Seminar für Sprachwissenschaft
EBERHARD KARLS UNIVERSITÄT TÜBINGEN

Morphological Disambiguation of Noun Phrases in German

Erhard W. Hinrichs and Julia S. Trushkina

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

Seminar für Sprachwissenschaft

University of Tübingen

Germany

GRIP: GeRman Incremental Parsing

Seminar für Sprachwissenschaft
EBERHARD KARLS UNIVERSITÄT TÜBINGEN

Research Context:

Robust, automatic annotation for large text corpora of German with dependency relations.

Implementation Platform:

Xerox Incremental Deep Parsing System (XIP).

XIP's Parsing Architecture

Seminar für Sprachwissenschaft
EBERHARD KARLS UNIVERSITÄT TÜBINGEN

Incremental Approach:

- Morphosyntactic Annotation
  - morphological analysis
  - POS tagging
- Syntactic annotation
  - chunking and shallow parsing
  - assignment of dependency relations

XIP's Parsing Architecture

Seminar für Sprachwissenschaft
EBERHARD KARLS UNIVERSITÄT TÜBINGEN

Incremental Approach:

- Morphosyntactic Annotation
  - morphological analysis
  - POS tagging
- Syntactic annotation
  - chunking and shallow parsing
  - assignment of dependency relations

Morphological Disambiguation with GRIP

- **Main Claim:**

  Morphological disambiguation is a crucial step for assignment of dependency structures.

- **Main Result:**

  Unique morphological analyses for the assignment of dependency relations to German NPs in 77% of all cases.

- **Method:**

  Customized disambiguation rules of Xerox Incremental Deep Parsing System (XIP)

Seminar für Sprachwissenschaft
EBERHARD KARLS UNIVERSITÄT TÜBINGEN

An Example from German

Die Politiker gaben verdienten Beamten und
the politicians gave worthy civil servants and
Lohnempfängern ein höheres Gehalt.
wage recipients a higher salary

‘The politicians gave worthy civil servants and wage recipients a higher salary.’

Seminar für Sprachwissenschaft
EBERHARD KARLS UNIVERSITÄT TÜBINGEN

Chunk Analysis and Dependency Relations

{VF {NP#1 Die Politiker}} {LK#2 gaben}
{MF {NP#3 verdienten Beamten} und
{NP#4 Lohnempfängern}
{NP#5 ein höheres Gehalt}}.

SUBJ(#2,#1), OBJ_dir(#2,#5), OBJ_indir(#2,#3),
OBJ_indir(#2,#4)

A simplified CELEX entry for *geben*

geben +Acc_Comp+Dat_Comp+Nom_Subj

Massive Morphological Ambiguity

verdienten	Adj+Fem+Sg+DatGen+Wk
verdienten	Adj+Masc+Sg+AccGen+StWk
verdienten	Adj+Masc+Sg+Dat+Wk
verdienten	Adj+Neut+Sg+Gen+StWk
verdienten	Adj+Neut+Sg+Dat+Wk
verdienten	Adj+FMN+Pl+NomAccDatGen+Wk
verdienten	Adj+Masc+Pl+Dat+St
Beamten	Noun+Masc+Sg+AccGen+StWk
Beamten	Noun+Masc+Sg+Dat+Wk
Beamten	Noun+Masc+Pl+NomAccDatGen+Wk
Beamten	Noun+Masc+Pl+Dat+St

Seminar für Sprachwissenschaft
EBERHARD KARLS UNIVERSITÄT TÜBINGEN

Two Types of Disambiguation Rules

- Concord Rules
- Syntactic Heuristics

Seminar für Sprachwissenschaft
EBERHARD KARLS UNIVERSITÄT TÜBINGEN

Massive Morphological Ambiguity

verdienten	Adj+Fem+Sg+DatGen+Wk
verdienten	Adj+Masc+Sg+AccGen+StWk
verdienten	Adj+Masc+Sg+Dat+Wk
verdienten	Adj+Neut+Sg+Gen+StWk
verdienten	Adj+Neut+Sg+Dat+Wk
verdienten	Adj+FMN+Pl+NomAccDatGen+Wk
verdienten	Adj+Masc+Pl+Dat+St
Beamten	Noun+Masc+Sg+AccGen+StWk
Beamten	Noun+Masc+Sg+Dat+Wk
Beamten	Noun+Masc+Pl+NomAccDatGen+Wk
Beamten	Noun+Masc+Pl+Dat+St

Seminar für Sprachwissenschaft
EBERHARD KARLS UNIVERSITÄT TÜBINGEN

Massive Morphological Ambiguity

verdienten	Adj+Masc+Sg+AccGen+StWk
verdienten	Adj+Masc+Sg+Dat+Wk
verdienten	Adj+Masc+Pl+NomAccDatGen+Wk
verdienten	Adj+Masc+Pl+Dat+St
Beamten	Noun+Masc+Sg+AccGen+StWk
Beamten	Noun+Masc+Sg+Dat+Wk
Beamten	Noun+Masc+Pl+NomAccDatGen+Wk
Beamten	Noun+Masc+Pl+Dat+St

