Introduction

Costanza Conforti
05.02.2015
Overview

- Bird’s eye view
- Preliminary steps:
  - Identification
  - Structure Analysis
- German Compounds
- Decompounding German words
  - Decompounding Pipeline
- Tasks
  - State-of-the-art model (Alfonseca 2008)
BIRD’S EYEVIEV
Bird’s eye view

- “There are probably no languages without either compounding, affixing, or both. In other words, there are probably no purely isolating languages. There are a considerable number of languages without inflection, perhaps none without compounding and derivation”

  Greenberg 1963. In Scalise 2010

- \([X \triangleright Y]_X\) \([N+A]_N\) *camposanto* (sp, ‘field+holly, graveyard’)

  \([N+[P+N]]_N\) *arc-en-ciel* (fr, ‘arch in sky, rainbow’)

- \([X \triangleright Y]_Y\) \([A+N]_N\) *green card* (engl), \([N+N]_N\) *Taxifahrer* (germ)

- \([X \triangleright Y]_Z\) \([V+N]_A\) *Quēdé* (jap, ‘lack+morals, immoral’)

  Scalise 2010

Decompounding German Words
Bird’s eye view

- Hybrid status:
  - Syntactically complex units which behave in a syntactically simple way
    Matthews 1947, In: Scalise 2010

- Phonetics + Syntax-Morphology Interface + Semantics
  - Complex interface domains of language competence
    Baroni 2010

- Orthography
  - Distinct words vs. one-words compounds
    - Unusual POS sequence vs. Lexical recognition
    - Simplify parsing vs. Segmentation

Decompounding German Words
PRELIMINARY STEPS

IDENTIFICATION

STRUCTURE ANALYSIS
Preliminary steps - Identification

- High type frequency/ low token frequency

- Productive compounding:
  
  Sizeable number of new orthographic words will constantly be added to the language

Baroni 2002

- Training corpus: Data sparseness

  Decompounding: compound as concatenation of its units
Preliminary steps - Structure analysis

- Bracketing/Labelling compound’s internal structure

<table>
<thead>
<tr>
<th></th>
<th>General %</th>
<th>Rom %</th>
<th>Germ %</th>
<th>Slav %</th>
<th>East A. %</th>
</tr>
</thead>
<tbody>
<tr>
<td>Right</td>
<td>66.7</td>
<td>40.7</td>
<td>87.0</td>
<td>61.9</td>
<td>57.5</td>
</tr>
<tr>
<td>No head</td>
<td>16.3</td>
<td>31.4</td>
<td>8.9</td>
<td>12.2</td>
<td>17.7</td>
</tr>
<tr>
<td>Left</td>
<td>6.8</td>
<td>20.3</td>
<td>1.9</td>
<td>6.0</td>
<td>6.8</td>
</tr>
<tr>
<td>Both</td>
<td>5.9</td>
<td>6.8</td>
<td>1.3</td>
<td>3.1</td>
<td>15.0</td>
</tr>
</tbody>
</table>

As can be seen, the Slavic group follows the general pattern more closely than the other groups. The Romance languages exhibit a relatively high percentage of exocentric constructions, while Germanic languages are more consistently right-headed. The East Asian languages are different in allowing a relatively high percentage of compounds with two heads, most of which are coordinate compounds.

In languages such as Italian, the exocentric pattern V+N is one of the most productive processes in compound formation. In some languages we also find a pattern that has been called 'absolute exocentricity' (Scalise, Fabregas & Forza 2009), from both a categorial and semantic point of view. In such compounds, the output category is entirely different from the categories of the constituents, as illustrated in (15) for Chinese, Turkish and Italian.

(15)  
A+A = N  Ch.  大小  dàxië or 'large small = size'
A+A = Adv  Tu.  補充 aptal aptal 'stupid stupid = in a stupid way'
V+V = N  It.  salì scendi 'go up, go down = elevator'

Despite their relative frequency, exocentric compounds have generally been viewed as a problem for morphological theory, since it is necessary to account for information present in the whole structure that is not present in the constituents. In fact, this has led to a number of analyses in which endocentric readings have been proposed for such compounds. For example, Bisetto (1999) claims that the Romance V+N compounds such as the Italian portalettere, seen above, is endocentric on the assumption that there is a null nominalizing suffix after the verb expressing the meaning of portalettere 'carrier of letters'. Booij (2005) claims that a different type of exocentric...
Preliminary steps - Structure analysis

- Recursion

\[
[\text{child} \text{ language}] \\
[[\text{child} \text{ language}] \text{acquisition}] \\
[[[\text{child} \text{ language}] \text{acquisition}] \text{research}] \\
[[[[\text{child} \text{ language}] \text{acquisition}] \text{research}] \text{group}] \\
[[[[[[\text{child} \text{ language}] \text{acquisition}] \text{research}] \text{group}] \text{member}]
\]

Baroni 2010

Decompounding German Words
Preliminary steps - Structure analysis

- L/R branching: Not Dichotomous!

