Automatic Generation of MultiNet Representations for German Text Corpora

Sven Hartrumpf

Joint work with
Ingo Glöckner, Hermann Helbig, Johannes Leveling, Rainer Osswald

Intelligent Information and Communication Systems (IICS)
http://pi7.fernuni-hagen.de/
FernUniversität in Hagen
58084 Hagen, Germany

2007-04-02, Prague
Overview

1 Motivation

2 Target Representation

3 Parsing

4 Disambiguation

5 Coreference Resolution

6 Semantic Assimilation

7 Applications

8 Conclusion and Future Work
1 Motivation

Many success stories for NLP applications relying on shallow methods only

But: there are many more NLP applications which call for deep methods

Here: deep semantic analysis of large text corpora

Many steps and methods spanning the range:

real-world text corpora
↓
formal semantic representations and
↓
their successful exploitation in applications
2 Target Representation

The semantic network formalism MultiNet (Helbig, 2006, 2001) provides means for homogenously representing the semantics of:

- words (also: multi-word expressions)
- phrases
- sentences
- texts
- text corpora
Target Representation (contd.)

Figure 1: Small multi-net (= semantic network of the MultiNet formalism) example:
Die Mutter gab ihrem Sohn ein neues T-Shirt.
The mother gave her son a new t-shirt.
3 Parsing

Syntactic-semantic parser (WOCADI, Helbig and Hartrumpf (1997); Hartrumpf (2003)) forms the center of automatic MultiNet construction

3.1 Morpholexical Analysis

Based on different lexica:

- deep lexicon (HaGenLex, Hartrumpf, Helbig, and Osswald (2003)):
  full information on orthography, morphology, syntactic-semantic valency
  26,000 entries: 13,500 nouns, 7,700 verbs, 3,300 adjectives

- flat lexicon:
  no semantics, no valency frames; sources: CELEX, UMLS
  47,500 additional entries

- name lexica: around 100 different lists with 500,000 entries
  for 46 different concept classes like city, country, region
  (some manually maintained; some extracted from Wikipedia etc.)
Figure 2: Abbreviated HaGenLex entry for the verb geben.1.2 used in Figure 1
Morpholexical Analysis (contd.)

Morphology: derived from CELEX

Compound analysis: semantically oriented

Compounds are very productive in German

For better disambiguation: a compound lexicon with 30,000 entries

Specifying only segmentation and semantic class of the compound

Example entries:
("Staubecke.1.1" ("stau.1.1" "ecke.1.1") assoc-compd "SH 2006-06-28")
("Staubecken.1.1" ("stauen.1.1" "becken.1.1") purp-compd "SH 2006-06-28")

Staubecken: dust corners or reservoir

Semi-automatically maintained
3.2 WCFA

Word Class Function Analysis (WCFA, Helbig (1994, 1986))

No clean separation of parser and grammar; not so fine for maintenance?

Yes, to some extent; but the WCFA parser is

- an *experienced workhorse*
- easily extendible, e.g. by disambiguation modules

New implementation in the Scheme programming language:
WOrd CIAss based DIsambiguating (WOCADI, Helbig and Hartrumpf (1997); Hartrumpf (2003))
**WCFA Characteristics**

For each word class (similar to PoS): a word class function (WCF) is implemented.

WCFs together with a central control mechanism (WCF machine) model two phases of word activity:

- **Opening act**: opening expectations about what might or should follow.
- **Completing act**: saturating expectations by connecting complements/adjuncts to heads etc.

→ Some similarity to chart parsing.

Oriented towards constructing multi-nets; useful by-product: dependency trees.

Emphasis on understanding acceptable sentences, not on perfectly grammatical sentences.

WCFA analyzes sentences on four levels:

- **Level 1**: elementary kernel (e.g. simple NP)
- **Level 2**: complex kernel (e.g. complex NP)
- **Level 3**: simple proposition (phrase containing only one main verb, i.e. a clause)
- **Level 4**: complex proposition (i.e. a complex sentence)
Example of a WCF: Simplified opening act of the WCF

`ADV`

```lisp
next-word-cat ← (fs-value *next-word '(syn cat))
adv-sort ← (fs-value *top '(semsel sem entity sort))
(follow-alternatives ; Each top-level if-clauses can lead to a parse alternative.
  if (and (equal? next-word-cat 'adv)
    (not (member? adv-sort '(l t)))))
    ; The adverb modifies another adverb.
    (fs-set-value *top '(open-cat) 'advp) ; expect an adv phrase
  fi
  if (and (equal? next-word-cat 'a)
    (not (member? adv-sort '(ng))))
    ; The adverb seems to modify an adjective.
    (fs-set-value *top '(open-cat) 'ap) ; expect an adjective phrase
  if (and (fs-value=? *top2 '(open-cat) 'advp)
    (fs-value=? *top2 '(complete-cat) 'cat)) ; inside an open advp
    (set-status 'completing)
  elif (not (type-subsumed-d? adv-sort '(gr ql t))) ; adv is not a graduator
    (set-status 'closed)
    (fs-set-value *top '(open-cat) 'advp)
    (fs-set-value *top '(complete-cat) 'advp)
  fi
  fi
```