Seminar für Sprachwissenschaft
EBERHARD KARLS UNIVERSITÄT TÜBINGEN

Massive Morphological Ambiguity

verdienten Adj+Masc+Sg+AccGen+St

verdienten Adj+Masc+Pl+Dat+St

Beamten Noun+Masc+Sg+AccGen+St

Beamten Noun+Masc+Pl+Dat+St

Seminar für Sprachwissenschaft
EBERHARD KARLS UNIVERSITÄT TÜBINGEN

Residual Ambiguity

Die	Det+Def+Masc+Pl+NomAcc+St
Politiker	Noun+Masc+Pl+NomAcc
verdienten	Adj+Masc+Pl+Dat+St
	Adj+Masc+Sg+AccGen+St
Beamten	Noun+Masc+Pl+Dat+St
	Noun+Masc+Sg+AccGen+St
Lohnempfängern	Noun+Masc+Pl+Dat
ein	Det+Indef+Neut+Sg+NomAcc+Wk
höheres	Adj+Neut+Sg+NomAcc+St
Gehalt	Noun+Neut+Sg+NomAcc

Results of Morphological Disambiguation

	percentage
1 reading	58.65%
2 readings	34.31%
$\geq$ 3 readings	7.04%

1.55 readings per token

Seminar für Sprachwissenschaft
EBERHARD KARLS UNIVERSITÄT TÜBINGEN

Some Syntactic Heuristics (1)

- The ambiguous NP is the only candidate subject NP in a finite clause $\rightarrow$ Nom
- A noun with feature `City` or `Country` is preceded by a preposition *in* $\rightarrow$ Dat
- Eliminate Nom reading on ambiguous NPs if there is a non-ambiguous Nom NP in a clause $\rightarrow \neg$ Nom
- The NP is an argument of a copula (*sein*) $\rightarrow$ Nom

Some Syntactic Heuristics (2)

- A nominative reading does not agree with a finite verb in number $\rightarrow \neg$ Nom
- The NP is neither preceded by a preposition nor by another NP $\rightarrow \neg$ Gen
- The NP is a second (third) NP in a Vorfeld position in V2 clause $\rightarrow$ Gen
- The NP is a complement of a zu-infinitive $\rightarrow \neg$ Nom
- NP conjuncts agree in case

Resolving the Residual Ambiguity

Die	Det+Def+Masc+Pl+NomAcc+St
Politiker	Noun+Masc+Pl+NomAcc
verdienten	Adj+Masc+Pl+Dat+St
	Adj+Masc+Sg+AccGen+St
Beamten	Noun+Masc+Pl+Dat+St
	Noun+Masc+Sg+AccGen+St
Lohnempfängern	Noun+Masc+Pl+Dat
ein	Det+Indef+Neut+Sg+NomAcc+Wk
höheres	Adj+Neut+Sg+NomAcc+St
Gehalt	Noun+Neut+Sg+NomAcc

Resolving the Residual Ambiguity

Die	Det+Def+Masc+Pl+NomAcc+St
Politiker	Noun+Masc+Pl+NomAcc
verdienten	Adj+Masc+Pl+Dat+St
Beamten	Noun+Masc+Pl+Dat+St
Lohnempfängern	Noun+Masc+Pl+Dat
ein	Det+Indef+Neut+Sg+NomAcc+Wk
höheres	Adj+Neut+Sg+NomAcc+St
Gehalt	Noun+Neut+Sg+NomAcc

Resolving the Residual Ambiguity

Die	Det+Def+Masc+Pl+NomAcc+St
Politiker	Noun+Masc+Pl+NomAcc
verdienten	Adj+Masc+Pl+Dat+St
Beamten	Noun+Masc+Pl+Dat+St
Lohnempfängern	Noun+Masc+Pl+Dat
ein	Det+Indef+Neut+Sg+Acc+Wk
höheres	Adj+Neut+Sg+Acc+St
Gehalt	Noun+Neut+Sg+Acc

Resolving the Residual Ambiguity

Die	Det+Def+Masc+Pl+Nom+St
Politiker	Noun+Masc+Pl+Nom
verdienten	Adj+Masc+Pl+Dat+St
Beamten	Noun+Masc+Pl+Dat+St
Lohnempfängern	Noun+Masc+Pl+Dat
ein	Det+Indef+Neut+Sg+Acc+Wk
höheres	Adj+Neut+Sg+Acc+St
Gehalt	Noun+Neut+Sg+Acc

Disambiguation with Syntactic Heuristics

	count of NPs	percentage
1 reading	1211	77.08%
2 readings	226	14.39%
$\geq$ 3 readings	134	8.53%