GermaNet 8.0
Henrich 2011
GERMAN COMPOUNDS
German compounds

- Statistics:
  APA newswire corpus (28 mill words)
  47% types
  7% tokens $\rightarrow$ hapax/rare $<5$
  Baroni 2002

- NN – AN – VN – PN – BN
  62% - 95%
  Baroni 2002 - Heinrich 2011

Decompounding German Words
German compounds

[[Modifier + (Linking morpheme) + (-)] + Head]

\[
\text{bound form:} \quad \downarrow \\
- \text{Hände-druck} \\
- \text{Farb-fernseher} \\
- \text{Doktor-/Doktoren-}
\]

\[
\downarrow \\
\text{Paradigmatic morphemes} \quad == \quad \text{Primitive words}
\]

Baroni 2002
Alfonseca 2008

\[
\text{Semantics-morphologic properties} \\
\downarrow \\
\emptyset \\
-\text{e} \\
+\text{s} \\
+\text{e} \\
+\text{en} \\
+\text{nen} \\
+\text{ens} \\
+\text{es} \\
+\text{ns} \\
+\text{er}
\]
# German compounds

<table>
<thead>
<tr>
<th>Anzahl der Lemmata</th>
<th>Suffix</th>
<th>Beispiel</th>
</tr>
</thead>
<tbody>
<tr>
<td>22759</td>
<td>∅</td>
<td>Kohlsuppe</td>
</tr>
<tr>
<td>9637</td>
<td>☞s</td>
<td>Staatsfeind</td>
</tr>
<tr>
<td>5307</td>
<td>☞n</td>
<td>Soziologenkongreß</td>
</tr>
<tr>
<td>4316</td>
<td>☞en</td>
<td>Straußenei</td>
</tr>
<tr>
<td>2610</td>
<td>☞nen</td>
<td>Wöchnerinnenheim</td>
</tr>
<tr>
<td>618</td>
<td>☞us ☞en</td>
<td>Aphorismenschatz</td>
</tr>
<tr>
<td>348</td>
<td>☞um ☞en</td>
<td>Museenverwaltung</td>
</tr>
<tr>
<td>255</td>
<td>☞um ☞a</td>
<td>Aphrodisiakaverkäufer</td>
</tr>
<tr>
<td>122</td>
<td>☞e</td>
<td>Kirchhof</td>
</tr>
<tr>
<td>95</td>
<td>☞a ☞en</td>
<td>Madonnenkult</td>
</tr>
<tr>
<td>87</td>
<td>☞e</td>
<td>Hundehalter</td>
</tr>
<tr>
<td>73</td>
<td>“ ☞e</td>
<td>Gänseklein</td>
</tr>
<tr>
<td>59</td>
<td>☞on ☞en</td>
<td>Stadionverbot</td>
</tr>
<tr>
<td>43</td>
<td>☞es</td>
<td>Geisteshaltung</td>
</tr>
<tr>
<td>38</td>
<td>“ ☞er</td>
<td>Blätterwald</td>
</tr>
<tr>
<td>33</td>
<td>☞en</td>
<td>Südwind</td>
</tr>
<tr>
<td>28</td>
<td>☞on ☞a</td>
<td>Pharmakaanalyse</td>
</tr>
<tr>
<td>25</td>
<td>☞er</td>
<td>Geisterstunde</td>
</tr>
<tr>
<td>19</td>
<td>☞ien</td>
<td>Prinzipienreiter</td>
</tr>
<tr>
<td>11</td>
<td>☞e ☞i</td>
<td>Carabinierischule</td>
</tr>
</tbody>
</table>

Langer 1998

Decompounding German Words
DECOMPOUNDING
GERMAN WORDS

Guidelines
Decompounding German words

- Decompounding:
  LINGUISTIC – CONSERVATIVE – AGGRESSIVE

Braschler 2003

- Overrecognize (Schweiner-ei)
  + Lexicalized Compounds (Frühstück, Freitag)

Goldsmith 2002
Decompounding Pipeline

1. **[Prior step: compound list]**
   - es. Bispo Santos 2014, GermaNet 8.0: semi-automatically generated

2. **Calculate every possible way of splitting a word in 1+ parts**
   - each of one is a known word > corpus + allowing for linking morphemes
   - only binary splits/recursively

