Seminar für Sprachwissenschaft
EBERHARD KARLS UNIVERSITÄT TÜBINGEN

A Hybrid Model for Morpho-Syntactic Annotation of German with a Large Tagset

Julia S. Trushkina and Erhard W . Hinrichs

{jul,eh}@sfs.uni-tuebingen.de

Seminar für Sprachwissenschaft

Eberhard-Karls-Universität Tübingen

Morpho-syntactic Annotation:

Annotation of lexical tokens with part-of-speech and inflectional morphology (case, number, gender, and person)

Morpho-syntactic Annotation

Morpho-syntactic Annotation:

Annotation of lexical tokens with part-of-speech and inflectional morphology (case, number, gender, and person)

An Example from German:

Siege	*gaben*	*Spielern*	*Selbstvertrauen*.
Victories	gave	players	self-confidence
NN	VVFIN	NN	NN

Morpho-syntactic Annotation

Morpho-syntactic Annotation:

Annotation of lexical tokens with part-of-speech and inflectional morphology (case, number, gender, and person)

An Example from German:

Siege	*gaben*	*Spielern*	*Selbstvertrauen*.
Victories	gave	players	self-confidence
NN	VVFIN	NN	NN
[NGA;pl]	[1,3;pl]	[D;pl]	[NGDA;sg]

Ambiguity Rates Across Languages

language and source of the statistics		average # analyses	ambiguous tokens	tagset size
German	(current paper)	7.10	68.87%	718
Czech	Hajič & Hladka (1997)	3.65	not avail.	1171
	Hajič & Hladka (1997)	2.36	not avail.	882
Turkish	Oflazer & Tür (1996)	1.83	50.66%	not avail.
English	Tapanainen & Voutilainen (1994)	1.77	not avail.	139
German (STTS)	(current paper)	1.77	39.57%	54
Romanian	Tufiş (2000)	1.71	38.17%	410
Hungarian	Tufiş et al. (2000)	1.33	31.90%	$>$ 1265

Morpho-syntactic Tagging of German

- **Goal:** Automatic, morpho-syntactic annotation of German with a large tagset that can be effectively trained on manually annotated data of moderate size.

Morpho-syntactic Tagging of German

- **Goal:** Automatic, morpho-syntactic annotation of German with a large tagset that can be effectively trained on manually annotated data of moderate size.
- **Design of Extended Tagset:**

Morpho-syntactic Tagging of German

- **Goal:** Automatic, morpho-syntactic annotation of German with a large tagset that can be effectively trained on manually annotated data of moderate size.
- **Design of Extended Tagset:**
  - Basis: Stuttgart-Tübingen (STTS) tagset with 54 part-of-speech categories for German.

Morpho-syntactic Tagging of German

- **Goal:** Automatic, morpho-syntactic annotation of German with a large tagset that can be effectively trained on manually annotated data of moderate size.
- **Design of Extended Tagset:**
  - Basis: Stuttgart-Tübingen (STTS) tagset with 54 part-of-speech categories for German.
  - Enrichment of the STTS labels by morpho-syntactic features such as case, number, person, gender, tense and mood.

Seminar für Sprachwissenschaft
Eberhard Karls Universität Tübingen

Morpho-syntactic Tagging of German

- **Goal:** Automatic, morpho-syntactic annotation of German with a large tagset that can be effectively trained on manually annotated data of moderate size.
- **Design of Extended Tagset:**
  - Basis: Stuttgart-Tübingen (STTS) tagset with 54 part-of-speech categories for German.
  - Enrichment of the STTS labels by morpho-syntactic features such as case, number, person, gender, tense and mood.
  - Size of resulting tagset: 718 possible tags.

The Method

Cascaded, hybrid Architecture, consisting of:

The Method

Cascaded, hybrid Architecture, consisting of:

- a rule-based component with manually written disambiguation rules

The Method

Cascaded, hybrid Architecture, consisting of:

- a rule-based component with manually written disambiguation rules
- a statistical component trained on the `taz` newspaper portion of the TüBa-D treebank (taz, 1999)

Main Results

- The combined model outperforms the rule-based and statistical modules applied in isolation.

Seminar für Sprachwissenschaft
EBERHARD KARLS UNIVERSITÄT TÜBINGEN

Main Results

- The combined model outperforms the rule-based and statistical modules applied in isolation.
- The best result of the model attains an accuracy of 92.04%, which corresponds to a 7.34% improvement of the best results reported by other researchers for the same task. (Lezius et al. (1996)

The Data

- `taz` newspaper portion of the Tübingen Treebank of German (TüBa-D/Z) used in:

The Data

- `taz` newspaper portion of the Tübingen Treebank of German (TüBa-D/Z) used in:
  - statistical and combined model experiments

The Data

- `taz` newspaper portion of the Tübingen Treebank of German (TüBa-D/Z) used in:
  - statistical and combined model experiments
  - evaluation of all modules

The Data

- taz newspaper portion of the Tübingen Treebank of German (TüBa-D/Z) used in:
  - statistical and combined model experiments
  - evaluation of all modules
- 11 361 tokens for test data
- 5 891 tokens for development data
- 104 049 tokens were used as training data

The Data

- taz newspaper portion of the Tübingen Treebank of German (TüBa-D/Z) used in:
  - statistical and combined model experiments
  - evaluation of all modules
- 11 361 tokens for test data
- 5 891 tokens for development data
- 104 049 tokens were used as training data
- The statistical component uses additional 115 098 tokens with no morphological information for weakly supervised training.

The Data

- taz newspaper portion of the Tübingen Treebank of German (TüBa-D/Z) used in:
  - statistical and combined model experiments
  - evaluation of all modules
- 11 361 tokens for test data
- 5 891 tokens for development data
- 104 049 tokens were used as training data
- The statistical component uses additional 115 098 tokens with no morphological information for weakly supervised training.

Motivation for Combined Model

- Tapanainen and Voutilainen's Law:
  Don't guess if you know!

Motivation for Combined Model

- Tapanainen and Voutilainen's Law:
  Don't guess if you know!
- In other words:

  If you can state cautious disambiguation rules that do not compromise recall, then this is preferable to surrendering control to a statistical model.

Motivation for Combined Model

- Tapanainen and Voutilainen's Law:
  Don't guess if you know!
- In other words:

  If you can state cautious disambiguation rules that do not compromise recall, then this is preferable to surrendering control to a statistical model.
- Then use a statistical model to resolve any remaining ambiguity.

Rule-Based Module

- Initial set of analyses for the rule-based disambiguation module provided by the Xerox morphological analyzer.
- Rule-based disambiguation module developed in the Xerox Incremental Parsing System (XIP) platform.
- XIP provides two types of disambiguation rules:
  - Concord Rules
  - Syntactic heuristics

Evaluation of Rule-based Module

module	precision	recall	F-measure	LE	DE	ambiguity tokens	ambiguity rate
morph. analyzer	13.61%	96.64%	23.86%	100%	0%	68.76%	9.87
POS disamb.	19.93%	96.11%	33.01%	86.01%	13.99%	59.79%	7.39
morph. disamb.	42.53%	94.86%	58.73%	64.51%	35.49%	31.05%	4.96
+ adding analyses	46.93%	95.64%	62.97%	60.12%	39.88%	30.13%	4.44

Evaluation of Rule-based Module

module	precision	recall	F-measure	LE	DE	ambiguity tokens	ambiguity rate
morph. analyzer	13.61%	96.64%	23.86%	100%	0%	68.76%	9.87
POS disamb.	19.93%	96.11%	33.01%	86.01%	13.99%	59.79%	7.39
morph. disamb.	42.53%	94.86%	58.73%	64.51%	35.49%	31.05%	4.96
+ adding analyses	46.93%	95.64%	62.97%	60.12%	39.88%	30.13%	4.44

- lexical errors (LE): output of morphological analyzer does not contain the correct analysis

Evaluation of Rule-based Module

module	precision	recall	F-measure	LE	DE	ambiguity tokens	ambiguity rate
morph. analyzer	13.61%	96.64%	23.86%	100%	0%	68.76%	9.87
POS disamb.	19.93%	96.11%	33.01%	86.01%	13.99%	59.79%	7.39
morph. disamb.	42.53%	94.86%	58.73%	64.51%	35.49%	31.05%	4.96
+ adding analyses	46.93%	95.64%	62.97%	60.12%	39.88%	30.13%	4.44

- lexical errors (LE): output of morphological analyzer does not contain the correct analysis
- disambiguation errors (DE): due to overapplication of disambiguation rules

Statistical Tagging Module (1)

- uses tagging mode of the PCFG parser LoPar (Schmid 2000)
- each tag is computed by the following formula:

$$\arg\max_j \alpha_j(k,k) P(N^j \rightarrow w_k)$$

- i.e. best tag sequence: sequence of those tags that yield the maximal product of the inside and outside probabilities among the candidate tags for a given token

Standard Method in Tagging

- **Standard Method:** n-gram models (e.g. TnT tagger (Brants 2000))

Standard Method in Tagging

- **Standard Method:** n-gram models (e.g. TnT tagger (Brants 2000))

- **Inefficiency of n-gram models:** n-gram taggers consider only sequences of **n** words and their candidate tags, i.e. very local contexts, as the basis for determining the most likely sequence of tags for the sentence.

A Sample TNT Error

Die	*Frage*	*nach*	*der*	*Form*	*beantwortet*
The	question	about	the	form	answers
	[nom,acc]				

er	*so*:
he	in this way
[nom]	

‘He answers the question about the form in this way:’

Statistical Tagging Module (2)

- Training corpus:
  - 104 048 tokens from TüBa-D/Z treebank with transformed tree representations and full morphology
  - 115 098 partially labelled tokens from TüBa-D/Z treebank with transformed tree representations
- Tagger Lexicon:
  - based on TüBa-D/Z training data
  - enriched by 60 901 tokens from Negra corpus.

Evaluation of Statistical Module

module	precision	recall	F-measure	no tag	LE	DE
statistical	89.20%	88.10%	88.68%	1.23%	11.55%	88.45%

- no tag: due to lack of PCFG parse
- lexical errors (LE): correct tag is missing from the lexicon.
- disambiguation errors (DE): correct tag is eliminated.

Combined Model

- Rule-based module as pre-filter for PCFG module.
- Two experiments conducted for PCFG module
  - Experiment 1:
    All analyses left after application of the rule-based module are provided as input to the statistical module.
  - Experiment 2:
    Input to the statistical model limited to the categories that are most reliably tagged by the rule-based module.

Evaluation of Combined Model

experiments	precision	recall	F-measure	no tag	LE	RBE	SE
full input	90.23%	86.29%	88.22%	4.37%	0%	42.98%	57.02%
partial input	90.59%	87.67%	89.11%	3.23%	8.12%	19.54%	72.34%

Error Analysis of Combined Model

errors	POS	case	number	gender	person	tense	mood
SE	27.54%	40.51%	2.01%	10.03%	0.94%	0.00%	0.80%
RBE	62.87%	20.79%	1.49%	2.48%	0.00%	0.00%	0.00%
LE	70.24%	5.95%	3.57%	1.19%	0.00%	0.00%	0.00%
all	37.91%	33.85%	2.03%	7.83%	0.68%	0.00%	0.58%

Seminar für Sprachwissenschaft
EBERHARD KARLS UNIVERSITÄT TÜBINGEN

Final Experiment with Perfect Lexicon

- Problem with previous experiments: lexical errors in rule-based and statistical module, due to deficiencies of the morphological analyser or of the tagger lexicon.

Final Experiment with Perfect Lexicon

- Problem with previous experiments: lexical errors in rule-based and statistical module, due to deficiencies of the morphological analyser or of the tagger lexicon.
- To test viability of the approach as such: assume a perfect lexicon.

Final Experiment with Perfect Lexicon

- Problem with previous experiments: lexical errors in rule-based and statistical module, due to deficiencies of the morphological analyser or of the tagger lexicon.
- To test viability of the approach as such: assume a perfect lexicon.

experiments	precision	recall	F-measure	no tag	LE	RBE	SE
full input	92.04%	88.35%	90.16%	4.00%	0%	22.93%	77.07%
partial input	91.82%	89.02%	90.40%	3.05%	0%	10.21%	89.79%

