KNOWLEDGE ENGINEERING: PRINCIPLES AND TECHNIQUES
Proceedings of the International Conference on Knowledge Engineering,
Principles and Techniques, KEPT2007
Cluj-Napoca (Romania), June 6–8, 2007, pp. 12–13

USING WIKIPEDIA FOR AUTOMATIC WORD SENSE DISAMBIGUATION

RADA MIHALCEA

Abstract

Ambiguity is inherent to human language. In particular, word sense ambiguity is prevalent in all natural languages, with a large number of the words in any given language carrying more than one meaning.

In this talk, I will present a new approach for building sense tagged corpora using Wikipedia as a source of sense annotations. Starting with the hyperlinks available in Wikipedia, I will show how one can generate sense annotated corpora that can be used for building accurate and robust sense classifiers. Through word sense disambiguation experiments performed on the Wikipedia-based sense tagged corpus generated for a subset of the Senseval ambiguous words, I will show that the Wikipedia annotations are reliable, and the quality of a sense tagging classifier built on this data set exceeds by a large margin the accuracy of an informed baseline that selects the most frequent word sense by default.

Biography

Rada Mihalcea is an Assistant Professor of Computer Science at the University of North Texas. Her research interests are in lexical semantics, graph-based algorithms for natural language processing, minimally supervised natural language learning, and multilingual natural language processing. She is currently involved in a number of research projects, including knowledge-based word sense disambiguation, (non-traditional) methods for building annotated corpora with volunteer contributions over the Web, graph-based algorithms for text processing, opinion and sentiment analysis, and computational humour. She has published a large number of articles in books, journals, and proceedings, in these and related areas. She is the president of the ACL Special Group on the Lexicon (SIGLEX), and a board member for the ACL Special Group on Natural Language Learning (SIGNLL). She serves on the editorial boards of the journal of Computational Linguistics, the journal of Language Resources and Evaluations, the Journal of Natural Language

2000 *Mathematics Subject Classification.* 68T35, 68T50, 91F20.

Engineering, the Journal of Research on Language and Computation, and the recently established journal of Interesting Negative Results in Natural Language Processing and Machine Learning.