Comments on Koskenniemi

Eva Ejerhed
Department of Linguistics
University of Umea
90187 Umea
Sweden
ejerhed@ling.umu.se

The paper concerns the problem of finding base forms of given word forms in the context of IR applications. Since errors in base form reduction lead to errors in retrieval of all inflected forms of a given keyword, the problem is an important one.

After a survey of various known approaches to this problem, and of the drawbacks of these approaches, the paper presents a new proposal for solving it by using a new lexicon transducer RsubH, based on morphological guessing.

The presentation of this transducer is very sketchy. In it "we replace the fixed list of stems with a general expression which is open enough to cover all anticipated proper names, foreign words, new terms, acronyms etc. which could occur in texts. It is worth while to note that only productive inflectional patterns need to be included in RsubH. Exceptions and idiosyncratic patterns are not likely to occur in new words."

I am not sure that the last assumption is correct, i.e. that unknown words are likely to follow the normal morphological patterns of a language where morphological guessing would be helpful. For unknown words that are names, acronyms and foreigh words, there can be considerable deviations from the productive morphotactics (and graphotactics) of a language.

However, for new terms, the strategy proposed could be quite useful. It would be interesting to have some quantitativa data on what proportion of unknown keywords in a given text (even if small) that are successfully covered by the proposal of the paper, and that would not be covered by the already existing approaches that the paper describes.

The proposed method, as I understand it, seems to me to have in common with other current proposals in NLP the general feature that it overcomes the sparseness of a finite amount of observed/known data by using "higher order" classes.

Unknown words, whether keywords or not, and unknown analyses of known words, are major problems that need to be addressed. More recent studies of (even) English text data bases have revealed that new word types keep ooccurring as the amount of text grows (Church, 1995, Ngrams, ACL-tutorial). There is no point at which growth stops, as is claimed in the paper.

Proceedings of the ECAI 96 Workshop
Extended Finite State Models of Language
Edited by A. Kornai.