Linguistic Data Mining with FCA

Uta Priss
ZeLL, Ostfalia University of Applied Sciences
Wolfenbüttel, Germany

www.upriss.org.uk

The use of lattice theory for linguistic data mining applications in the widest sense has been independently suggested by different researchers. For example, Masterman (1956) suggests using a lattice-based thesaurus model for machine translation. Mooers (1958) describes a lattice-based information retrieval model which was included in the first edition of Salton’s (1968) influential textbook. Sladek (1975) models word fields with lattices. Dyvik (2004) generates lattices which represent mirrored semantic structures in a bilingual parallel corpus. These approaches were later translated into the language of Formal Concept Analysis (FCA) in order to provide a more unified framework and to generalise them for use with other applications (Priss (2005), Priss & Old (2005 and 2009)).

Linguistic data mining can be subdivided into syntagmatic and paradigmatic approaches. Syntagmatic approaches exploit syntactic relationships. For example, Basili et al. (1997) describe how to learn semantic structures from the exploration of syntactic verb-relationships using FCA. This was 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.

Paradigmatic relationships are semantic in nature and can, for example, be extracted from bilingual corpora, dictionaries and thesauri. FCA neighbourhood lattices are a suitable means of mining bilingual data sources (Priss & Old (2005 and 2007)) and monolingual data sources (Priss & Old (2004 and 2006)). Experimental results for neighbourhood lattices have been computed for Roget’s Thesaurus, WordNet and Wikipedia data (Priss & Old 2006, 2010a and 2010b).

Previous overviews of linguistic applications of FCA were presented by Priss (2005 and 2009). This presentation summarises previous results and provides an overview of more recent research developments in the area of linguistic data mining with FCA.

References


