POS Tagging for German: How Important is the Right Context?

Paper by:
Steliana Ivanova and Sandra Kübler
Outline

• Introduction
• Problem with the German Language
• Experimental Setup
• Results
• Conclusion
Outline

• Introduction
• Problem with the German Language
• Experimental Setup
• Results
• Conclusion
Introduction

- Usually POS tagging from left to right
- POS Taggers based on Markov Models generally only use the left context
- Only 1 or 2 words to the left
- Many languages can only be disambiguated by the context to the right of the word
Outline

• Introduction

• Problem with the German Language

• Experimental Setup

• Results

• Conclusion
Problem with the German Language

- Determiners can also serve as relative and demonstrative pronouns
Problem with the German Language

• Determiners can also serve as relative and demonstrative pronouns
  ▪ **Determiner (is a word that modifies a noun):**

    **The** boy is **a** teacher.  
    **Der Junge** ist **ein** Lehrer.
Problem with the German Language

• Determiners can also serve as relative and demonstrative pronouns
  ▪ **Determiner (is a word that modifies a noun):**
    
    The boy is a teacher. Der Junge ist ein Lehrer.

  ▪ **Relative pronoun (substitutes a noun in a sentence):**
    
    The Girl bought bread which she ate afterwards. Das Mädchen kaufte Brot das sie danach aß.
Problem with the German Language

• Determiners can also serve as relative and demonstrative pronouns
  ▪ Determiner (is a word that modifies a noun):
    The boy is a teacher. Der Junge ist ein Lehrer.

  ▪ Relative pronoun (substitutes a noun in a sentence):
    The Girl bought bread which she ate afterwards.
    Das Mädchen kaufte Brot das sie danach aß.

  ▪ Demonstrative pronoun (replaces a noun that is near):
    You take these papers, I’ll take those.
    Der da hat Schuld.
Beide wissen, der Anpassungsdruck an den High-Schools ist bereits enorm hoch - der, den Mitschüler ausüben. Now enormously high - the one which classmates exert.
Problem with the German Language

Beide wissen, der Anpassungsdruck an den High-Schools ist bereits jetzt enorm hoch - der, den Mitschüler ausüben. Both know, the peer pressure at the high schools is already now enormously high - the one which classmates exert.
Outline

• Introduction
• Problem with the German Language
  • Experimental Setup
• Results
• Conclusion
Experimental Setup - Corpus

• Data taken from Tübingen Treebank of Written German (TüBa-D/Z)
  ▪ Annotated corpus
  ▪ Tagged with STTS tag set
• Consists of newspaper articles from a German newspaper
• 90% of data for training – 10% for testing
Experimental Setup - Tagger

• **Memory-based POS tagger-generator (MBT) was used**

• **Proceeds in two phases:**
  1. Generating a tagger using a memory-based learner – in this case: TiMBL (Tilburg Memory-Based Learner)
  2. Tagging text with tagger generated in (1)

• **Learning:** Learning methods assume that decisions are made on previously seen events

• **When new word has to be classified:**
  - k nearest neighbors with similar context are retrieved
  - It picks the majority class (the one with more tags of the same kind)
Experimental Setup - Tagger

- **Memory-based POS tagger-generator (MBT) was used**
- Proceeds in two phases:
  1. Generating a tagger using a memory-based learner – in this case: TiMBL (Tilburg Memory-Based Learner)
  2. Tagging text with tagger generated in (1)

- Learning: Learning methods assume that decisions are made on previously seen events
- When new word has to be classified:
  - $k$ nearest neighbors with similar context are retrieved
  - It picks the majority class (the one with more tags of the same kind)

- TiMBL was run with default settings first
- Test was conducted with left context and right context from 0 to 2 words
Outline

• Introduction
• Problem with the German Language
• Experimental Setup
  • Results
• Conclusion
Results

- At first they tested different context sizes
- TnT had an accuracy of 97.04% on the same corpus
- MBT, without any improvement:  Bigram 93.91%  Trigram 94.00%
Results

• At first they tested different context sizes
• TnT had an accuracy of 97.04% on the same corpus
• MBT, without any improvement:       Bigram 93.91%     Trigram 94.00%

<table>
<thead>
<tr>
<th></th>
<th>ddfWaa</th>
<th>dfWaa</th>
<th>ddfWa</th>
<th>fWaa</th>
<th>ddfW</th>
<th>dfW</th>
<th>fWa</th>
</tr>
</thead>
<tbody>
<tr>
<td>forward</td>
<td>96.08</td>
<td>96.05</td>
<td>96.06</td>
<td>95.27</td>
<td>94.00</td>
<td>93.81</td>
<td>95.29</td>
</tr>
<tr>
<td>backward</td>
<td>96.06</td>
<td>96.05</td>
<td>96.02</td>
<td>95.97</td>
<td>95.26</td>
<td>95.28</td>
<td>95.39</td>
</tr>
</tbody>
</table>
Results

• At first they tested different context sizes
• TnT had an accuracy of 97.04% on the same corpus
• MBT, without any improvement: Bigram 93.91% Trigram 94.00%

<table>
<thead>
<tr>
<th></th>
<th>ddfWaa</th>
<th>dfWaa</th>
<th>ddfWa</th>
<th>fWaa</th>
<th>ddfW</th>
<th>dfW</th>
<th>fWa</th>
</tr>
</thead>
<tbody>
<tr>
<td>forward</td>
<td>96.08</td>
<td>96.05</td>
<td>96.06</td>
<td>95.27</td>
<td>94.00</td>
<td>93.81</td>
<td>95.29</td>
</tr>
<tr>
<td>backward</td>
<td>96.06</td>
<td>96.05</td>
<td>96.02</td>
<td>95.97</td>
<td>95.26</td>
<td>95.28</td>
<td>95.39</td>
</tr>
</tbody>
</table>
Results

- At first they tested different context sizes
- TnT had an accuracy of 97.04% on the same corpus
- MBT, without any improvement: Bigram 93.91% Trigram 94.00%

<table>
<thead>
<tr>
<th></th>
<th>ddfWaa</th>
<th>dfWaa</th>
<th>ddfWa</th>
<th>fWaa</th>
<th>ddfW</th>
<th>dfW</th>
<th>fWa</th>
</tr>
</thead>
<tbody>
<tr>
<td>forward</td>
<td>96.08</td>
<td>96.05</td>
<td>96.06</td>
<td>95.27</td>
<td>94.00</td>
<td>93.81</td>
<td>95.29</td>
</tr>
<tr>
<td>backward</td>
<td>96.06</td>
<td>96.05</td>
<td>96.02</td>
<td>95.97</td>
<td>95.26</td>
<td>95.28</td>
<td>95.39</td>
</tr>
</tbody>
</table>

- After optimization an accuracy of 96.73% was reached
- Many ambiguities were disambiguated by the tagger
Outline

• Introduction
• Problem with the German Language
• Experimental Setup
• Results
• Conclusion
Conclusion

The right context is very important for the German Language
Thank you for your attention