Coling 2010

23rd International Conference on Computational Linguistics

Proceedings of the 2nd Workshop on
Cognitive Aspects of the Lexicon

Workshop chairs:
Michael Zock and Reinhard Rapp

22 August 2010
Beijing International Convention Center
Beijing, China

Produced by
Chinese Information Processing Society of China

To order the CD of Coling 2010 and its Workshop Proceedings, please contact:

Chinese Information Processing Society of China
No.4, Southern Fourth Street
Haidian District, Beijing, 100190
China
Tel: +86-010-62562916
Fax: +86-010-62562916
cips@iscas.ac.cn

Introduction

Whenever we read a book, write a letter or launch a query on a search engine, we always use words, the shorthand labels and concrete forms of abstract notions (concepts, ideas and more or less well specified thoughts). Yet, words are not only vehicles to express 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, refining, expanding or illustrating our possibly underspecified thoughts. Only words have these unique capabilities, which is why they are so important.

Obviously, a good dictionary should contain many entries and a lot of information associated with each one of them. Yet, the quality of a dictionary depends not only on coverage, but also on accessibility of information. Access strategies vary with the task (text understanding vs. text production) and the knowledge available at the moment of consultation (words, concepts, speech sounds). Unlike readers who look for meanings, writers start from them, searching for the corresponding words. While paper dictionaries are static, permitting only limited strategies for accessing information, their electronic counterparts promise dynamic, proactive search via multiple criteria (meaning, sound, related words) and via diverse access routes. Navigation takes place in a huge conceptual lexical space, and the results are displayable in a multitude of forms (e.g. as trees, as lists, as graphs, or sorted alphabetically, by topic, by frequency).

Many lexicographers work nowadays with huge digital corpora, using language technology to build and to maintain the lexicon. But access to the potential wealth of information in dictionaries remains limited for the common user. Yet, the new possibilities of electronic media in terms of comfort, speed and flexibility (multiple inputs, polyform outputs) are enormous. Computational resources are not prone to the same limitations as paperbound dictionaries. The latter were limited in scope, being confined to a specific task (translation, synonyms, ...) for economical reasons, but this limitation is not justified anymore.

Today, by exploiting the advantages of the digital form, we can perform all tasks via one single resource, which may comprise a dictionary, a thesaurus and even more. The goal of this second CogALex workshop, which follows the first edition at COLING 2008 in Manchester, is to perform the groundwork for the next generation of electronic dictionaries, that is, to study the possibility of integrating the different resources, as well as to explore the feasibility of taking the users needs, knowledge and access strategies into account. To reach this goal, we have invited researchers from fields such as computational lexicography, psycholinguistics, cognitive psychology, language learning and ergonomics to address one or several of the following topics:

1. *Conceptual input* of a dictionary user. What is in the authors' minds when they are generating a message and looking for a word? Do they start from partial definitions, i.e. underspecified input (bag of words), conceptual primitives, semantically related words, something akin to synsets, or something completely different? What does it take to bridge the gap between this input, incomplete as it may be, and the desired output (target word)?
2. *Organizing the lexicon and indexing words*. Concepts, words and multi-word expressions can be organized and indexed in many ways, depending on the task and language type. For example, in Indo-European languages words are traditionally organized in alphabetical order, whereas in

Chinese they are organized by semantic radicals and stroke counts. The way words and multi-word expressions are stored and organized affects indexing and access. Since knowledge states (i.e. knowledge available when initiating search) vary greatly and in unpredictable ways, indexing must allow for multiple ways of navigation and access. Hence the question: what organizational principles allow the greatest flexibility for access?

3. *Access, navigation and search strategies* based on various entry types (modalities) and knowledge states. Words are composed of meanings, forms and sounds. Hence, access should be possible via any of these components: via meanings (bag of words), via forms, simple or compound ('hot, dog' vs. 'hot-dog'), and via sounds (syllables). Access should be possible even if input is given in an incomplete, imprecise or degraded form. Furthermore, to allow for natural and efficient access, we need to take the users' knowledge into account (search space reduction) and provide adequate navigational tools, metaphorically speaking, a map and a compass. How do existing tools address these needs, and what could be done to go further?

4. *NLP applications*: Contributors can also demonstrate how such enhanced dictionaries, once embedded in existing NLP applications, can boost performance and help to solve lexical and textual-entailment problems, such as those evaluated in SEMEVAL 2007, or, more generally, generation problems encountered in the context of summarization, question-answering, interactive paraphrasing or translation.

Quite a few of these issues are dealt with in the papers we received. The accepted papers present a rich selection of ideas on the crossroads of semantics, cognition, lexicography, and language learning, thereby emphasizing the interdisciplinary character of the workshop. These are the topics: generating semantic networks, encoding commonsense knowledge in WordNet, textual entailment, sentiment analysis, corpus-based extraction of conceptual classes, parsing of thesauri, term extraction, determining noun classifiers, requirements when using the dictionary of an authoring tool, and the problem of word access.

In sum, there is an active community of researchers working on cognitive aspects of the lexicon, and there is a real awareness concerning the importance of the problems presented in our call for papers.

We would like to thank all the people who in one way or another have helped us to make this workshop a success. Our special thanks go to Eduard Hovy for having accepted to give the invited presentation, and to the members of the program committee who did an excellent job in reviewing the submitted papers. Their reviews were important not only to assure a good selection of papers, but also for the authors, helping them to improve their work. We would also like to express our gratitude to the COLING organizers, in particular to the general workshop chairs and the publication chairs. Last but not least, we would like to thank our authors for their papers and presentations and the participants of the workshop for their interest and their contributions to the discussions.