...
3.3 MultiNet Construction in the Parser

Example multi-net for the CLEF corpus sentence:
Nachdem die Kommunisten ihr Machtmonopol aufgegeben hatten, wechselte Karadzic in die Politik und gründete 1990 die Serbische Demokratische Partei (SDS).
After the communists gave up their power monopoly, Karadzic moved to politics and founded in 1990 the Serbian Democratic Party (SDS).
Relational Layer

Where do the multi-net parts come from?

1. WCFs
   e.g. PROP, PROPR, *PMOD from WCF a(djective)

2. Valency frame of verbs, nouns, and adjectives
   Each argument specifies one or more relations leading from the mother node to the argument node
   Set-valued feature semsel|select|N|REL
   e.g. verb geben.1.2 (give; from Figure 1):
   AGT, ORNT, OBJ

3. Feature semsel|sem|net of verbs, nouns, and adjectives
   Used if simple relations do not suffice for connecting arguments
   e.g. if the semantics consists of relations connecting the arguments directly
   e.g. verb bestehen.1.2 (The table consists of wood.): net (origm x1 x2)
   A metaverb because it is the NL expression for a MultiNet primitive

4. Feature semsel|sem|net of prepositions (similarly for conjunctions)
   e.g. the shot in the morning versus the shot in the darkness
   Can also be seen as PP interpretation rules (see Hartrumpf, Helbig, and Osswald (2006))
Relational Layer (contd.)

Other HaGenLex features with multi-net content:

- **ENT-NET**: filled automatically from semi-formal entailments
- **EQU-NET**: equivalences (special case of ENT-NET)
- **LEX-NET**: lexical relations to other lexical concepts (SYNO, ANTO, etc.)
- **PRE-NET**: presuppositions (planned)
- **SEL-NET**: selectional restrictions (planned)
Layer Features of Nodes

MultiNet nodes have layer features depending upon their sort:

Intensional layer features:
- gener: genericity (generic vs. specific)
- quant: quantification
- refer: referentiality (determinate, indeterminate)
- varia: variable reference

Extensional layer features:
- card: cardinality
- etype: extension type (individual, set, set of sets)
- fact: facticity (real, hypothetical, nonreal)
Layer Features of Nodes (contd.)

Sources:

- Lexicon
  - esp. entries for determiners and quantifiers
  - nouns: count nouns versus group nouns

- WCFs
  - e.g. WCF n(oun): ETYPE modification depending upon NUM(erus)

- PP interpretation rules: CARD, ETYPE

- **Layer unification principle**: enforced in WCFs participating in NPs (Hartrumpf and Helbig, 2002)

But: some postprocessing needed (e.g. generic uses not handled in parser)
MWR (MultiNet Workbench: layer module)
Layer Features of Relations

Each MultiNet edge characterized by knowledge type (k-type) at start node and end node

\[ ktype \rightarrow descr \rightarrow imman \rightarrow categ \n\]
\[ restr \rightarrow situa \rightarrow proto \n\]

Sources:

- Defaults for each relation (Helbig, 2006)
- Refinement in some WCFs, mainly refinements to categ(orical)
- Explicit k-types in PP interpretation rules

But: some postprocessing needed \(\rightarrow\) MWR layer module
3.4 Evaluation of the Parser

<table>
<thead>
<tr>
<th>Corpus</th>
<th># Texts</th>
<th># Sentences</th>
<th># Words</th>
</tr>
</thead>
<tbody>
<tr>
<td>TüBa-D/Z (Tübingen Treebank of Written German)</td>
<td>27,125</td>
<td></td>
<td></td>
</tr>
<tr>
<td>HaGenLex examples</td>
<td>24,000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CLEF (Cross-Language Evaluation Forum)</td>
<td>277,000</td>
<td>4,900,000</td>
<td>90,000,000</td>
</tr>
<tr>
<td>Wikipedia (German snapshot from 2006-09-25)</td>
<td>475,000</td>
<td>11,000,000</td>
<td>198,000,000</td>
</tr>
</tbody>
</table>

Coverage: around 60% for correct sentences (HaGenLex: near 100%)

Why not higher?