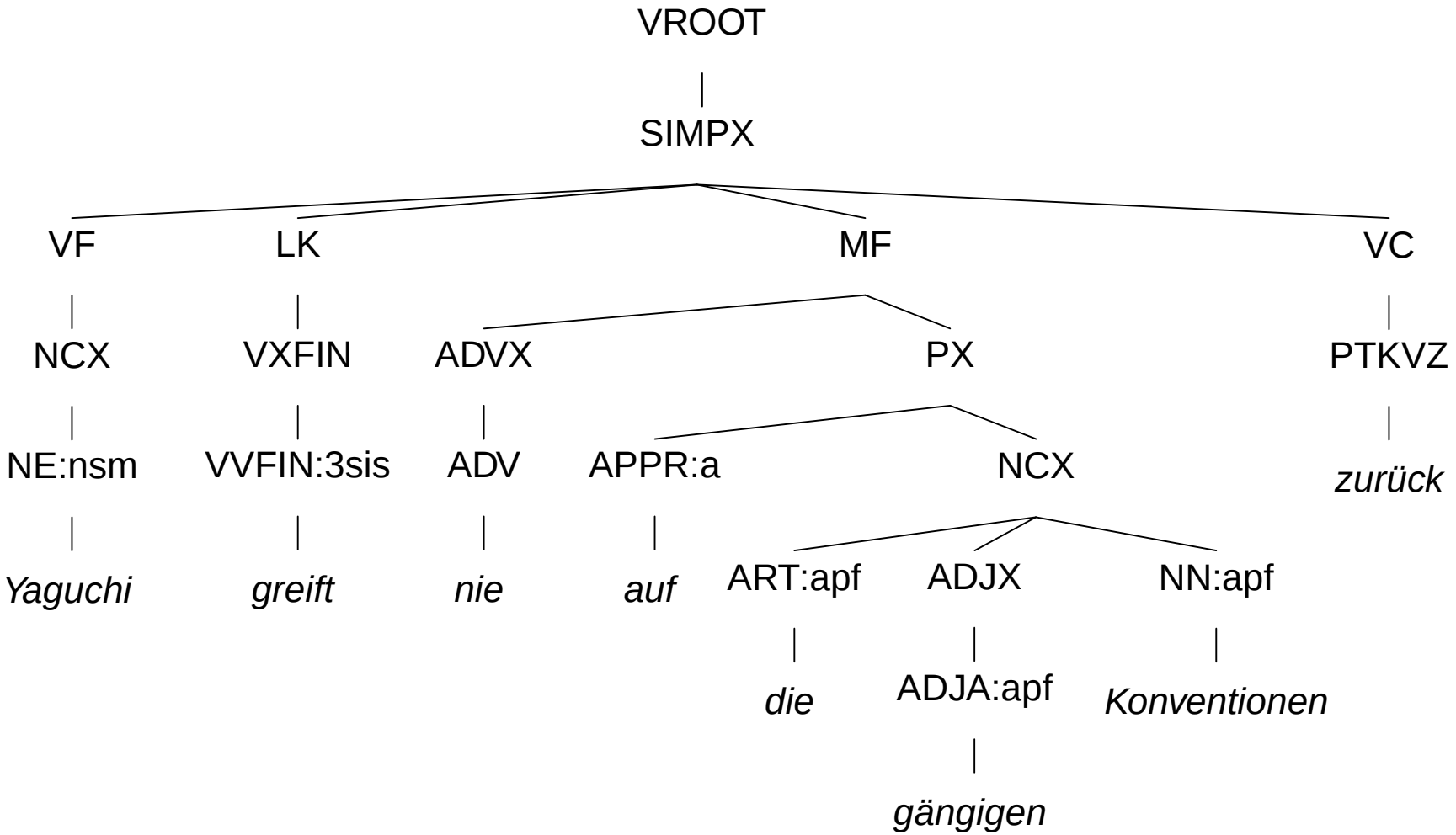
Main Results

- The combined model outperforms the rule-based and statistical modules applied in isolation.
- The best result of the model attains an accuracy of 92.04%, which corresponds to a 7.34% improvement of the best results reported by other researchers for the same task. (Lezius et al. (1996)

LoPar PCFG experiments

		precision	unparsed sent.
1.	baseline:	66.40	5
2.	back-up lexicon added	77.65	5
3.	topological fi elds deleted (except for VC and C)	77.58	5
4.	case passed up to NX and NCX	84.23	6
5.	gramm. functions (-ON, -OA) added & passed to SIMPX + rules binarized	84.98	5
6.	morph. info passed to NXs & VXFINs	87.62	7
7.	FIN label with number passed to SIMPX	88.21	9
8.	results on test data	87.69	11

A Sample Tree from TüBa-D treebank



An Example of a Transformed Tree