefan Bott, Nana Khvtisavrishvili, Max Kisselew and Sabine Schulte im Walde

Discovering Potential Terminological Relationships from Twitter's Timed Content
Mohammad Daoud and Daoud Daoud

Lexfom: a lexical functions ontology model
Alexsandro Fonseca, Fatiha Sadat and François Lareau

A Proposal for combining "general" and specialized frames
Marie-Claude L' Homme, Carlos Subirats and Benoît Robichaud

Antonymy and Canonicity: Experimental and Distributional Evidence
Andreana Pastena and Alessandro Lenci

Categorization of Semantic Roles for Dictionary Definitions
Vivian Silva, Siegfried Handschuh and André Freitas

Corpus and dictionary development for classifiers/quantifiers towards a French-Japanese machine translation
Mutsuko Tomokiyo and Christian Boitet

17:50–18:00 ***Wrap-up and sayonara***