Exploring Corpus Data Using Data Mining: A Project of Automatic Term Recognition

Dominika Kováříková (Institute of the Czech National Corpus, Charles University in Prague, Czech Republic)

The corpus data are so vast and rich that they cannot be manually processed by a person. For a corpus-based research, we need concordancers and other textual analysis tools. Data mining is the next level of corpus data processing. It is a (semi)automatic process of discovering meaningful patterns in substantial quantities of data using computer algorithms (Witten and Frank 2005, 5). Data mining is very useful for a corpus-based exploration of large amounts of texts, especially when enriched by details behind the concordances: word frequencies and distribution, contextual information, or linguistic information about each of the tokens.

Data mining can be used in various linguistic research tasks such as text classification (Sebastiani 2002), topic identification (Clifton et al. 2004), study of linguistic properties of different types of texts (i.e. scientific texts, Teich and Fankhauser 2010) or examination of human processing of lexicon (Baayen 2005). It is also able to automatically identify specific types of lexical items such as terms (Kováříková 2014), collocations or idioms.

The presented research focuses on automatic term recognition using data mining. Thanks to the opportunity to work with very large and diverse data of the Czech National Corpus (100 million words in an easily accessible corpus SYN2010 that includes academic texts of almost 40 disciplines), and thanks to the chosen approach of data mining, the scope of the research was quite wide: single-word terms as well as multi-word terms were identified in a number of academic disciplines. The research provides not only a list of automatically identified terms in Czech academic texts but also valuable findings about terms and terminology in general. In addition, the disciplines can now be compared based on the retrieved terms (number of terms, lists of the most frequent terms, terms shared by two or more disciplines).

For the project of automatic term identification, the data mining platform WEKA (Witten and Eibe 2005) required data in form of a very large matrix with thousands of lines and tens of columns. Each line was assigned to one token/text position (or a bigram for multi-word term identification), whereas each column contained values of one of the characteristics relevant for recognizing terms in a text. The features suitable for term identification are usually linguistic (e.g. POS, word structure), or are based on word frequency and distribution in texts (e.g. frequency in a given academic discipline, distribution in disciplines), or they monitor the context of the word (e.g. average distance of two words in texts). Data mining is able to track complex non-linear relations between individual term characteristics which results in effective term identification.

In addition, data mining has the capacity to identify the most relevant of the examined features (feature ranking). The highest-ranked characteristics of single word terms are derived from distribution and the frequency of a given lexical item:

1. the ratio of the relative frequency in a given academic discipline to the relative frequency in the reference corpus (containing fiction and journalism);
2. average reduced frequency (shows evenness of distribution throughout the corpus as well as the frequency of the word in corpus, see Savický and Hlaváčová 2003);
3. relative distribution in academic disciplines;

4. standard deviation of the relative distance of two neighboring occurrences of the word.

Based on the highest-ranked features, a term can be simply described as a lexical item specific (and often unique) to a particular discipline. Such description of a term can complement standard term definitions such as ISO 1087-1:2000: Term is a „verbal designation of a general concept in a specific subject field."

The features evaluated as the most important for multi-word term recognition[1] are:

1. lexical association measure t-score;
2. the first word in bigram is a term;
3. the second word in bigram is a term;
4. lexical association measure mutual information (MI-score).

A multi-word term can be described as a multi-word lexical item (typically a collocation), usually with at least one of its components being a single-word term.

The data mining process was evaluated by standard statistical measures, i.e. accuracy, precision and recall[2] (Manning and Schütze 2000). The results are quite promising (see Table 1) – the high success rates guarantee the reliability of findings that are based on automatically identified terms, such as the estimation of percentage of terms in academic texts or the analysis of relations between academic disciplines based on shared terms.

Table 1. Results of the term recognition method using data mining

	Single-word terms	Multi-word terms (bigrams)
Accuracy	94%	97%
Precision	85%	81%
Recall	72%	75%

The estimate percentage of text positions occupied by terms differs for individual academic disciplines, ranging from less than 10% in psychology and philosophy to almost 35% in botany and geography. In humanities, the texts contain considerably less terms (app. 17%) than in natural and formal sciences (app. 27%). The average number of terms in academic texts is app. 22%.

Another way to compare disciplines, especially in terms of finding similarities and relations between them, is to examine shared terms – in this case a great number of automatically identified terms in various academic fields. Disciplines sharing terms is a common phenomenon. There are several main reasons why a term occurs in more than one discipline[3]:

1. There is a relationship of some kind between disciplines, such as in case of biology and medicine, or engineering and computer science.

1 Data are in the form of bigrams.

2 Accuracy is a statistical measure that is able to assess the proportion of the correctly labeled words in the text (terms and non-terms). One hundred percent accuracy means that all the true terms were classified as terms and all the true non-terms were classified as non-terms. Precision is the ratio of the correctly identified terms to all words labeled as terms (correctly and incorrectly). Recall is the ratio of the correctly identified terms to all true terms (labeled as terms and as non-terms).

3 This interesting topic is of course more complex and would deserve further research.

2. The term is used metaphorically or with different meaning in one or more areas of study (e.g. *valency* in chemistry and in linguistics, *communication* in linguistics and transportation engineering).
3. A term in one field is a non-term in other disciplines (e.g. *author* in literature and in medicine).

