DGfS-CL

Sourjikova, Eva & Frank, Anette
Computational Linguistics Department, Heidelberg University

Identification of Domain-Specific Named Entity Classes

The objective of our research is to automatically detect fine-grained named entity classes in any domain. Given a text collection (e.g. texts about astronomy), our system identifies an inventory of semantic classes (e.g. astronomer, asteroid, planet, star, constellation etc.) of named entities that are characteristic in this domain (e.g. “Edwin Hubble”, “Gaspra”, “Venus”, “Zeta Leporis”, “Pegasus”, etc.). The inventory of named entity classes we detect for a domain using a representative domain text collection implicitly represents a concept structure of the domain. Inducing such a structure would be beneficial for many semantic text processing tasks such as fine-grained named entity classification, text categorization, ontology induction etc. Research in named entity classification typically involves a small, predefined set of broad classes (e.g. person, location, organization) that are present in any domain. In contrast, in our work we focus on automatic corpus-based induction of fine-grained named entity classes in order to capture the diversity of named entity classes across different domains.

Our corpus-based approach is theoretically grounded on the referential and formal properties of proper names: proper names single out important entities in a domain and these entities are often introduced in the texts jointly with their semantic category. We use linguistically motivated part-of-speech patterns to automatically extract these class-instance pairs from the corpus and then employ their cooccurrence-based features as well as ontological features induced from WordNet to select domain-specific named entity classes from the overall set of acquired classes. Our classification model distinguishes named-entity classes (e.g. “scientist”) and classes of entities that are not name bearing (e.g. "summer", "curtain", "laughter") and thus reliably eliminates incorrect extractions. Experimental results using several domain-specific corpora show that our method is able to identify named entity classes with high accuracy in diverse domains of different granularity.

Baroni, M., Bernardini, S., Ferraresi, A. and Zanchetta, E.: The wacky wide web: A collection of very large linguistically processed web-crawled corpora. 2009

Evert, S.: The Statistics of Word Cooccurrences: Word Pairs and Collocations. 2004, Stuttgart

Fellbaum, C.: Wordnet: An Electronic Lexical Database. 1998, Cambridge

Hearst, M. A.: Automatic Acquisition of Hyponyms from Large Text Corpora.1992

Langendonck, W. V.: Theory and Typology of proper names. 2007, Berlin

Snow, R., Jurafsky, D. And Ng, A.Y.: Learning syntactic patterns for automatic hypernym discovery. 2004, Vancouver