Multi-label Classification of Semantic Relations in German Nominal Compounds using SVMs

Daniil Sorokin, Corina Dima and Erhard Hinrichs
CCLCC Workshop @ ESSLLI 2014

Thanks to:
Our colleagues: Verena Henrich and Christina Hoppermann
Our student annotators: Kathrin Adlung, Nadine Balbach and Tabea Sanwald
What is a Nominal Compound?

“A nominal compound is a sequence of nouns which function as a single noun.”

- *On the creation and use of the English compound nouns*, Language 53(4)
  Downing, 1977

Holzbank ‘wooden bench’
Bierglas ‘beer glass’
Fischsuppe ‘fish soup’
Muttersprache ‘native language/mother tongue’ lit. ‘mother language’
Interpreting Noun Compounds

Steinhaus
‘石屋’

Stein ‘stone’

Haus ‘house’

R(x,y)

modifier

head
Specifying compound-internal relations

Möbelhaus
‘furniture store’

Modehaus
‘fashion house’

Steinhaus
‘stone house’

Schneehaus
‘igloo’ lit. ‘snow house’

Holzhaus
‘wood[en] house’

Baumhaus
‘tree house’

Landhaus
‘country house’

Eckhaus
‘corner house’

Autohaus
‘car dealership’

Geburtshaus
‘birth house’

Konzerthaus
‘concert hall’

Auktionshaus
‘auction house’

Gästehaus
‘guest house’

Armenhaus
‘poorhouse’

Waisenhaus
‘orphanage’
Specifying compound-internal relations

Möbelhaus
‘furniture store’

Autohaus
‘car dealership’

Geburtshaus
‘birth house’

Modehaus
‘fashion house’

Konzerthaus
‘concert hall’

Schneehaus
‘igloo’ lit. ‘snow house’

Auktionshaus
‘auction house’

Steinhaus
‘stone house’

Gästehaus
‘guest house’

Holzhaus
‘wood[en] house’

Waisenhaus
‘orphanage’

Baumhaus
‘tree house’

Armenghaus
‘poorhouse’

Landhaus
‘country house’

Eckhaus
‘corner house’

Haus
‘house’

für
‘for’

aus
‘from’

an
‘on’

auf
‘in’

in
‘in’
Specifying compound-internal relations

- **Möbelhaus** 'furniture store'
- **Autohaus** 'car dealership'
- **Modehaus** 'fashion house'
- **Geburtshaus** 'birth house'
- **Konzerthaus** 'concert hall'
- **Auktionshaus** 'auction house'
- **Schneehaus** 'igloo' lit. 'snow house'
- **Steinhaus** 'stone house'
- **Auktionshaus** 'auction house'
- **Holzhaus** 'wood[en] house'
- **Solidhaus** 'solid house'
- **Baumhaus** 'tree house'
- **Landhaus** 'country house'
- **Eckhaus** 'corner house'
- **Armenhaus** 'poorhouse'
- **Gästehaus** 'guest house'
- **Waisenhaus** 'orphanage'

**goods** für 'for'

**usage** für 'for'

**material** aus 'from'

**location** in 'in'/ auf 'in'/ an 'on'
Compound Dataset

- per head annotation, started with a list of concrete nouns
- (already split!) compounds (Henrich and Hinrichs, 2011) from the German wordnet GermaNet, release 9

Dataset statistics:
- 5082 compounds (4607 used in the experiments)
- 360 distinct heads
- 2171 distinct modifiers
- 12.8 compounds per head (avg.)
- 2.1 compounds per modifier (avg.)

Annotation scheme (for German)
- 38 semantic properties
- 17 prepositions
- 76.4% agreement - property annotation
- 79.5% agreement - preposition annotation
- 68.6% agreement - property and preposition annotation

for more details see (Dima et al., LREC 2014)
**Compound modelling**

**Premise:** The meaning of a compound is derived from the meaning of its constituents and the world knowledge as to how these constituents can interact.

*Baumhaus* ‘tree house’

Therefore: model individual constituents and predict the possible relation based on their combination.

\[
\begin{align*}
&\text{[Baum]} \quad , \quad \text{[Haus]} \quad , \quad \text{[Baum] + [Haus]} \\
&\text{individually} \quad \text{individually} \quad \text{in conjunction}
\end{align*}
\]
Compound modelling

**Corpus-based features**

Web-news corpus
(1.6 billion tokens)

**Knowledge-based features**

German wordnet GermaNet
(Literals: 110738)
Compound modelling

Corpus-based features

\[
\left[ \underline{\text{Baum}}, \quad \underline{\text{Haus}}, \quad \underline{\text{Baum} + \text{Haus}} \right] \\
\text{individually} \quad \text{individually} \quad \text{in conjunction}
\]

Co-occurrence distributions of the target word.

- 1000 most frequent German words (lemmas)
- 17 prepositions

\[ \cdots \cdots \cdots \underline{\text{Baum}} \cdots \cdots \cdots \]

3 tokens

3 tokens
Compound modelling

Corpus-based features

\[
\left[ \begin{array}{c}
[\text{Baum}], \quad [\text{Haus}], \\
\text{individually, individually}
\end{array} \right]
\begin{array}{c}
[Baum] + [Haus] \\
\text{in conjunction}
\end{array}
\]

Co-occurrence distributions of the pair of target words (words should appear together in one sentence).