Michael Zock and Reinhard Rapp

Organizers:

Michael Zock, LIF-CNRS, Marseille (France)
Reinhard Rapp, University of Tarragona (Spain)

Invited Speaker:

Eduard Hovy, Information Sciences Institute, University of Southern California (USA)

Program Committee:

Slaven Bilac, Google Tokyo (Japan)
Pierrette Bouillon, ISSCO, Geneva (Switzerland)
Dan Cristea, University of Iasi (Romania)
Katrin Erk, University of Texas (USA)
Olivier Ferret, CEA LIST (France)
Thierry Fontenelle, EU Translation Centre (Luxembourg)
Sylviane Granger Université Catholique de Louvain (Belgium)
Gregory Grefenstette, Exalead, Paris (France)
Ulrich Heid, IMS, University of Stuttgart (Germany)
Erhard Hinrichs, University of Tübingen (Germany)
Graeme Hirst, University of Toronto (Canada)
Eduard Hovy, ISI, University of Southern California, Los Angeles (USA)
Chu-Ren Huang, Hong Kong Polytechnic University (China)
Terry Joyce, Tama University, Kanagawa-ken (Japan)
Philippe Langlais, DIRO/RALI University of Montreal (Canada)
Marie-Claude L'Homme, University of Montreal (Canada)
Verginica Mititelu, RACAI, Bucharest (Romania)
Alain Polguère, ATILF - CNRS / Université Nancy 2 (France)
Reinhard Rapp, University of Tarragona (Spain)
Sabine Schulte im Walde, University of Stuttgart (Germany)
Gilles Sérasset, IMAG, Grenoble (France)
Serge Sharoff, University of Leeds (UK)
Anna Sinopalnikova, FIT, BUT, Brno (Czech Republic)
Carole Tiberius, Institute for Dutch Lexicology (The Netherlands)
Takenobu Tokunaga, TITECH, Tokyo (Japan)
Dan Tufis, RACAI, Bucharest (Romania)
Piek Vossen, Vrije Universiteit Amsterdam (The Netherlands)
Yorick Wilks, Oxford Internet Institute (UK)
Michael Zock, LIF-CNRS, Marseille (France)
Pierre Zweigenbaum, LIMSI-CNRS, Orsay (France)

Table of Contents

Distributional Semantics and the Lexicon
Eduard Hovy 1

SemanticNet-Perception of Human Pragmatics
Amitava Das and Sivaji Bandyopadhyay 2

Exploiting Lexical Resources for Therapeutic Purposes: the Case of WordNet and STaRS.sys
Gianluca E. Lebani and Emanuele Pianta 12

Textual Entailmaint Recognition using Word Overlap, Mutual Information and Subpath Set
Yuki Muramatsu, Kunihiro Uduka and Kazuhide Yamamoto 18

The Color of Emotions in Texts
Carlo Strapparava and Gozde Ozbal 28

How to Expand Dictionaries by Web-Mining Techniques
Nicolas Béchet and Mathieu Roche 33

An Optimal and Portable Parsing Method for Romanian, French, and German Large Dictionaries
Neculai Curteanu, Alex Moruz and Diana Trandabat 38

Conceptual Structure of Automatically Extracted Multi-Word Terms from Domain Specific Corpora: a Case Study for Italian
Elisa Lavagnino and Jungyeul Park 48

Computational Lexicography: A Feature-based Approach in Designing an E-dictionary of Chinese Classifiers
Helena Gao 56

In Search of the 'Right' Word
Stella Markantonatou, Aggeliki Fotopoulou, Maria Alexopoulou and Marianna Mini 66

Lexical Access, a Search-Problem
Michael Zock, Didier Schwab and Nirina Rakotonanahary 75

Conference Program

Sunday, August 22, 2010

9:00–9:15 Opening Remarks

Invited Keynote Presentation

9:15–10:30 *Distributional Semantics and the Lexicon*
Eduard Hovy

10:30–11:00 Coffee break

Session 1: Semantics and Cognition

11:00–11:30 *SemanticNet-Perception of Human Pragmatics*
Amitava Das and Sivaji Bandyopadhyay

11:30–12:00 *Exploiting Lexical Resources for Therapeutic Purposes: the Case of WordNet and STaRS.sys*
Gianluca E. Lebani and Emanuele Pianta

12:00–12:30 *Textual Entailmaint Recognition using Word Overlap, Mutual Information and Sub-path Set*
Yuki Muramatsu, Kunihiro Uduka and Kazuhide Yamamoto

12:30–13:00 *The Color of Emotions in Texts*
Carlo Strapparava and Gozde Ozbal

13:00–14:00 Lunch break

Sunday, August 22, 2010 (continued)

Session 2: Lexicography

14:00–14:30 *How to Expand Dictionaries by Web-Mining Techniques*
Nicolas Béchet and Mathieu Roche

14:30–15:00 *An Optimal and Portable Parsing Method for Romanian, French, and German Large Dictionaries*
Neculai Curteanu, Alex Moruz and Diana Trandabat

15:00–15:30 *Conceptual Structure of Automatically Extracted Multi-Word Terms from Domain Specific Corpora: a Case Study for Italian*
Elisa Lavagnino and Jungyeul Park

15:30–16:00 Coffee break

Session 3: Word Access and Language Learning

16:00–16:30 *Computational Lexicography: A Feature-based Approach in Designing an E-dictionary of Chinese Classifiers*
Helena Gao

16:30–17:00 *In Search of the 'Right' Word*
Stella Markantonatou, Aggeliki Fotopoulou, Maria Alexopoulou and Marianna Mini

Keynote Presentation

17:00–17:45 *Lexical Access, a Search-Problem*
Michael Zock, Didier Schwab and Nirina Rakotonanahary

17:45–18:00 Wrap Up Discussion

18:00 End of the Workshop