Abs-64

Lyne Da Sylva

Extracting a vocabulary for structured indexing: comparison among three corpora

This project aims to develop lexical resources for automatic indexing. We are particularly interested in automatically producing back-of-the-book style indexes, which exhibit structured entries expressing various semantic relationships between main headings and subheadings and whose creation can be quite challenging by automatic methods.

A type of useful vocabulary for this type of indexing is defined, the basic scholarly vocabulary (BSV). It contains words which are used across all disciplines with roughly the same meaning. Examples are: “development”, “structure”, “onset”, “absence”, etc. It is akin to Ogden’s Basic English (Ogden, 1930) and to Coxhead’s Academic Word List (Coxhead, 2000), but with a different purpose and properties. This type of word can be combined with specialized vocabulary items to form evocative, structured index entries such as the following: “combustion engine, structure” or “First World War, onset”. An automatic indexing prototype has been developed which combines such specialized terms with BSV terms both occurring within a set window of text (Da Sylva and Doll, 2005). To improve on the system’s internal list of manually-compiled BSV words, an experiment of semi-automatic extraction of the basic scholarly vocabulary lexical items from a large English corpus of 14 million words was devised and reported on earlier (Da Sylva, 2009); it consists of abstracts of scholarly articles in pure and applied sciences as well as in the humanities and social sciences.

The present paper describes the results of a further experiment of semi-automatic extraction from two additional corpora of abstracts of scholarly articles. The goal was to extract BSV lists for both English and French; two parallel corpora were used, containing abstracts of articles in pure and applied science only. The abstracts were (human) translations of each other and represented approximately 2 million words (2,5 million for the French one). The new extraction task for French was successful in doubling the size of a previously manually compiled list. Moreover, it has proved more efficient than the equivalent task on its parallel English corpus yielding a greater proportion of BSV in the top-ranking words.

A comparison among the BSV extracted from each of the three corpora has yielded interesting results. Specifically, the extraction task on the two parallel scientific corpora has favoured words typical of the pure and applied sciences (measurements such as “rate”, “increase” and “value”), scarcer in social sciences and humanities. This has confirmed our initial choice for a wider-ranging corpus. Also, the results on the smaller scientific English corpus are inferior to those on the larger, wide-ranging one (where a greater proportion of extracted words belong to the BSV). This suggests that the results on the French corpus could be improved upon, with an appropriate corpus.

References

Coxhead, Averil. (2000). ‘A New Academic Word List’. TESOL Quarterly, 34(2), p. 213-238.

Da Sylva L. (2009). ‘Corpus-based derivation of a "basic scientific vocabulary" for indexing purposes’. In *Proceedings of the Corpus Linguistics Conference*, Univ. Liverpool, 21-23 July 2009.

Da Sylva, Lyne; Doll, Frédéric. ‘A Document Browsing Tool: Using Lexical Classes to Convey Information’. In Lapalme, Guy; Kégl, Balász. *Advances in Artificial Intelligence: 18th Conference of the Canadian Society for Computational Studies of Intelligence, Canadian AI 2005* (Proceedings), New York : Springer-Verlag, 2005, pp. 307-318.

Ogden, Charles K. (1930) *Basic English: A General Introduction with Rules and Grammar*. London:

Paul Treber & Co., Ltd.