Quantifying Document Dissimilarity within and across Languages: a Benchmarking Trial

Richard Forsyth
Centre for Translation Studies
School of Modern Languages & Cultures
University of Leeds
Leeds LS2 9JT
U.K.
(+44) 113 3433195
R.S.Forsyth@leeds.ac.uk

Serge Sharoff
Centre for Translation Studies
School of Modern Languages & Cultures
University of Leeds
Leeds LS2 9JT
U.K.
(+44) 113 3437287
S.Sharoff@leeds.ac.uk

Quantifying Document Dissimilarity within and across Languages: a Benchmarking Trial

Abstract

Quantifying similarity or dissimilarity between documents is an important problem in authorship attribution (Juola, 2006) corpus composition (Kilgarriff, 2001), information retrieval (Salton and McGill, 1983), plagiarism detection (Clough and Gaizauskas, 2009), term extraction (Li and Gaussier, 2010), text mining (Weiss et al., 2005) and many other natural-language processing tasks.

Many indices have been used for these purposes, but little effort has been devoted to calibrating such indices by systematically comparing the outputs of various textual dissimilarity functions with a text-external standard. One reason for this is that, although texts can be placed into categories on the basis of genre, register, topic and other discourse-level attributes, there is no widespread agreement on how to deal quantitatively with the fact that some categories differ more than others. The present paper addresses this issue by proposing a method for establishing a text-external similarity space, on the basis of readers' judgements, within which texts can be located. This then provides the grounding for a benchmarking study that compares several text-based measures of dissimilarity with the dissimilarities derived from the text-external framework, in five different languages. Some widely used measures perform comparatively poorly in this trial. In particular, rank correlation consistently outperforms cosine similarity on our data.

A method of generalizing such within-language dissimilarity measures across languages is also proposed.

References

Clough, P. & Gaizauskas, R. (2009). Corpora and text re-use. In A. Lüdeling, M. Kytö and A. McEnery, eds., *Handbook of Corpus Linguistics*, 1249-1271, Mouton de Gruyter.

Juola, P. (2006). Authorship attribution. *Foundations & Trends in Information Retrieval*, 1(3), 233-334.

Kilgarriff, A. (2001). Comparing corpora. *International Journal of Corpus Linguistics*, 6(1), 1-37.

Li, B. & Gaussier, E. (2010). Improving corpus comparability for bilingual lexicon extraction from comparable corpora. In *Proc. 23rd International Conference on Computational Linguistics (Coling 2010)*, Beijing.

Salton, G. & McGill, M.J. (1983). *Introduction to Modern Information Retrieval*. McGraw-Hill.

Weiss, S., lndurkhya, N., Zhang, T. & Damerau, F. (2005). *Text Mining: Predictive Methods for Analyzing Unstructured Information*. Springer.