Beyond Multidimensional Analysis: Probabilistic Register Induction for Large Corpora

Felix Bildhauer[1], Roland Schäfer[2]

[1] Leibniz-Institut für Deutsche Sprache (IDS), Mannheim, [2] Humboldt-Universität zu Berlin, SFB1412/A04

bildhauer@ids-mannheim.de , roland.schaefer@fu-berlin.de

Mittwoch,
04.03.2020
15:45–16:30
ESA1 W Foyer

The automatic analysis of the register in which a corpus document is written is prominently associated with Biber's (1988; 1995) Multidimensional Analysis (MDA). We present an approach superficially similar to MDA but which solves three major conceptual problems of MDA by using Bayesian inference to uncover registers – or rather potential registers (pregisters). First, in MDA, registers are associated discretely with documents, and each document can only instantiate one specific register, whereas we allow registers to be associated probabilistically with documents, and we allow mixtures of registers in single documents. Given that many linguistic phenomena are now understood as being probabilistic in nature (cf. Schäfer 2018), we suggest that this is a much more realistic assumption. Second, we assume the surface features to be associated with registers in a probabilistic manner for similar reasons. Third, we do not use a catalogue of registers assumed to exist a priori but we merely infer pregisters via clusters of surface features. The question of which pregisters actually correspond to actual registers with an identifiable situational-communicative setting will be dealt with in a future stage of the project using theory-driven evaluation and experimental validation. Given our assumptions about the nature of the mapping between features and pregisters as well as pregisters and documents, an obvious algorithm to use is Bayesian inference in the form of Latent Dirichlet Allocation (LDA; Blei et al. 2003; Blei 2012) as used in Topic Modelling. In our approach, we deal with pregisters instead of topics and with distributions of lexico-grammatical surface features instead of lexical words. The LDA algorithm otherwise performs an exactly parallel inference task. We first show how we extended the COReX feature extraction framework (Bildhauer & Schäfer in prep.) developed at FU Berlin and the IDS Mannheim in order to provide a large enough number of features for the LDA algorithm to work. We then present first results and discuss how we tuned the LDA algorithm and the feature set to lead to interpretable results. In order to be able to interpret the pregisters found by LDA, we extract the documents which most strongly instantiate the inferred pregisters. We introduce the PreCOx20 sub-corpus of the DECOW German web corpus, in which those prototypical documents are collected for further analysis w.r.t. their situational communicative setting.

References: Biber, D. (1988). Variation across Speech and Writing. CUP. Biber, D. (1995). Dimensions of Register Variation: A Cross-Linguistic Comparison. CUP. Bildhauer, F. & R. Schäfer (in prep.) Automatic register annotation and alternation modelling. Blei, D. M (2012). Probabilistic topic models. Communications of the ACM 55(4), 77–84. Blei, D. M., A. Y. Ng & M. I. Jordan (2003). Latent Dirichlet Allocation. Journal of Machine Learning Research 3, 993–1022. Schäfer, R. (2018). Probabilistic German Morphosyntax. Habilitation thesis. HU Berlin.