Feature Exploration for the prediction of the German Vorfeld

Yulia Pilkevich & Heike Zinsmeister

Fachbereich Sprachwissenschaft, Universität Konstanz

Universität Konstanz

Motivation

Manchmal denkt Patrick Kundmüller, daß er seinen Doktorkittel an den Nagel hängen sollte. **Dann** aber glaubt er doch wieder an seinen Traum: eine Marktlücke entdeckt und einen Job zu haben, der auch noch Spaß macht. Eine Arbeit mit gelegentlichen Sternstunden. **Eine** erlebte er, als ihm eine Kundin überschwenglich siebzig Mark Trinkgeld in die Hand drückte. Der Bremer Kundmüller übt einen Beruf aus, der noch gar nicht lange existiert - er ist Bike-Doctor. (TüBa-D/Z v5: s19451-s19455)

- 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)
  - influence of information structure (previously mentioned elements, frame-setting elements, cf. Speyer 2007, Filippova & Strube 2007)

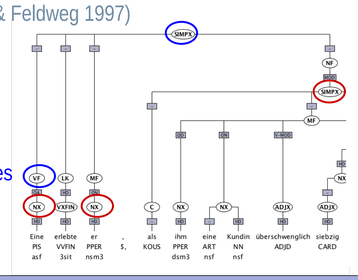
Vorfeld 'prefield'	sentence bracket	middle field	sentence bracket	postfield
one constituent	finite verb	constituents	verbal complex	some constituent types
Manchmal 'Sometimes'	denkt / hat 'thinks / has'	Patrick K. 'Patrick K.'	Ø / gedacht ' / thought'	dass … 'that …'
Eine 'one$_{ACC}$ [a magic moment]	erlebte 'experienced'	er 'he$_{NOM}$'		als ihm eine Kundin 70 Mark Trinkgeld gab. 'when a customer gave him a tip of 70 marks.'

- 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/Z 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/Z
- Semantic classes from GermaNet (Hamp & Feldweg 1997)



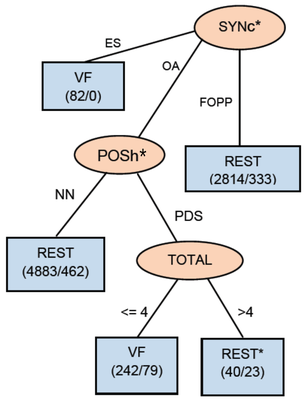
- Extraction of various features from sentence and constituent levels (Mousser & Zinsmeister 2009)
  - 28,102 declarative Verb Second clauses
  - 97,242 major constituents

Constituent-based classification

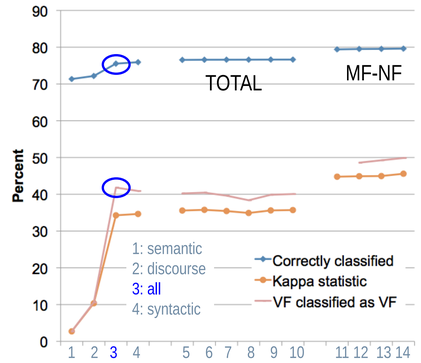
- Weka's (online) implementation of C4.5 decision tree classifier (J48)
  - binary classification: 'Vorfeld' versus 'rest'
  - automatically pruned trees; 10-fold cross validation
  - starting with all features; then systematic variation
  - example training instance (simplified):

LEXh	TOTAL	SYNc	POSh	LENGTHc	ANA-TYPE	ANA-POS	VLEX	VOICE	CLASSc
ein 'a / one'	3	OA accusative object	PIS subst. indef. pronoun	1 in tokens	instance instantiation of a class of entities	NN normal noun	erleben 'experience'	active	VF 'prefield'

- Decision tree (fragment):



Ten-fold cross validation:
Baseline "all as 'rest'": accuracy = 71.1%
w/ #TOTAL: accuracy = 76.6% ***
w/ #MF-NF: accuracy > 79.4%***



Feature groups

(adapted from Filippova & Strube 2007)

- Lexical features
  - the lemma of the root clause (VLEX)
  - the lemma of the head of the constituent (LEXh)
- Syntactic features
  - part of speech of the head of the constituent (POSh)
  - grammatical function of the constituent (SYNc or SYNc-REP with finer graded subject types)
  - length of a constituent in words (LENGTHc)
  - number of nodes between the maximal constituent level and the head of a constituent (DEPTHc)
  - whether the head of the constituent is modified by a relative clause (RELc)
  - the number of the modifiers of the head of the constituent (MODh)
  - the syntactic category of the constituent (CATc)
- Semantic features
  - voice of the verb (VOICE)
  - semantic class of the head of the constituent (SEMc)
- Discourse-related features
  - whether the head of the constituent appeared in the previous sentence (REP)
  - type of anaphoric/coreference relation (ANA-TYPE)
  - part of speech of the antecedent (ANA-POS), head of the antecedent (ANA-HEAD)
  - determiner type modifying the head of the constituent (DETc)
  - lexical form of the determiner (DETcform)

Sentence-based classification

- Most probable VF constituent per sentence
  - Perl 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
  - lower than results on Wikipedia biographies corpus (cf. Filippova & Strube 2007, Cheung & Penn 2009)
  - outperforms other studies on an earlier versions of the TüBa-D/Z 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 *expletive* 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 (Duden, chapter 2.3.4)
  - beginning of a text: subjective, temporal or local anchor
  - within a text: coreferential 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 optionality and variance of Vorfeld into account

References

- Bader M. & J. Häussler. 2010. Word order in German: a corpus study. *Lingua: 120*/3. 717-762
- Cheung, J. & G. Penn. 2009. Entity-based local coherence modelling using topological fields. In *Proceedings of the 48th ACL*,186–195.
- Duden Bd. 4. Die Grammatik. 7th edition. Mannheim a.o.: Dudenverlag.
- Filippova K. & M. Strube. 2007. Generating constituent order in German clauses. In *Proceedings of the 45th ACL*, 320–327.
- Hamp B. & H. Feldweg. 1997. GermaNet - a Lexical-Semantic Net for German. In *Proceedings of ACL workshop Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications.*
- Mousser J. & H. Zinsmeister (2009). Experiment: generating constituent order in German. Manuscript. University of Konstanz.
- Naumann, K. 2006. Manual of the annotation of in-document deferential relations. Technical Report. University of Tübingen.
- Schmid, H. 1995. Improvements in Part-of-Speech Tagging with an Application to German. *Proceedings of the ACL SIGDAT-Workshop.*
- Telljohann et al. 2009. Stylebook for the Tübingen Treebank of Written German (TüBa-D/Z). Technical Report. University of Tübingen.
- TüBa-D/Z (online) http://www.sfs.uni-tuebingen.de/tuebadz.shtml. [accessed, 03-03-2012]
- Weka (online) http://www.cs.waikato.ac.nz/ml/weka. [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.