Abs-181

Richard Forsyth and Serge Sharoff (both University of Leeds)

From crawled collections to comparable corpora: An approach based on automatic archetype identification

With the rise of the "web as corpus" (Kilgarriff & Grefenstette, 2003) many corpora are being compiled by "crawling" the world-wide web using a variety of search strategies (Baroni, et al, 2009). This approach has several advantages, not least the fact that it offers a rapid way for collecting documents in new and rapidly developing fields in a large range of languages, so that we can compare usages across them. Such corpora can be made closer to traditional corpora such as the BNC by applying metadata (Sharoff, 2007).

The present study forms part of a project whose ultimate aim is to facilitate the rapid development of bilingual terminological lexicons in emerging technological fields by compiling comparable corpora in a number of languages (see www.ttc-project.eu). To achieve this goal we first need to ensure monolingual comparability. This paper describes an initial step in that process, the partitioning of the crawled collection into emergent subgroups using unsupervised machine learning, and then the identification of archetypal texts, representative of their subgroups, which can be used as probes in their own right to find additional documents of a similar type.

The procedure has a number of novel aspects. Firstly, the features used to characterize textual similarity and difference pay more respect to the inescapably sequential nature of language than the more conventional term-vector (or "bag of words") approach. Our feature-finding technique is based what Sinclair (1991) calls the "idiom principle", namely the tendency for speakers and writers, as well as listeners and readers, to work with chunks of language rather than isolated words. The results of such chunking have been referred to by a variety of terms, such as "collocations", "concgrams", "lexical bundles", "multi-word expressions", "prefabricated phrases", "skipgrams", among other designations (Cheng et al., 2006). All are generalizations of the basic notion of an n-gram, but different authors have generalized this concept in slightly different ways, and thus the meanings of these terms overlap in a somewhat confusing manner. As the terminology for flexible multi-element linguistic units is not yet standardized, we refer in this paper to "flexigrams" (Min & McCarthy, 2010).

Secondly we use an evolutionary algorithm to find archetypes, by optimizing an objective function which generalizes that used in Ward's method of agglomerative clustering (Ward, 1963).

Thirdly, our software can optionally reduce the dimensionality of the archetype data by finding a subset of informative attributes, using essentially the same algorithm.

We will describe the results of applying these methods to collections gathered from web-crawls designed to gather texts concerning renewable energy in Chinese, English and Russian, among others.

References

Baroni, M., Bernardini, S., Ferraresi, A., and Zanchetta, E. (2009). The WaCky wide web: a collection of very large linguistically processed web-crawled corpora. Language Resources and Evaluation, 43(3), 209–226.

Cheng, W., Greaves, C. and Warren, M. (2006). From n-gram to skipgram to concgram. International Journal of Corpus Linguistics, 11(4), 411-433.

Kilgarriff, A. & Grefenstette, G. (2003). Introduction to the special issue on web as corpus. Computational Linguistics 29 (3), 333-347.

Min, H.C. & McCarthy, P.M. (2010). Identifying Varietals in the Discourse of American and Korean Scientists: A Contrastive Corpus Analysis Using the Gramulator. Proc. 23rd International Florida AI Society Conference (FLAIRS-2010), 247-252.

Sharoff, S. (2007). Classifying web corpora into domain and genre using automatic feature identification. In Proc. of Web as Corpus Workshop, Louvain-la-Neuve.

Sinclair, J. (1991). Corpus, Concordance and Collocation. Oxford: OUP.

Ward, J.H. (1963). Hierarchical grouping to optimize an objective function. J. American Statistical Association, 58, 236-244.