Feature Exploration for the prediction of the German Vorfeld
Yulia Pikveich & Heike Zinsmeister
Fachbereich Sprachwissenschaft, Universität Konstanz

Motivation

German Vorfeld (‘prefield’)
- about 50% of declarative main clauses in newspaper texts do not start with the subject
- first position vs. order of other constituents in the clause (cf. Filippova & Strube 2007, Bader & Häussler 2010)

Challenge for automatic generation of contiguous text
- choosing an appropriate sentence beginning to support the fluency of a generated text (local coherence)
- applications: (multi-document) summarization, machine translation

Research question:
What kind of features are relevant for automatically determining the sentence beginning in German?

Data

- TüBa-D/2 treebank of written German (v.5)
  - daily issues of the newspaper “die tageszeitung” (taz)
  - annotation:
    - topological fields, constituency and grammatical functions (Telljohann et al. 2009)
    - coreference and anaphoric relations (Naumann 2006)
- Lemmatization by TreeTagger (Schmid 1995) – is included in newer Version of TüBa-D/2
- Semantic classes from GermNet (Hamp & Feldkamp 1997)

- Extraction of various features from sentence and constituent levels (Mousser & Zinsmeister 2009)
  - 28,102 declarative Verb Second clauses
  - 97,242 major constituents

Constituent-based classification

- Weka’s (online) implementation of C4.5 decision tree classifier (J48)
  - binary classification: Vorfeld versus ‘rest’
  - automatically pruned trees; 10-fold cross validation
  - starting with all features; then systematic variation
  - example training instance (simplified):

    | LEXh | TOTAL | SYNc | POSh | LENGTHc | ANA-TYPE | ANA-POS | VLEX | VOICE | CLASSc | VP | Vorfeld |
    |-----|-------|------|------|---------|----------|---------|------|-------|--------|---|----------|
    | 544 | 681   | 25   | 298  | 220     | 101      | 99      | 1     | 1     | 175    | 1 | Vorfeld  |

- Ten-fold cross validation:
  - accuracy: 71.1% w/o TOTAL: accuracy = 76.6%*** w/ MMF-NF: accuracy > 79.4%***

Sentence-based classification

- Most probable VF constituent per sentence
  - Perf script (input = decision tree’s constituent-based classification):
    for each target sentence s, if there are constituents classified as VF, constituent c with highest VF probability becomes Vorfeld else, constituent c with lowest rest probability becomes Vorfeld

- Results of pilot study
  - training set (1000 s) and test set (190 s): Baseline: subject in VF = 50%
  - 64% per-sentence accuracy
  - outperforms other studies on an earlier versions of the TüBa-D/2 treebank without manual anaphoric and coreference annotation (cf. Filippova & Strube 2007)

Discussion

What kind of features have an impact on the Vorfeld?
“More features are better features”?
- Lexical features
- improve Vorfeld precision; but: model is overfitting
- Discourse and semantic features
- anaphoric, coreferential and explosive correlate with Vorfeld position
- boost Vorfeld recall, but: evidence is too sparse (large group of ‘none’)
- Features that do not distinguish between VF and ‘rest’
- each feature was used in some pruned decision tree

Further considerations

- Prior classification whether a ‘postfield’ is expected; would improve most Vorfeld classifiers
- Taking text structure into account:
  - Vorfeld preferences are dependent on the position in the text; determined by information structure (Oudenhoven, chapter 2.3.4)
  - beginning of a text: subjective, temporal or local anchor
  - within a text: referential bridge to previously mentioned referents (kind of topic/theme)
- Better classification results by other methods?
  - machine learning state-of-the-art: Support Vector Machines
  - significance of interactions: regression model
- Evaluation that takes optimality and variance of Vorfeld into account

References

- TüBa-D/2 online http://www.exs.uni-tuebingen.de/darstellung/ [accessed, 03-03-2012]
- Weka (online) http://www.cs.waikato.ac.nz/ml/weka.html [accessed, 03-03-2012]

The research was partly funded by Europäischer Sozialfonds in Baden-Württemberg and by the Young Scholar Fund of the University of Konstanz.

What kind of features are relevant for automatically determining the sentence beginning in German?