Multilingual Word Sense Disambiguation and Entity Linking

Roberto Navigli and **Andrea Moro**
Dipartimento di Informatica, Sapienza Università di Roma
Viale Regina Elena 295, 00161 Roma, Italy
{moro, navigli}@di.uniroma1.it

Tutorial Motivation and Description

Nowadays the textual information available online is provided in an increasingly wide range of languages. This language explosion clearly forces researchers to focus on the challenging problem of being able to analyze and understand text written in any language. At the core of this problem lies the lexical ambiguity of language, an issue which is addressed by two key tasks in computational lexical semantics: multilingual Word Sense Disambiguation (WSD) and Entity Linking (EL).

WSD (Navigli, 2009) is a historical task in Computational Linguistics aimed at explicitly assigning meanings to word occurrences within text, while EL (Erbs et al., 2011; Rao et al., 2013; Cornolti et al., 2013) is a recent task focused on finding mentions of entities within a text and linking them to the most suitable entry in a knowledge base, if one exists. The two main differences between WSD and EL are in the kind of inventory used, i.e. dictionary vs. encyclopedia, and the assumption that the mention is complete or potentially incomplete, respectively. Notwithstanding these differences, the tasks are pretty similar in nature, in that they both involve the disambiguation of textual fragments in a given language according to a reference inventory. Nevertheless, the research community has tackled the two tasks separately, often duplicating efforts and solutions. Moreover, the vast majority of the state-of-the-art approaches only marginally take into account languages different from English.

In this tutorial, we present the two tasks of multilingual WSD and EL, by surveying the challenges involved and the most effective solutions, both in isolation by illustrating the state of the art in the two fields, and when the tasks are addressed in a unified, multilingual way.

In particular, this tutorial covers three key aspects of the multilingual WSD and EL tasks:

1. Multilingual inventories of word senses and named entities;
2. State-of-the-art methods to perform disambiguation and linking;
3. Evaluation of the systems: gold standard datasets and performance measures.

The tutorial is aimed at stressing the key challenges of the tasks of WSD and EL when moving from a monolingual to a multilingual setup. The tutorial includes examples and discussions intended to illustrate and clarify the major challenges of the tasks and which solutions are most appropriate to which problem.

Organization of the tutorial

The half-day tutorial contains sessions on the following topics:

1. Introduction (30 mins) This first session will provide the necessary background, definitions and examples for the two considered tasks: multilingual WSD and EL.
2. The multilingual inventory for word senses and named entities (45 mins) In this session we will overview the definitions of the inventories used in state-of-the-art approaches both for WSD and

Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Tutorial Abstracts, pages 5–7, Dublin, Ireland, August 23-29 2014.

EL. We will then discuss the key aspects of partial matching for EL and, finally, we will describe multilingual inventories of word senses and named entities, among which Open Multilingual WordNet (Bond and Foster, 2013), Wikipedia[1], DBpedia (Auer et al., 2007), BabelNet (Navigli and Ponzetto, 2012).

3. State of the art in WSD and EL (75 mins) This session will introduce the key challenges to the tasks of WSD and EL and the well-known approaches, such as IMS (Zhong and Ng, 2010) and UKB (Agirre et al., 2013) for WSD, and Babelfy (Moro et al., 2014), AIDA (Hoffart et al., 2011; Hoffart et al., 2012), Tagme (Ferragina and Scaiella, 2010; Ferragina and Scaiella, 2012), Illinois Wikifier (Cheng and Roth, 2013) and DBpedia Spotlight (Mendes et al., 2011; Daiber et al., 2013), that can partially address them. Challenges include: the lack of training data for non-English languages, the granularity of the sense inventory, the high level of ambiguity in EL, the most frequent sense baseline challenge, etc.

4. Evaluation measures and gold standard datasets (30 mins) We will conclude the tutorial by describing the standard performance measures for these tasks and well-known datasets for multilingual WSD and EL together with a discussion of the results.

