Letter Sequence Labeling for Compound Splitting

Jianqiang Ma, Verena Henrich and Erhard Hinrichs
University of Tübingen, SFB 833 & Department of Linguistics

Compound Splitting

Task

<table>
<thead>
<tr>
<th>Input: Ortname ‘place name’</th>
<th>Output: Ort Name</th>
</tr>
</thead>
</table>

- compounds are written as contiguous strings without word delimiters in many languages
- treating each compound as a unique word would dramatically increase the vocabulary size
- impact: identifying compounds and their constituents can benefit NLP tasks such as machine translation

Frequency-based approach

- idea: the splitting hypothesis that has the highest geometric mean of frequencies of constituent words is the best splitting (Koehn and Knight, 2003)
- many state-of-the-art splitters for German (Popović et al., 2006; Weller and Heid, 2012) are based on it
- limitation: high frequency does not guarantee correct splitting; relying on other NLP tools for better results

Proposed Method

Motivation

- learning to make splitting decisions
- exploiting rich word form features such as -ung (a German suffix) implicitly as letter ngrams
- avoiding dependencies on external morphological analyzer and/or POS tagger as in current methods

Letter Sequence Labeling

- splitting outputs are encoded as string sequences with white spaces added between constituent words e.g. Ort name → Ort Name
- compound splitting is formulated as labeling each letter in terms of its positional role within words

<table>
<thead>
<tr>
<th>Input: O r t s n a m e</th>
<th>Label: B M M E B M M E</th>
</tr>
</thead>
</table>

- label set: {Begin, Middle, End, Singleton}
- model: linear chain conditional random fields (CRF, Lafferty et al., 2001)
- features: functions describing letter ngrams in the input and the label for the current/previous letter

Experiments

Compiling the GermaNet dataset

- extraction of 51,667 unique compounds from GermaNet 9.0 (Henrich and Hinrichs, 2011); each compound has up to 5 constituents (avg. 2.1)
- sampling of 31,076 unique non-compounds from the rest words in the GermaNet with the constraint that the word length is no more than 10 letters
- the total set consists of 82,743 words

Experiments with GermaNet dataset

- disjoint training/development/test sets (7:1:2)

Results of our method with different features

<table>
<thead>
<tr>
<th>model</th>
<th>precision</th>
<th>recall</th>
<th>accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>uni- &amp; bi-grams</td>
<td>0.873</td>
<td>0.833</td>
<td>0.857</td>
</tr>
<tr>
<td>+ trigrams</td>
<td>0.937</td>
<td>0.920</td>
<td>0.925</td>
</tr>
<tr>
<td>+ 4-grams</td>
<td>0.952</td>
<td>0.940</td>
<td>0.942</td>
</tr>
<tr>
<td>+ 5-grams</td>
<td>0.955</td>
<td>0.941</td>
<td>0.943</td>
</tr>
</tbody>
</table>

Experiments with PE dataset

- the PE dataset: 342 compound tokens and 3,009 non-compounds from Parra Escartín (2014); each compound has 2-5 constituents (avg. 2.3)
- training with the GermaNet data except those words appearing in PE data; testing on PE data

Comparison with the state-of-the-art

<table>
<thead>
<tr>
<th>model</th>
<th>precision</th>
<th>recall</th>
<th>accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Popović et al. (2006)</td>
<td>0.961</td>
<td>0.752</td>
<td>0.972</td>
</tr>
<tr>
<td>Weller &amp; Heid (2012)</td>
<td>0.992</td>
<td>0.757</td>
<td>0.975</td>
</tr>
<tr>
<td>this work</td>
<td>0.855</td>
<td>0.930</td>
<td>0.980</td>
</tr>
</tbody>
</table>

* Results are from Parra Escartín (2014)

Error analysis of the precision score

- half of the “non-compounds” that our model “wrongly” splits are adjective/verbal compounds
- the rest of true wrong split errors can be reduced by using higher quality training cases of non-compound

Conclusion

- letter sequence labeling can split compounds accurately without using external NLP modules
- letter ngrams can capture morpho/orthographic regularities without manually encoding knowledge