(Invited Talk) The Web as an Implicit Training Set: Application to Noun Compounds Syntax and Semantics

Preslav Nakov
Qatar Computing Research Institute
pnakov@qf.org.qa

Abstract

The 60-year-old dream of computational linguistics is to make computers capable of communicating with humans in natural language. This has proven hard, and thus research has focused on sub-problems. Even so, the field was stuck with manual rules until the early 90s, when computers became powerful enough to enable the rise of statistical approaches. Eventually, this shifted the main research attention to machine learning from text corpora, thus triggering a revolution in the field.

Today, the Web is the biggest available corpus, providing access to quadrillions of words; and, in corpus-based natural language processing, size does matter. Unfortunately, while there has been substantial research on the Web as a corpus, it has typically been restricted to using page hit counts as an estimate for n-gram word frequencies; this has led some researchers to conclude that the Web should be only used as a baseline.

In this talk, I will reveal some of the hidden potential of the Web that lies beyond the n-gram, with focus on the syntax and semantics of English noun compounds. First, I will present a highly accurate lightly supervised approach based on surface markers and linguistically-motivated paraphrases that yields state-of-the-art results for noun compound bracketing: e.g., "[[liver cell] antibody]" is left-bracketed, while "[liver [cell line]]" is right-bracketed. Second, I will present a simple unsupervised method for mining implicit predicates that can characterize the semantic relations holding between the nouns in noun compounds, e.g., "malaria mosquito" is a "mosquito that carries/spreads/causes/transmits/brings/infects with/... malaria". Finally, I will show how these ideas can be used to improve statistical machine translation.

About the Speaker:
Preslav Nakov is a Senior Scientist at the Qatar Computing Research Institute (QCRI). He received his Ph.D. in Computer Science from the University of California at Berkeley in 2007 (supported by a Fulbright grant and a UC Berkeley fellowship). Before joining QCRI, Preslav was a Research Fellow at the National University of Singapore. He has also spent a few months at the Bulgarian Academy of Sciences and the Sofia University, where he was an honorary lecturer. Preslav's research interests include lexical semantics (in particular, multi-word expressions, noun compounds syntax and semantics, and semantic relation extraction), machine translation, Web as a corpus, and biomedical text processing.

Preslav was involved in many activities related to lexical semantics. He is a member of the SIGLEX board, he is co-chairing SemEval'2014, SemEval'2015, and SemEval'2016, and he has co-organized several SemEval tasks, e.g., on the semantics of noun compounds, on semantic relation extraction, on sentiment analysis on Twitter, and on community question answering. He has co-chaired MWE in 2009 and 2010, as well as other semantics workshops such as RELMS, and he was an area chair of *SEM'2013. He was also a guest co-editor for the 2013 special issue of the journal of Natural Language Engineering on the syntax and semantics of noun compounds, and he is currently a guest co-editor of a special issue of LRE on SemEval-2014 and Beyond. In 2013, he has published a Morgan & Claypool book on semantic relation extraction; he has given a tutorial on the same topic at RANLP'2013, and he is giving a similar one at EMNLP'2015.

Proceedings of the ACL 2015 Workshop on Novel Computational Approaches to Keyphrase Extraction, page 18,
Beijing, China, July 30, 2015.