Evaluation of neural coreference annotation of simplified German

Sarah Jablotschkin, Nele Benz, Heike Zinsmeister
Universität Hamburg

Simplified German

- Easy German (Leichte Sprache) and Plain German (einfache Sprache)
- Simplified lexicon, morphology and syntax
- Various target groups (Bredel & Maaß 2016, Baumert 2019)
- Players: Netzwerk Leichte Sprache, social and governmental institutions, broadcasting services (e.g. AWO, Bundestag, NDR)

Coreference


Fig. 1: Easy German by NDR

Research Questions

- How well does a neural coreference resolver (c2f, Schröder et al. 2021) trained on Standard German newspaper texts (TüBa-D/Z) detect mentions in simplified German texts?
- What kind of errors (error sources) occur in simplified and Standard German texts?

Annotation Study

LeiKo core corpus, complemented by standard German texts (taz Standard)

Data

1. Automatic coreference annotation by c2f
2. Manual correction and extension of the annotation in order to capture phenomena typical of simplified German (see guidelines on Zenodo, LeiKo v1.5)
3. Mentions: Identification of true/false positives and false negatives
4. Annotation of possible error sources: Linguistic context and noun phrase types (1,412 mentions in 28 texts).

Mentions

Referring expressions (e.g. noun phrases with definite or indefinite articles, personal pronouns, proper nouns)

Coreference links

Links between mentions referring to the same entity

Results

Fig. 2: Coreference evaluation scores

Fig. 3: Association plot of false negative mentions

Conclusions

- Quality of mention detection and coreference resolution in simplified German is comparable to Standard German given the modified guidelines (on avg. F1 LEA: 73.72, CoNLL: 74.52).
- Easy German has high proportions of def_NPs (NDR) or indef_NPs (tazleicht).
- All subcorpora have high error proportions of indef_NP. Statistically significant proportion in tazleicht (\( \chi^2_{\text{without}} = 65.136, df=12, p=2.573e-09 \)) → hypothesis: effect of genericity.

- No annotation of singleton mentions → follow up errors in identification.
- Simplified variants: Idiosyncratic spellings do not pose a problem (e.g. midpoint).
- Infrequent error issues: direct speech, personal pronouns, segmentation.
- Further investigation of correlations between mention type and formal entity characteristics. Hypothesis: more restricted variants have more uniform mentions/entities that are easier to resolve.

References: