Donnerstag
09.03.2017
09:00 – 10:00
B4 1, 0.04

Automatic register annotation for linguistic research?

Felix Bildhauer
IDS Mannheim
bildhauer@ids-mannheim.de

Roland Schäfer
Freie Universität Berlin
roland.schaefer@fu-berlin.de

Corpus linguists often require document-level meta data such as document registers. Since manual annotation of registers is infeasible for very large corpora (such as crawled web corpora), the only viable alternative is automatic classification. Most approaches to automatic register annotation rely on a (usually high) number of linguistic features extracted from the documents. In our talk, we explore the usefulness of automatically annotated register categories for models of alternation phenomena in morpho-syntax. An obvious conceptual problem of such models is circularity if the registers are operationalized in terms of document-internal morpho-syntactic features. However, we discuss a technical aspect, namely that models can be of much higher quality if the raw features are used instead of aggregated register categories. To this end, we conducted two case studies on German: a) case variation after prepositions and b) inflection of adjectives after pronominal adjectives. We created an ad-hoc corpus of approx. 0.6 m documents/ 0.4 bn tokens sampled from the DECOW16 web corpus (Schäfer and Bildhauer, 2012) and the DeReKo corpus (Kupietz et al., 2010) and extracted a large number of features at the document level. First, we used these as predictors in a generalized linear model (GLM) modeling the two alternation phenomena. Second, we used the features to induce document categories in the corpus, then fitting a number of alternative models with these categories as predictors. Additionally, we used Monte Carlo simulations to demonstrate how aggregation of features into register categories can systematically affect the quality of GLMs. We compare the quality of the different models thus obtained and discuss the implications of our findings for using automatically annotated high-level categories in research on grammatical and morphological alternation phenomena.

References: • Kupietz, M., Belica, C., Keibel, H. and Witt, A. (2010). The German Reference Corpus DeReKo: A Primordial Sample for Linguistic Research. In N. Calzolari, et al. (eds.), Proceedings of LREC'10, pages 1848–1854, Valletta, Malta: ELRA. • Schäfer, R. and Bildhauer, F. (2012). Building large corpora from the web using a new efficient tool chain. In N. Calzolari, et al. (eds.), Proceedings of LREC'12, pages 486–493, Istanbul: ELRA.