Abstract

Johansson Kokkinakis, Sofie, *En studie över påverkande faktorer i ordklasstaggning. Baserad på taggning av svensk text med EPOS.* Göteborg: 2002, 221 s. Språk: svenska. Doktorsavhandling i språkvetenskaplig databehandling. Institutionen för svenska språket, Göteborgs universitet. *A study of factors influencing performance in part of speech tagging. Based on tagging of Swedish text with EPOS.* Language: Swedish. Doctoral dissertation in Natural Language Processing. Department of Swedish Language, Göteborg University. 2002. 221 pp.

Part of speech (POS) tagging is known to be one of the first steps in natural language processing. The aims of the study is to determine to what degree it is possible to enhance performance in POS tagging and at what cost. These issues are addressed by examining five different factors bearing on the performance of part of speech tagging in Swedish. These factors are divided into two subgroups depending on how directly or indirectly they seem to affect performance. The *direct* factors are lexical as well as textual initial text material and computational method, whereas the *indirect* ones are tagset and evalutation method.

In order to determine the most adequate approach to improving performance in POS tagging, results from the statistical method were compared with approaches focusing on well-established problem areas. It is concluded that statistical methods are certainly very important, up to a point, when it comes to correctly tagged text. In the end, however the approaches which focused on specific problem areas seemed to have greater effect.

Special attention was paid to evaluation. The performance of the tagger used was tested on a text of 100,000 words, and the errors discovered (5,000) were examined in detail. Since ordinary evaluation measures seem far too superficial to account for the relevant factors underlying the errors, a new evaluation method is proposed, $H_{measure}$, in order to handle homography in the materials more adequately.

Key words: part of speech tagging, lexical databases, ambiguity, homography, disambiguation, tagsets, training and test corpora, machine learning, evaluation.

Distribution: Inst. för svenska språket (Dept. of Swedish Language)
Box 200, SE-405 30 Göteborg, Sweden