Speakers

Roberto Navigli is an associate professor in the Department of Computer Science at the Sapienza University of Rome. He is the recipient of an ERC Starting Grant in computer science and informatics on multilingual word sense disambiguation (2011-2016) and a co-PI of a Google Focused Research Award on Natural Language Understanding. His research interests lie in the field of Word Sense Disambiguation and Induction, multilingual knowledge acquisition and applications of lexical semantics.
Andrea Moro is a PhD student in the Department of Computer Science at the Sapienza University of Rome. His research interests focus on Natural Language Understanding with an emphasis on Unsupervised Relation Extraction, Word Sense Disambiguation and Entity Linking.

Acknowledgments

The authors gratefully acknowledge the support of the ERC Starting Grant MultiJEDI No. 259234.

References

Eneko Agirre, Oier Lopez de Lacalle, and Aitor Soroa. 2013. Random Walks for Knowledge-Based Word Sense Disambiguation. *Computational Linguistics*.

Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. 2007. DBpedia: A Nucleus for a Web of Open Data. In *Proc. of ISWC/ASWC*, pages 722–735.

Francis Bond and Ryan Foster. 2013. Linking and extending an open multilingual wordnet. In *Proc. of ACL*, pages 1352–1362.

Xiao Cheng and Dan Roth. 2013. Relational Inference for Wikification. In *Proc. of EMNLP*.

Marco Cornolti, Paolo Ferragina, and Massimiliano Ciaramita. 2013. A framework for benchmarking entity-annotation systems. In *Proc. of WWW*, pages 249–260.

Joachim Daiber, Max Jakob, Chris Hokamp, and Pablo N. Mendes. 2013. Improving Efficiency and Accuracy in Multilingual Entity Extraction. In *Proc. of I-Semantics*.

Nicolai Erbs, Torsten Zesch, and Iryna Gurevych. 2011. Link Discovery: A Comprehensive Analysis. In *Proc. of ICSC*, pages 83–86.

[1] http://wikipedia.org

Paolo Ferragina and Ugo Scaiella. 2010. TAGME: on-the-fly annotation of short text fragments (by wikipedia entities). In *Proc. of CIKM*, pages 1625–1628.

Paolo Ferragina and Ugo Scaiella. 2012. Fast and accurate annotation of short texts with wikipedia pages. *Software, IEEE*, 29(1):70–75, Jan.

Johannes Hoffart, Mohamed Amir Yosef, Ilaria Bordino, Hagen Fürstenau, Manfred Pinkal, Marc Spaniol, Bilyana Taneva, Stefan Thater, and Gerhard Weikum. 2011. Robust Disambiguation of Named Entities in Text. In *Proc. of EMNLP*, pages 782–792.

Johannes Hoffart, Stephan Seufert, Dat Ba Nguyen, Martin Theobald, and Gerhard Weikum. 2012. KORE: keyphrase overlap relatedness for entity disambiguation. In *Proc. of CIKM*, pages 545–554.

Pablo N. Mendes, Max Jakob, Andrés García-Silva, and Christian Bizer. 2011. DBpedia spotlight: shedding light on the web of documents. In *Proc. of I-Semantics*, pages 1–8.

Andrea Moro, Alessandro Raganato, and Roberto Navigli. 2014. Entity Linking meets Word Sense Disambiguation: A Unified Approach. *Transactions of the 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. *Artificial Intelligence*, 193:217–250.

Roberto Navigli. 2009. Word sense disambiguation: A survey. *ACM Comput. Surv.*, 41(2):1–69.

Delip Rao, Paul McNamee, and Mark Dredze. 2013. Entity linking: Finding extracted entities in a knowledge base. In *Multi-source, Multilingual Information Extraction and Summarization*, pages 93–115. Springer.

Zhi Zhong and Hwee Tou Ng. 2010. It Makes Sense: A Wide-Coverage Word Sense Disambiguation System for Free Text. In *Proc. of the ACL 2010 System Demonstrations*, pages 78–83, Uppsala, Sweden, July. Association for Computational Linguistics.