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

1. Basili, R.; Pazienza, M.; Vindigni, M. (1997). *Corpus-driven unsupervised learning of verb subcategorization frames.* AI*IA-97.
2. 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.
3. Dyvik, H. (2004). *Translations as semantic mirrors: from parallel corpus to wordnet.* Language and Computers, 49, 1, Rodopi, p. 311-326.

4. Masterman, Margaret (1956). *Potentialities of a Mechanical Thesaurus.* MIT Conference on Mechanical Translation, CLRU Typescript. [Abstract]. In: Report on research: Cambridge Language Research Unit. Mechanical Translation 3, 2, p. 36. Full paper in: Masterman (2005).
5. Mooers, Calvin N. (1958). *A mathematical theory of language symbols in retrieval.* In: Proc. Int. Conf. Scientific Information, Washington D.C.
6. Priss, Uta; Old, L. John (2004). *Modelling Lexical Databases with Formal Concept Analysis.* Journal of Universal Computer Science, 10, 8, p. 967-984.
7. 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.
8. Priss, Uta; Old, L. John (2005). *Conceptual Exploration of Semantic Mirrors.* In: Ganter; Godin (eds.), Formal Concept Analysis: Third International Conference, ICFCA 2005, Springer Verlag, LNCS 3403, p. 21-32.
9. Priss, Uta; Old, L. John (2006). *An application of relation algebra to lexical databases.* In: Schaerfe, Hitzler, Ohrstrom (eds.), Conceptual Structures: Inspiration and Application, Proceedings of the 14th International Conference on Conceptual Structures, ICCS'06, Springer Verlag, LNAI 4068, p. 388-400.
10. Priss, Uta; Old, L. John (2007). *Bilingual Word Association Networks.* In: Priss, Polovina, Hill (eds.), Proceedings of the 15th International Conference on Conceptual Structures, ICCS'07, Springer Verlag, LNAI 4604, p. 310-320.
11. Priss, Uta (2009). *Formal Concept Analysis as a Tool for Linguistic Data Exploration.* In: Hitzler, Pascal; Scharfe, Henrik (eds.), Conceptual Structures in Practice, Chapman & Hall/CRC studies in informatics series, p. 177-198.
12. 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.
13. Priss, Uta; Old, L. John (2010a). *Concept Neighbourhoods in Knowledge Organisation Systems.* In: Gnoli; Mazzocchi (eds.), Paradigms and conceptual systems in knowledge organization. Proceedings of the 11th International ISKO Conference, p. 165-170.
14. Priss, Uta; Old, L. John (2010b). *Concept Neighbourhoods in Lexical Databases.* In: Kwuida; Sertkaya (eds.), Proceedings of the 8th International Conference on Formal Concept Analysis, ICFCA'10, Springer Verlag, LNCS 5986, p. 283-295.
15. Salton, Gerard (1968). *Automatic Information Organization and Retrieval.* McGraw-Hill, New York.
16. Stepanova, Nadezhda A. (2009). *Automatic acquisition of lexico-semantic knowledge from corpora.* SENSE'09 Workshop. Available at http://ceur-ws.org/Vol-476/.
17. Sladek, A. (1975). *Wortfelder in Verbänden.* Gunter Narr Verlag, Tübingen.