ANALYSIS OF PART-OF-SPEECH TAGGING OF HISTORICAL GERMAN TEXTS

DATECH 2017

Markus Paluch, Gabriela Rotari, David Steding, Maximilian Weß, Maria Moritz, Marco Büchler

June 1 2017, University of Göttingen
# Table of Contents

1. Introduction
2. Data
3. Experimental Setup
4. Results
5. Conclusion
INTRODUCTION
RQ: Should POS-Taggers be trained on a certain epoch/period?
POS-Tagging: The process of marking up the words in a text to a particular part of speech (tag).
**POS-Tagging**: The process of marking up the words in a text to a particular part of speech (tag).

<table>
<thead>
<tr>
<th>Word / Token</th>
<th>Tag</th>
<th>Wordclass</th>
</tr>
</thead>
<tbody>
<tr>
<td>Money</td>
<td>NN</td>
<td>noun</td>
</tr>
<tr>
<td>does</td>
<td>DOZ</td>
<td>does</td>
</tr>
<tr>
<td>not</td>
<td>*</td>
<td>negation</td>
</tr>
<tr>
<td>smell</td>
<td>VB</td>
<td>verb</td>
</tr>
<tr>
<td>.</td>
<td>.</td>
<td>punctuation</td>
</tr>
</tbody>
</table>

Not all words correspond to a single wordclass.

- mobile  JJ  adverb
- mobile  NN  noun
POS-Tagger trained on dataset X: A computer model which learned to perform POS-Tagging on texts in X.
German ≠ English, it is known that:
A German trained tagger processing English texts performs badly and vice versa.

Historical German ≠ Contemporary German, we ask:
Does a tagger trained on contemporary German processing historical German texts performs badly and vice versa?
RQ: Should POS-Taggers be trained on a certain epoch/period?
DATA
German Text Archive (Deutsches Text Archiv, DTA)¹

- comprises 1598 texts
- dating from 1050 to 1926

### Table 1: Datasets for the experiment

<table>
<thead>
<tr>
<th>Period</th>
<th>Texts</th>
<th>Tokens</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baroque</td>
<td>1600-1720</td>
<td>76</td>
</tr>
<tr>
<td>Romanticism</td>
<td>1810-1840</td>
<td>264</td>
</tr>
<tr>
<td>Modernism</td>
<td>1880-1920</td>
<td>87</td>
</tr>
</tbody>
</table>
EXPERIMENTAL SETUP
Included POS-Tagger algorithms¹:

- **Unigram**
- **Hidden Markov Model (HMM)**
- **Conditional Random Field (CRF)**
- **Perceptron**

¹. All used algorithm implementations are from the natural language toolkit (NLTK)
Procedure:

1. Training of taggers on data
2. Testing of taggers (Results)
METHODOLOGY

![Diagram showing the relationship between Baroque, Romanticism, and Modernism through training and testing taggers. The diagram illustrates how results from each period feed into the next phase of the methodology.]
RESULTS
RESULTS OF MODERNISM TAGGERS

Accuracy of Taggers on DTA data

Taggers are trained on Modernism

Accuracy in %

Tested data period

Baroque

Romanticism

Modernism

Perceptron

Unigram

CRF

HMM
RESULTS OF ROMANTICISM TAGGERS

Accuracy of Taggers on DTA data
Taggers are trained on Romanticism

Accuracy in %

Baroque  Romanticism  Modernism
Tested data period

Perceptron  Unigram  CRF  HMM
RESULTS OF BAROQUE TAGGERS

Baroque

Accuracy of Taggers on DTA data
Taggers are trained on Baroque

Accuracy in %
100.0% 95.0% 90.0% 85.0% 80.0% 75.0% 70.0% 65.0% 60.0%

Baroque  Romanticism  Modernism
Tested data period

Perceptron  Unigram  CRF  HMM
We handtagged about 300 tokens of 1 text per period.

What happens if a tagger trained on non goldstandard data (DTA) is tested against goldstandard data (handtagging)?

<table>
<thead>
<tr>
<th>Taggers</th>
<th>Accuracy</th>
<th>Handtagging</th>
</tr>
</thead>
<tbody>
<tr>
<td>Modernisim</td>
<td>94.1%-98.3%</td>
<td>91.7%-95.6%</td>
</tr>
<tr>
<td>Romanticism</td>
<td>87.7%-97.7%</td>
<td>93.6%-96.8%</td>
</tr>
<tr>
<td>Baroque</td>
<td>89.9%-97.3%</td>
<td>88.1%-90.5%</td>
</tr>
</tbody>
</table>
CONCLUSION
1. Using a POS-Tagger trained on a different period of the same language can dramatically decrease its performance!
   • Higher time differences between periods increase the performance decrease.

2. DTA POS-Tags for Baroque are more erroneous than POS-Tags of Romanticism or Modernism on our handtagged examples.

RQ: Should POS-Taggers be trained on a certain epoch/period? Yes!
Visit us

http://www.etrap.eu
contact@etrap.eu

Stealing from one is plagiarism, stealing from many is research
(Wilson Mitzner, 1876-1933)
The theme this presentation is based on is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License. Changes to the theme are the work of eTRAP.