Good sentences, bad sentences

Marie-Pauline Krielke
Universität des Saarlandes
mariepauline.krielke@uni-saarland.de

Jörg Knappen
Universität des Saarlandes
j.knappen@mx.uni-saarland.de

During syntactic parsing many stumbling blocks can jeopardize the quality of the output. As others (e.g., Ortmann et al. 2019) have observed before, a key bottleneck in parsing is correct end-of-sentence recognition. In this paper, we present a set of measures to distinguish good sentences from bad sentences and show how filtering for good sentences improves accuracy in Universal Dependencies (UD) parses of historical German texts.

Our corpus (DTAW) features German scientific texts from the Deutsches Textarchiv (DTA, Geyken et al. 2018) between 1650 and 1899. To detect 'bad sentences' we use a similar approach as Didakowsky et al. (2012), who develop rules to extract good example sentences for a lexical resource. We look for sentences beginning with a word in lower case (in this case we mark also the preceding sentence as bad), sentences with fewer than 8 tokens, as well as sentences lacking a verb and all sentences in other languages than German. We retain approximately 74 million 'good' tokens out of initially 82 million tokens (rejection rate 9.42%).

We conduct three different types of evaluation: 1) General parsability 2) Number and accuracy of roots per sentence 3) Correctness of the assigned UD tag per token, correctness of the syntactic head of each token, and correctness of both labels (UD tag and head) per token. In all evaluations the parsing accuracy for good sentences improved significantly. Grammatical interpretablity is 100% for the good sentences and 71% for the bad ones. Evaluation of number of roots came out negative for the good sentences as on average a good sentence has 1.54 roots (1.48 for a bad one). Correctness of root labels results in a mean accuracy of 62% for good sentences and 40% for bad ones. Label accuracy is 92% for the good sentences (65% for the bad ones) and head correctness is 87% for the good sentences (65% for the bad ones). Our procedure can be useful for NLP applications such as syntactic parsing and other tasks where sentence splitting plays a role.

References: Didakowski, J. et al. 2012. Automatic example sentence extraction for a contemporary German dictionary. 15th EURALEX International Congress. • Geyken, A. et al. 2018. Das Deutsche Textarchiv als Forschungsplattform für historische Daten in CLARIN. In *Germanistische Sprachwissenschaft um 2020*, vol. 6. • Ortmann, K. et al. 2019. Evaluating Off-the-Shelf NLP Tools for German. In *KONVENS*.