Enhancing Roget's Thesaurus with Semantic Tags[1]

Uta Priss, L. John Old

(Edinburgh Napier University)

In the 1950s Margaret Masterman was a pioneer of natural language processing (NLP) and artificial intelligence (AI). Her research is mainly published in technical reports and was to some degree forgotten until Wilks published an edition of her work (Masterman, 2005). One of her ideas was to use Roget's Thesaurus (RT) as a tool for representing what she called "semantic transformations" (and what would be called "conceptual structures" in modern terminology) for the purposes of machine translation (or "mechanical translation" in the terminology of her time). Recently, the relevance of her thesaurus research has been investigated with respect to modern applications and found to be highly relevant from a modern viewpoint (Priss & Old, 2009).

In this talk, we are revisiting Masterman's idea of attaching "semantic tags" to the entries of RT. In our usage "semantic tags" refers predominantly to frames, but includes the tags that already exist in RT. Although RT contains a similar coverage of English words to WordNet, its classification structure is predominantly suited for human usage, and not NLP. This is because many linguistic details are implied, not explicit. Currently the only systematically applied tags in RT indicate the part of speech of a word and a division of words into antonymic categories (which is already less consistently represented). RT further contains a long list of connotational tags, such as "colloquial, philosophical, slang, dialectic, jocular, etc", which, however, only apply to a small subset of the words. In order for RT to become a more general-purpose NLP resource, similar in style to WordNet (which is a long-term goal of our research), the tags need to be more systematically and consistently applied.

Modern search engines, such as Google, often fail to accurately retrieve documents for prepositional queries. For example, a search for "food Provence" or for "definition semantics" will retrieve documents relating to "food of the Provence" and "definition of semantics", respectively. But it is difficult to retrieve an "example of componential semantics" if there is no document which contains that exact phrase because a non-phrasal search for "componential semantics" and "example" will retrieve documents that contain the words "componential semantics" and "for example". These kinds of queries need NLP techniques for grammatical analysis, but also dictionaries that contain basic semantic tags or frames for each word. We believe that a semantically tagged electronic dictionary is an essential tool for improving the NLP capabilities of search engines.

[1]This paper was presented at the Second Conference on Concept Types and Frames in Language, Cognition and Science, Düsseldorf, Germany, August 2009.

The idea for this research is to investigate automatic means for adding semantic tags to RT. These means are based on the use of Formal Concept Analysis (FCA) lattices and subcategorisation frames. This approach was first suggested by Basili (1997) and subsequently used in similar form by Cimiano (2003) for ontology construction, by Priss (2005) for semantic classification and by Stepanova (2009) for the acquisition of lexico-semantic knowledge from corpora. Basili's idea is to identify verb subcategorisation frames by classifying verbs based on their argument structures as derived from a corpus. Basili then constructs FCA concept lattices (Ganter & Wille, 1999) from the verbs and their arguments. These mathematical lattices are a means for deriving classifications (in this case of verbs) in a manner that naturally groups similar items and forms a generalisation/specialisation hierarchy. Similar FCA techniques have been successfully used in a variety of linguistic applications (Priss, 2005), for example, by Dyvik (2004) for constructing dictionaries from a bilingual corpus. With respect to the FCA techniques, Stepanova's (2009) approach is similar to Basili's, but she uses these techniques in combination with a different set of other linguistic tools. She mainly exploits the implied categorical information of genetive phrases from corpora to build a concept-oriented lexicon for use in a question answering system. Our idea is to apply similar FCA-based techniques to the problem of enhancing the semantic tags of RT.

We are predominantly interested in techniques that are reasonably fast and cheaply to implement (making use of already existing linguistic and FCA software and freely available resources, such as WordNet, dictionaries and corpora). Since RT contains about 200,000 entries (i.e. words that occur in a specific sense in RT), manual tagging of RT would be very expensive and time consuming. Automatically derived semantic tags are not expected to have the same kind of quality as manually derived tags, but can be obtained quickly and cheaply. The main research question of this talk is to determine the most promising techniques that result in meaningful semantic tags for RT. In particular, it is important to determine which tags are easy to derive, which tags already exist in RT and only need to be extended, and which tags are promising for the use in NLP tasks.

This talk starts with a brief overview of the historical background, the FCA technologies that are used and of the current structures that are available in an electronic version of RT. The details of FCA-based linguistic corpus processing will be discussed, and our up-to-date results in this research will be presented.

References.

Basili, R.; Pazienza, M.; Vindigni, M. (1997). Corpus-driven unsupervised learning of verb subcategorization frames. AI*IA-97.

Cimiano, P.; Staab, S.; Tane, J. 2003. Automatic Acquisition of Taxonomies from Text: FCA meets NLP. Proceedings of the ECML/PKDD Workshop on Adaptive Text Extraction and Mining, p. 10-17.

Dyvik, H. 2004. Translations as semantic mirrors: from parallel corpus to wordnet. Language and Computers, Vol. 49, iss. 1, Rodopi, p. 311-326.

Ganter, Bernhard, Wille, Rudolf (1999). Formal Concept Analysis. Mathematical Foundations. Berlin-Heidelberg-New York: Springer.

Masterman, Margaret (2005). Language, Cohesion and Form. Edited by Yorick Wilks. Cambridge University Press.

Priss, Uta (2005). Linguistic Applications of Formal Concept Analysis. In: Ganter; Stumme; Wille (eds.), Formal Concept Analysis, Foundations and Applications. Springer Verlag. LNAI 3626, p. 149-160.

Priss, Uta; Old, L. John (2009). Revisiting the Potentialities of a Mechanical Thesaurus. In: Ferre; Rudolph (eds.), Proceedings of the 7th International Conference on Formal Concept Analysis, ICFCA'09, Springer Verlag.

Stepanova, Nadezhda A. (2009). Automatic acquisition of lexico-semantic knowledge from corpora. SENSE'09 Workshop (in publication).