- Not only syntax, but also semantics
- Coverage rates multiply: e.g. $0.8[\text{syntactic coverage}] \cdot 0.8[\text{semantic coverage}] = 0.64$
- Main problems:
  - Verb readings missing in the lexicon
  - Complex sentences comprising many clauses
  - Analysis depth
  - Corpus-specific encoding of special characters like quotation marks

Accuracy: only investigated on small samples
But: the need to construct full semantics is often a good filter/disambiguator

Speed: depends on parser parameter *maximum number of WCF calls* $\rightarrow$ coverage
typically: 1–2 seconds per sentence on current PCs
4 Disambiguation

WOCADI parser uses specialized disambiguation methods:

- prepositional phrase (PP) attachment and interpretation
- coreference resolution
- word sense disambiguation (WSD; in the context of PPs)

4.1 General Hybrid Disambiguation Approach

**hybrid** method *(Hartrumpf, 2003)*

1. linguistic rules ← manual

2. disambiguation statistics ← automatic (from annotated corpora)
   new statistical model: multidimensional back-off
Multidimensional Back-off Models

$R \in \mathbb{R}$ representation of an ambiguity problem (instance) as a list of alternatives

$\mathbb{R}$ set of all theoretically possible ambiguity problems

Training Corpus:

(1) $\mathbf{R}_{\text{train}}$: list of elements $(R_j, i_j)$ with $R_j \in \mathbb{R}, i_j \in \{0, 1, \ldots, |R_j| - 1\}$

Initial counts:

(2) $C(R_j, i_j) :=$ absolute frequency of $(R_j, i_j)$ in $\mathbf{R}_{\text{train}}$

Abstraction function:

(3) $a : \mathbb{L} \times \mathbb{R} \rightarrow \mathbb{A}$
(4) $a(l, R) = A_l$ with $l \in \mathbb{L}, r \in \mathbb{R}, A_l \in \mathbb{A}$
Abstraction function from dimension abstraction functions:

(5) \[ a(l, R) := a_0(l, R, a_1, a_2, \ldots, a_d) \]

(6) \[ a_i : L_i \times R_i \rightarrow A_i, L_i := \{\text{min}l_i, \text{min}l_i + 1, \ldots, \text{max}l_i\} \text{ for } 1 \leq i \leq d \]

Back-off levels:

(7) \[ L := L_1 \times L_2 \times \ldots L_d \]
New compared to standard back-off models: **detail back-off** vs. **alternative back-off** as two dimensions

Example abstractions from coreference resolution

Detail back-off with three levels: \( L_1 = \{1, 2, 3\} \)
Alternative back-off with two levels: \( L_2 = \{1, 2\} \)

<table>
<thead>
<tr>
<th>Level</th>
<th>Abstraction</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1, 1)</td>
<td>((ident.n_perspro, entity=substance, case=nom), (ident.n_perspro, entity=substance, case=acc), (ident.perspro_perspro, entity=object, case=nom))</td>
</tr>
<tr>
<td>(1, 2)</td>
<td>((ident.n_perspro, entity=substance), (ident.n_perspro, entity=substance), (ident.perspro_perspro, entity=object))</td>
</tr>
<tr>
<td>(1, 3)</td>
<td>((ident.n_perspro, entity=con-object), (ident.n_perspro, entity=con-object), (ident.perspro_perspro, entity=object))</td>
</tr>
<tr>
<td>(2, 1)</td>
<td>((ident.n_perspro, entity=substance, case=nom), (ident.n_perspro, entity=substance, case=acc))</td>
</tr>
<tr>
<td>(2, 2)</td>
<td>((ident.n_perspro, entity=substance), (ident.n_perspro, entity=substance))</td>
</tr>
<tr>
<td>(2, 3)</td>
<td>((ident.n_perspro, entity=con-object), (ident.n_perspro, entity=con-object))</td>
</tr>
</tbody>
</table>
Learning phase: $|L|$ statistical models by multidimensional back-off

Back-off direction function:

(8) $b : \{0, 1, \ldots, |L| - 1\} \rightarrow L$

(9) $\forall j, k \in \{0, 1, \ldots, |L| - 1\} (j \neq k \rightarrow b(j) \neq b(k))$

Statistical models:

(10) $C_l(A, i) := \sum_{R \in R \text{ with } a(l, R) = A \land s(l, R) = false} C(R, i) + acc(l, R, i)$

with $l \in L$, $A \in A$, $0 \leq i < \text{Zahl der Alternativen in } A$

$acc$ (accumulation): also count problems with additional alternative
Application phase: find relevant model, relative frequencies, probabilities