1.2 readings per token

Seminar für Sprachwissenschaft
EBERHARD KARLS UNIVERSITÄT TÜBINGEN

Disambiguation for Non-single-element NPs

1 reading	82.33%
2 readings	17.18%
$\geq$ 3 readings	0.49%

Seminar für Sprachwissenschaft
EBERHARD KARLS UNIVERSITÄT TÜBINGEN

Precision & Recall for GRIP disambiguator

	NPs	NP lexical nodes
recall	76.26%	78.32%
precision	98.93%	99.02%

Seminar für Sprachwissenschaft
EBERHARD KARLS UNIVERSITÄT TÜBINGEN

An Alternative Approach

- Morphological Disambiguation by POS tagging

Seminar für Sprachwissenschaft
EBERHARD KARLS UNIVERSITÄT TÜBINGEN

An Alternative Approach

- Morphological Disambiguation by POS tagging
- Experiment with Brants' TNT Tagger

Seminar für Sprachwissenschaft
EBERHARD KARLS UNIVERSITÄT TÜBINGEN

An Alternative Approach

- Morphological Disambiguation by POS tagging
- Experiment with Brants' TNT Tagger
  - trained on 60 052 lexical tokens from the `taz` newspaper corpus, using two different tagsets:

An Alternative Approach

- Morphological Disambiguation by POS tagging
- Experiment with Brants' TNT Tagger
  - trained on 60 052 lexical tokens from the `taz` newspaper corpus, using two different tagsets:
  - STTS tagset (54 distinct tags)

An Alternative Approach

- Morphological Disambiguation by POS tagging
- Experiment with Brants' TNT Tagger
  - trained on 60 052 lexical tokens from the `taz` newspaper corpus, using two different tagsets:
  - STTS tagset (54 distinct tags)
  - STTS tags combined with morphological features for case, number, gender, tense, mood, and person (718 distinct tags)

Accuracy of TnT Tagger

	percentage
STTS tagset	93.39%
full tagset	70.78%
full tagset for NPs only	50.70%

Seminar für Sprachwissenschaft
EBERHARD KARLS UNIVERSITÄT TÜBINGEN

Error Analysis of TnT Tagger

	percentage
morphology only	81.98%
pos plus morphology	13.68%
pos only	4.34%

Seminar für Sprachwissenschaft
EBERHARD KARLS UNIVERSITÄT TÜBINGEN

Ordinary Disambiguation Rules

readings_filter = |left_context| selected_readings
|right_context|.

Seminar für Sprachwissenschaft
EBERHARD KARLS UNIVERSITÄT TÜBINGEN

Ordinary Disambiguation Rules

readings_filter = |left_context| selected_readings
|right_context|.

det,pron = det |adj*, noun|.

Seminar für Sprachwissenschaft
EBERHARD KARLS UNIVERSITÄT TÜBINGEN

Double Reduction Rules

$|\text{node\_sequence}| \Rightarrow \text{boolean\_constraints}.$

Seminar für Sprachwissenschaft
EBERHARD KARLS UNIVERSITÄT TÜBINGEN

Double Reduction Rules

$|\text{node\_sequence}| \Rightarrow \text{boolean\_constraints}.$

$|\text{adj\#1, noun\#2}| \Rightarrow \#1[\text{agr}] :: \#2[\text{agr}].$

Seminar für Sprachwissenschaft
EBERHARD KARLS UNIVERSITÄT TÜBINGEN

Double Reduction Rules

$|\text{node\_sequence}| \Rightarrow \text{boolean\_constraints}.$

$|\text{adj\#1, noun\#2}| \Rightarrow \#1[\text{agr}] :: \#2[\text{agr}].$

$|\text{adj*, adj\#1, adj*, noun\#2}| \Rightarrow (\#1[\text{agr}] :: \#2[\text{agr}]).$

Double Reduction Rules

$|\text{node\_sequence}| \Rightarrow \text{boolean\_constraints}.$

$|\text{adj\#1, noun\#2}| \Rightarrow \#1[\text{agr}] :: \#2[\text{agr}].$

$|\text{adj*, adj\#1, adj*, noun\#2}| \Rightarrow (\#1[\text{agr}] :: \#2[\text{agr}]).$

$|\text{det\#1, adj*, adj\#2, adj*, noun}| \Rightarrow$
$(\#1[\text{agr}] :: \#2[\text{agr}]) \;\&\; (\#1[\text{decl}] \sim: \#2[\text{decl}]).$

Seminar für Sprachwissenschaft
EBERHARD KARLS UNIVERSITÄT TÜBINGEN

Double Reduction Rules

$|\text{node\_sequence}| \Rightarrow \text{boolean\_constraints}.$

$|\text{adj\#1, noun\#2}| \Rightarrow \#1[\text{agr}] :: \#2[\text{agr}].$

$|\text{adj*, adj\#1, adj*, noun\#2}| \Rightarrow (\#1[\text{agr}] :: \#2[\text{agr}]).$

$|\text{det\#1, adj*, adj\#2, adj*, noun}| \Rightarrow$
$(\#1[\text{agr}] :: \#2[\text{agr}]) \;\&\; (\#1[\text{decl}] \sim: \#2[\text{decl}]).$

$|?[\text{det}:\sim], \text{adj*, adj\#1, adj*, noun\#2}| \Rightarrow$
$(\#1[\text{agr}] :: \#2[\text{agr}]) \;\&\; (\#1[\text{decl}: \text{St}]) \;\&$
$(\#2[\text{decl}: \text{St}]).$