---

**Diagram:**

```
  Aktionsplan  →  actionplan
     /\     \\
   /  \    /  \
  Aktion s  plan
     /  \    /  \
   /    \  /    \  
  Akt  ion s plan  →  act ion plan
```
3. Score those parts
   Weighting function:
   - Frequency-Based Methods
   - Probability-Based Methods

4. Take the highest-scoring decomposition.
   (if it contains just one part, it means that word is not a compound)
TASKS
Tasks

- **Prediction in AAC systems** Baroni et al. 2002
- **Speech Understanding Systems** Adda et al. 2009
- **Machine Translation**
- **Information Retrieval** Braschler 2003 - ecc
  - **Query Keywords** Alfonseca 2008
State-of-the-art model
Alfonseca 2008

- Query Keywords  ➔  noisy data

<table>
<thead>
<tr>
<th>Word</th>
<th>Probable interpretation</th>
</tr>
</thead>
<tbody>
<tr>
<td>achzigerjahre</td>
<td>achtzig+jahre (misspelled)</td>
</tr>
<tr>
<td>activelor</td>
<td>unknown word</td>
</tr>
<tr>
<td>adminslots</td>
<td>admin slots (English)</td>
</tr>
<tr>
<td>agilitydog</td>
<td>agility dog (English)</td>
</tr>
<tr>
<td>akkuwarner</td>
<td>unknown word</td>
</tr>
<tr>
<td>alabams</td>
<td>alabama (misspelled)</td>
</tr>
<tr>
<td>almyghtyzeus</td>
<td>almyghty zeus (English and misspelled)</td>
</tr>
<tr>
<td>amaryllo</td>
<td>amarillo (Spanish and misspelled)</td>
</tr>
<tr>
<td>ampihbienn</td>
<td>amphibian (misspelled)</td>
</tr>
</tbody>
</table>

- Unknown Compounds: useful information
  *Turingmaschine, Blumenstrause, Bismarkausseenpolitik*

Decompounding German Words
State-of-the-art model
Alfonseca 2008

\[
\text{Precision} = \frac{\text{correct splits}}{\text{correct splits} + \text{wrong faulty splits} + \text{wrong splits}}
\]

\[
\text{Recall} = \frac{\text{correct splits}}{\text{correct splits} + \text{wrong faulty splits} + \text{wrong no splits}}
\]

\[
\text{Accuracy} = \frac{\text{correct splits}}{\text{correct splits} + \text{wrong splits}}
\]

Koehn 2003
## State-of-the-art model
### Alfonseca 2008

<table>
<thead>
<tr>
<th>Method</th>
<th>Precision</th>
<th>Recall</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Never split</td>
<td>-</td>
<td>0.00%</td>
<td>64.09%</td>
</tr>
<tr>
<td>Geometric mean of frequencies</td>
<td>39.77%</td>
<td>54.06%</td>
<td>65.58%</td>
</tr>
<tr>
<td>Compound probability</td>
<td>60.41%</td>
<td>80.68%</td>
<td>76.23%</td>
</tr>
<tr>
<td>Mutual Information</td>
<td>82.00%</td>
<td>48.29%</td>
<td>80.52%</td>
</tr>
<tr>
<td>Support-Vector Machine</td>
<td>83.56%</td>
<td>79.48%</td>
<td>87.21%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Language</th>
<th>P</th>
<th>R</th>
<th>A</th>
</tr>
</thead>
<tbody>
<tr>
<td>German</td>
<td>83.56%</td>
<td>79.48%</td>
<td>87.21%</td>
</tr>
<tr>
<td>Dutch</td>
<td>78.99%</td>
<td>76.18%</td>
<td>83.45%</td>
</tr>
<tr>
<td>Danish</td>
<td>81.97%</td>
<td>87.12%</td>
<td>85.36%</td>
</tr>
<tr>
<td>Norwegian</td>
<td>88.13%</td>
<td>93.05%</td>
<td>90.40%</td>
</tr>
<tr>
<td>Swedish</td>
<td>83.34%</td>
<td>92.98%</td>
<td>87.79%</td>
</tr>
<tr>
<td>Finnish</td>
<td>90.79%</td>
<td>91.21%</td>
<td>91.62%</td>
</tr>
</tbody>
</table>

### Metrics

Table 3 shows the results reported for German. An ablation study, reported in that paper, indicated that the contribution of the web anchor texts is minimal. Concerning the second step, there is some work apart from a few exceptions, the results are rather confidence interval are highlighted. Interestingly, the highest value and those which are inside its 95% all the other languages together. For each row, the another experiment to check whether the models languages. Table 5 shows the results obtained in this another experiment to check whether the models languages. Table 5 shows the results obtained in this method, or learning a language model from a cor-

## Decompounding German Words
References


Thank you - Danke - Grazie