Probabilities for first level $l_0$:

\begin{equation}
 f(l_0, R, i) := \frac{C_{l_0}(a(l_0, R), i)}{\sum_{j=0}^{n} C_{l_0}(a(l_0, R), j)} \quad \text{with } i \in \{0, 1, \ldots, n\}
\end{equation}

\begin{equation}
 p(i|R) \approx f(l_0, R, i)
\end{equation}

Stopping criterion: absolute threshold $t_a$

\begin{equation}
 c(l, R) := \begin{cases} 
 true & \text{if } \exists i \in \{0, \ldots, m - 1\} (C_l(a(l, R), i) \geq t_a) \\
 false & \text{otherwise}
\end{cases}
\end{equation}

$m$: number of alternatives in $a(l, R)$

instead (or in addition): relative threshold
Back-off to next model $C_{l_k}$:

\[ f(l_k, R, i) := \frac{C_{l_k}(A, i)}{\sum_{j=0}^{m-1} C_{l_k}(A, j)} \]

with $A = a(l_k, R)$, $m = \text{number of alternatives in } A, i \in \{0, \ldots, m - 1\}$

\[ p(i|R) \approx f(l_k, R, i) \]
5 Coreference Resolution

WOCADI parses a text sentence by sentence

Resolves only absolutely necessary coreferences (involving relative pronouns or relative determiners)

Coreference resolution afterwards by the coreference module (CORUDIS):
determines antecedent of each mention (markable) as defined in MUC

Leads to additional EQU relations
between nodes of sentence multi-nets or
between nodes of a single multi-net

Trained and evaluated on a corpus annotated by us: GeCCo (German Coreference Corpus)
6 Semantic Assimilation

Further aspects for semantics of a whole text:
during assimilation as a post-interpretation step after parsing (refinement components)

Rationale: delay all semantic and pragmatic aspects that need
discourse model, dialog model, or domain knowledge

Only some parts implemented:

- Turning the EQU relations from CORUDIS into modified multi-nets:
  only one node for all mentions in a coreference chain
- Deixis resolution: temporal part implemented
  40 MultiNet rules
  Successfully employed for question answering (CLEF 2006)
- Intertextual coreference resolution (e.g. event tracking): planned
7 Applications

Implemented applications that build on deep semantic representations:


- Information retrieval (IR): CLEF track GIRT (German Indexing and Retrieval Test database) parsing documents/queries (Leveling, 2006b)

- Geographic information retrieval (GIR): CLEF track GeoCLEF parsing documents/queries; metonymy recognition for location names (Leveling and Veiel, 2006)

- Natural language interfaces (NLIs): for bibliographic databases (Z39.50) and SQL databases (Leveling, 2006a)

- Measuring and criticizing readability of NL texts (EU project *Benchmarking Tools and Methods for the Web*, BenToWeb) e.g. ambiguous pronoun references, distant antecedents, semantic complexity (Jenge, Hartrumpf, Helbig, and Osswald, 2006; Hartrumpf et al., 2006)

- Support for semi-automatic knowledge assimilation
Semantics at Work: Details from Question Answering

Question (variant of CLEF question qa06.079):
In welchem Jahr verschied Charles de Gaulle?
In which year did Charles de Gaulle pass away?

Document sentence (SDA.951109.0236.88):
France’s chief of state Jacques Chirac acknowledged the merits of general and statesman Charles de Gaulle, who died 25 years ago.

Answer from InSicht:
im Jahr 1970
in 1970

Found by comparing semantic networks and inferencing:

◊ Related verb readings are identified by lexical relations from HaGenLex:
verscheiden.1.1 and sterben.1.1

◊ Temporal deixis is resolved:
vor 25 Jahren and date of publication (from article start or article metadata)
8 Conclusion and Future Work

Work at the FernUniversität in Hagen (IICS):

- Deep semantic approach to NL understanding
- MultiNet as integrating language across all linguistic levels
- Implemented for use on large corpora
- Basis for NLP applications, e.g. successful question answering system for QA@CLEF

Plans for the future:

- Extension of analysis to English:
  - English lexicon is being constructed in parallel to German (HaGenLex)
  - WCFs are being parametrized for English
- German: extend coverage
- Further parts of semantic assimilation, e.g. intertextual coreferences (event and entity tracking)

Questions?
References

Hartrumpf, Sven (2003). *Hybrid Disambiguation in Natural Language Analysis*. Osnabrück, Germany: Der Andere Verlag. 5, 8, 21


Leveling, Johannes (2006a). *Formale Interpretation von Nutzeranfragen für natürlichsprachliche Interfaces zu Informationsangeboten im Internet*. Der andere Verlag, Tönning, Germany.