- 1000 most frequent German words (lemmas)
- 17 prepositions
Compound modelling

Knowledge-based features

\[
\begin{bmatrix}
[Baum] & , & [Haus] & , & [Baum] + [Haus] \\
\text{individually} & \text{individually} & \text{in conjunction}
\end{bmatrix}
\]

• Beginner category

House → Artifact

• Binary indicators for hypernyms

House → abode, construction, building

• Gloss terms (binary indicators for 1000 most frequent German words)

House → A structure serving as an abode of human beings.
Compound modelling

Knowledge-based features

\[
\left[ \begin{array}{c}
\text{[Baum]} \quad , \quad \text{[Haus]}
\end{array} \right] \quad \text{individually individually}
\]

\[
\left[ \begin{array}{c}
\text{[Baum]} + \text{[Haus]}
\end{array} \right] \quad \text{in conjunction}
\]

- Binary indicators for relations between constituents
  - Synonymy, Hyponymy, Meronymy, etc.
- *Hirst-St.Onge* relatedness measure
  - Numeric value
- Beginner category of the least common subsumer of the two constituents
  - Artifact, Substance, etc.
Compound modelling

6943-dimensional vector

- Corpus-based: 43.9% (3054)
- Knowledge-based: 56.1% (3889)

- [Baum]: 42.6% (2958)
- [Haus]: 42.6% (2958)
- [Baum] + [Haus]: 14.79% (1027)

- modifier individually
- head individually
- constituents in conjunction
### Experimental setup

<table>
<thead>
<tr>
<th>Single-label learning</th>
<th>Multi-label learning</th>
</tr>
</thead>
<tbody>
<tr>
<td>Predict preposition and property <strong>in isolation</strong></td>
<td>Predict preposition and property <strong>simultaneously</strong></td>
</tr>
<tr>
<td>Tools: SVM, Weka implementation with 10-fold cross validation</td>
<td>Tools: Mulan multi-label learning library with 10-fold cross validation</td>
</tr>
</tbody>
</table>
## Results

### Preposition prediction

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Accuracy</th>
<th>F-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline (most freq.)</td>
<td>30.09%</td>
<td>0.182</td>
</tr>
<tr>
<td>SVM</td>
<td>62.32%</td>
<td>0.616</td>
</tr>
<tr>
<td>Lauer (1995) EN</td>
<td>47.00%</td>
<td>—</td>
</tr>
</tbody>
</table>

### Single-label learning

**Classifier**

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline (most freq.)</td>
<td>30.09%</td>
</tr>
</tbody>
</table>

### Multi-label learning

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVM</td>
<td>62.32%</td>
</tr>
<tr>
<td>Lauer (1995) EN</td>
<td>47.00%</td>
</tr>
</tbody>
</table>
## Results

### Single-label learning

### Multi-label learning

#### Semantic property prediction

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Accuracy</th>
<th>F-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline (most freq.)</td>
<td>23.30%</td>
<td>0.084</td>
</tr>
<tr>
<td><strong>SVM</strong></td>
<td><strong>60.74%</strong></td>
<td><strong>0.601</strong></td>
</tr>
<tr>
<td>Tratz and Hovy (2010) EN</td>
<td>79.30%</td>
<td>—</td>
</tr>
<tr>
<td>Ó Séaghdha (2013) EN</td>
<td>65.40%</td>
<td>—</td>
</tr>
<tr>
<td>Verhoeven (2012) NL</td>
<td>47.80%</td>
<td>0.490</td>
</tr>
</tbody>
</table>
## Results

<table>
<thead>
<tr>
<th>Classifier</th>
<th>F-score preposition</th>
<th>F-score sem. property</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline (most freq.)</td>
<td>0.182</td>
<td>0.084</td>
</tr>
<tr>
<td>Single-label preposition</td>
<td>0.616</td>
<td>—</td>
</tr>
<tr>
<td>Single-label property</td>
<td>—</td>
<td>0.601</td>
</tr>
<tr>
<td><strong>Multi-label classifier</strong></td>
<td><strong>0.639</strong></td>
<td><strong>0.601</strong></td>
</tr>
</tbody>
</table>
# Results

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Combined label accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline (most freq.)</td>
<td>22.66%</td>
</tr>
<tr>
<td>IAA</td>
<td>68.60%</td>
</tr>
<tr>
<td>Single-label preposition + single-label property</td>
<td>48.44%</td>
</tr>
<tr>
<td>Multi-label classifier</td>
<td><strong>59.61%</strong></td>
</tr>
</tbody>
</table>

**Single-label learning**

**Multi-label learning**
Conclusions

• The first results for the task of automatically classifying German compounds
• The first hybrid, multi-label annotation scheme for compounds
• The hybrid annotation scheme provides more semantic information and improves automatic classification results

Dataset release: in 2015
Detailfrage
Kernfrage
Forschungsfrage
Kostenfrage
Sachfrage
Schlüsselfrage
References


