

Affiliation Details

Harald Hammarström
Dept of Comp Sci
Chalmers University
412 96 Gothenburg
SWEDEN

email: harald2@cs.chalmers.se
homepage: <http://www.cs.chalmers.se/~harald2>
phone: +46-70-7184339

Also member of GSLT and NGSLT.

Description of PhD project

I do resesarch in the area of unsupervised induction of morphology (and to a lesser extent unsupervised induction of part-of-speech categories), i.e., input raw (unannotated) text data and output some kind of descripton of the morphology of that language. This is to be achieved with as little human tuning, built-in language specific knowledge etc. as possible. See my Lic. Thesis for motivation and specifics (Hammarström 2007). A concrete target application area is multilingual IR.

I have now reached the stage where I can successfully output lists of probable stems, lists of probable suffixes (and/or prefixes) and lists of probably paradigms, i.e., sets of suffixes that tend to appear on the same stems. However, my methods are completely ignorant as to semantics, so they cannot even make a guess as to, e.g., if two suffixes are allomorphs or whether a paradigm occurs with animate nouns only. Things like LSA are relevant here and such techniques are reasonably well-understood. Another path appears to be with ontologies, description logics and so on, of which I know too little. The course in question appears to give a state-of-the-art picture of ontologies and how they may be exploited (especially in IR).

References

- Hammarström, H. (2007). Unsupervised learning of morphology: Survey, model, algorithm and experiments. Thesis for the Degree of Licentiate of Engineering, Department of Computer Science and Engineering, Chalmers University, 91 pp.