The disciplines that shared the highest number of automatically identified terms were engineering and computer science, law and economics, sociology and history, and chemistry and biology. However, even quite distant fields such as transportation engineering and biology shared a small number of terms, e.g. *reflexní* (*reflective* and *reflexive*), *signály* (*signals*), *transport* (*transport*), *ventilace* (*vent* and *ventilation*). Although shared terms have usually rather high frequencies in texts and are interesting from linguistic point of view, it is important to understand that the majority of the terms are unique to one discipline[4] (or to a cluster of closely related disciplines such as chemistry and biochemistry, biology and medicine etc.), e.g. *N-nitrosofenylhydroxylamin, regurgitace triskupidální chlopně* (*triscupid valve regurgiation*).

Conlusion

It is evident that exploring corpus using data mining can be fruitful and valuable. Corpus data are a great match for data mining because they are very large and diverse in terms of usable features. On the other hand, data mining can provide not only a high success rate or a list of identified items (in this case terms), but also deeper insight into the linguistic theory and characteristics of the analysed linguistic phenomena.

References

Baayen, R. H. (2005). Data mining at the intersection of psychology and linguistics. In Cutler, A. (ed), *Twenty-First Century Psycholinguistics: Four Cornerstones*. Erlbaum.

Bečka, J.V. (1972). The lexical composition of specialized texts and its quantitative aspect. *Prague Studies in Mathematical Linguistics,* 4, 47-64.

Čermák, F. (2010). *Lexikon a sémantika*. Praha: NLN.

Clifton, C., Cooley, R. and J. Rennie (2004). TopCat: Data Mining for topic identification in a Text Corpus. *IEEE transaction on Knowledge and data engineering*.

Czech National Corpus - SYN2010 (2010). Praha: Institute of the Czech National Corpus, FF UK. Accessed November 19, 2014, http://www.korpus.cz.

Chung, T. M. (2003). A corpus comparison approach for terminology extraction. *Terminology,* 9(2), 221-246.

Cvrček, V. (2013). Kvantitativní analýza kontextu. Praha: NLN/ÚČNK.

Gamper, H. and O. Stock (1998/1999). Corpus-based terminology. *Terminology,* 5(2), 147-159.

Hall, M. et al. (2009). The WEKA Data Mining Software: An Update. *SIGKDD Explorations,* 11(1).

Hearst, M.A. (1999). Untangling Text Data Mining. *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*. Accessed December 2016, http://www.aclweb.org/anthology/P99-1001.

Heid, U. (1998/1999). A linguistic bootstrapping approach to the extraction of term candidates from German text. *Terminology,* 5(2), 161-181.

4 Particularly due to the fact that the majority of terms are multi-word lexical items.

Kageura, K. and B. Umino (1996). Methods of automatic term recognition: A review. *Terminology,* 3(2), 259-289.

Kit, C. and X. Liu (2008). Measuring mono-word termhood by rank difference via corpus comparison. *Terminology,* 14(2), 204–229.

Kováříková, D. (2014). *Kvantitativní charakteristiky termínů (Quantitative Characteristics of terms)*. Ph.D. Thesis, Charles University in Prague, Czech Republic.

L'Homme, M., U. Heid and J. C. Sager (2003). Terminology during the past decade (1994-2004): An editorial statement. *Terminology,* 9(2), 151–161.

Lauriston, A. (1995). Criteria for measuring term recognition. *EACL '95 Proceedingsof the seventh conference on European chapter of the Association for Computational Linguistics*. San Francisco: Morgan Kaufmann Publishers.

Manning, C. D. and H. Schütze (2000). *Foundations of Statistical Natural Language Processing*. Cambridge/London: The MIT Press.

Savický, P. and J. Hlaváčová (2003). Measures of word commonness. *Journal of Quantitative Linguistics,* 9(3), 215–231.

Sebastiani, R. (2002). Machine learning in automated text categorization. *ACM Computing Surveys,* 34(1), 1-47.

Šrajerová, D., O. Kovářík and V. Cvrček (2009). Automatic term recognition based on data-mining techniques. *Proceedings of Computer Science and Information Engineering - CSIE*. Los Angeles.

Teich, E. and P. Fankhauser (2010). Exploring a corpus of scientific texts using data mining. *Corpus-linguistic applications. Current studies, new directions*. Gries, S.Th., Wulff, S. and Mark Davies (Eds.). Amsterdam/New York, NY, 2010, VI, Rodopi.

Ville-Ometz, F., Royauté, J. and A. Zasadzinski (2007). Enhancing in automatic recognition and extraction of term variants with linguistic features. *Terminology,* 13(1), 35–59.

Wermter, J. and U. Hahn (2005). Finding new terminology in very large corpora. *Proceedings of the 3rd International Conference on Knowledge Capture (KCAP 2005)*.

Witten, Ian H. and E. Frank (2005). *Data Mining: Practical Machine Learning Tools and Techniques*. Amsterdam: Elsevier.

Yang, H. (1986). A new technique for identifying scientific/technical terms and describing science texts. *Literary and Linguistic Computing,* 1(2), 93-103.