Automatic Morphology

John Goldsmith, Svetlana Soglasnova, and Derrick Higgins
The University of Chicago

Summary

The purpose of our experiment was to get an estimate of the usefulness of an automatic morphological analyzer that we have been working on. The automatic morphological analyzer (called AutoMorphology) accepts a text of any length and automatically determines the stems and suffixes of the words in the language. For a full description of the algorithm, see http://humanities.uchicago.edu/faculty/goldsmith/Linguistica2000/Paper/paper.html. There has been discussion in the IR literature of the tradeoffs between using automatic language-independent means for stemming and language-dependent (hand-written) stemming algorithms{Xu and Croft 1995, Lennon et al.1981}; our goal has been to get the best of both approaches by using an automatic but language-particular stemming algorithm.

In a language such as English, where good stemmers exist already and where morphology is relatively simple, we do not expect any improvement over the current art from our system. {The usefulness of stemming for English for IR purposes is a controversial issue; while it has been claimed to be overly beneficial [Sparck Jones 1999, Kowalski 1997, Frakes 1992], the gain in recall is often offset by the loss in precision [Strzalkowski et al. 1999:124, Harman 1991]. But we hope to find improvements over the current arts as we look at more languages with richer and more complex morphologies for which there do Not currently exist stemmers for use in IR research and applications, {but for whose likes stemming improves IR significantly, as shown for French, Slovene, Finnish, Dutch and Russian, [Jacquemin & Tsoukerman 1999, Popovic and Willett 1992, Koskenniemi 1996, Kraaij and Pohlmann 1996, Nozhov 1998.]

We implemented the SMART system as our retrieval engine, with standard tf*idf term weighting. The primary modification we made to the system was to incorporate our custom stemmer, which was automatically derived from the corpora for each language, although for German we also did some corpus-driven breaking of compound words.

The stemmer incorporated into SMART's indexing function basically consisted of a hash-table lookup from a file of previously stemmed words. This file was produced by running Automorphology on the document collection for each language, with no human intervention or language-specific configuration. So the retrieval model we used was the standard vector-space model, with stemming done both on indexed document terms and query terms.

References

Frakes, W.B. 1992. Stemming Algorithms. In Frakes, W.B., Ricardo Baeza-Yates (Eds.) Information Retrieval Data Structures and Algorithms. Prentice Hall: New Jersey, p.131-160

Harman, D. 1991. How effective is suffixing? Journal of the American Society for Information Science, 42(1), 7-15

Jacquemin, Christian and Evelyne Tsoukermann. 1999. NLP for term variant extraction: synergy between morphology, lexicon, and syntax. In Strzalkowski, Tomek (Ed.)Natural Language Information Retrieval. Dordrecht: Kluwer Academic Publishers p.25-74

Koskenniemi, K. 1996 Finite-state morphology and information retrieval. In Proceedings of the ECAI-96 Workshop on Extended Finite State Models of Language,pp.42-5, ECAI, Budapest, Hungary (Available on CD-ROM in Kornai 1999).

Kraaij, W. and Pohlmann, R. 1996 Viewing stemming as recall enhancement. In Proceedings, 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '96) pp.40-8, Zurich

Krovetz, R. 1993. Viewing morphology as an inference process. In Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp.191-202

Lennon, M., Pierce, D.C., and Willett, P. 1981. An evaluation of some conflation algorithms. Journal of Information Science, 3, pp.177-183

Nozhov, Igor'. 1998. Prikladnoi morpfologicheskii analiz [Applied morphological analysis].In: Pravovaia Informatika. 1998, v.4. Moscow

Popovic and Willett 1992 . The effectiveness of stemming for natural-language access to Slovene textual data. Journal of the American Society for Information Science, 43(5) pp.384-390

Sparck Jones, Karen. 1999. What is the role of NLP in text retrieval? In Strzalkowski, Tomek (Ed.) Natural Language Information Retrieval. Dordrecht: Kluwer Academic Publishers, p.1-24

Strzalkowski, Tomek (Ed.) Natural Language Information Retrieval. Dordrecht: Kluwer Academic Publishers

Xu, Jinxi and W.Bruce Croft. 1995. Corpus-Based Stemming using Co-occurrence of Word Variants. In Proceedings for the Fourth Annual Symposium on Document Analysis and Information Retrieval, pages 147-159, Las Vegas, Nevada, April 1995.