Improving historical spelling normalization with bi-directional LSTMs and multi-task learning

Marcel Bollmann\textsuperscript{1}  
Anders Søgaard\textsuperscript{2}

\textsuperscript{1}Ruhr-Universität Bochum, Germany \hspace{1cm} \textsuperscript{2}University of Copenhagen, Denmark

COLING 2016
December 13, 2016
Motivation

Sample of a manuscript from Early New High German

Marcel Bollmann, Anders Søgaard

Historical spelling normalization with bi-LSTMs and MTL
A corpus of Early New High German

- Medieval religious treatise
  “Interrogatio Sancti Anselmi de Passione Domini”

- > 50 manuscripts and prints (in German)

- 14th–16th century

- Various dialects
  - Bavarian
  - Middle German
  - Low German
  - …

Sample from an Anselm manuscript

http://www.linguistics.rub.de/anselm/

Marcel Bollmann, Anders Søgaard

Historical spelling normalization with bi-LSTMs and MTL
Examples for historical spellings

**Frau** *(woman)*  
fraw, frawe, fräwe, frauwe, fraüwe, frow, frouw, vraw, vrow, vorwe, vrauwe, vrouwe

**Kind** *(child)*  
chind, chinde, chindt, chint, kind, kinde, kindi, kindt, kint, kinth, kynde, kynt

**Mutter** *(mother)*  
moder, moeder, mueter, müeter, muoter, muotter, muter, mutter, mvoter, mvter, mweter
Dealing with spelling variation

The problems…

▶ Difficult to annotate with tools aimed at modern data
▶ High variance in spelling
▶ None/very little training data

Normalization

▶ Removes variance
▶ Enables re-using of existing tools
▶ Useful annotation layer (e.g. for corpus query)

Normalization as the mapping of historical spellings to their modern-day equivalents.
Dealing with spelling variation

The problems…

- Difficult to annotate with tools aimed at modern data
- High variance in spelling
- None/very little training data

Normalization…

- Removes variance
- Enables re-using of existing tools
- Useful annotation layer (e.g. for corpus query)

**Normalization** as the mapping of historical spellings to their modern-day equivalents.
Our approach

- Character-based sequence labelling

\[ Hist \quad vrow \]
\[ Norm \quad frau \]
Our approach

- Character-based sequence labelling

\[ \text{Hist} \quad \text{v r o w} \]
\[ \text{Norm} \quad \text{f r a u} \]
Our approach

➢ Character-based sequence labelling

\[ Hist \quad v \quad r \quad o \quad w \]
\[ Norm \quad f \quad r \quad a \quad u \]

➢ Not all examples are so straightforward…
Our approach

Hist vsfuret

Norm ausführt
Our approach

- Hist vs f u r e t
- Norm a u s f ü h r t

- Iterated Levenshtein distance alignment (Wieling et al., 2009)
Our approach

- Iterated Levenshtein distance alignment (Wieling et al., 2009)
- Epsilon label for "deletions"
Our approach

- Iterated Levenshtein distance alignment (Wieling et al., 2009)
- Epsilon label for “deletions”
- Leftward merging of “insertions”
Our approach

Iterated Levenshtein distance alignment (Wieling et al., 2009)

- Epsilon label for “deletions”
- Leftward merging of “insertions”
- Special “beginning of word” symbol
Our model

- **Prediction layer**
  - ε
  - f
  - r
  - a
  - u

- **Stack of bi-LSTM layers**
- **Embedding layer**
  - <BOS>
  - v
  - r
  - o
  - w

Marcel Bollmann, Anders Søgaard

Historical spelling normalization with bi-LSTMs and MTL
## Evaluation

- **44 texts** from the Anselm corpus
  - ≈ 4,200 – 13,200 tokens per text (average: 7,353 tokens)
- **1,000 tokens** for evaluation
- **1,000 tokens** for development (not used)
- Remaining tokens for training

- **Pre-processing**
  - Remove punctuation
  - Lowercase all words
Methods for comparison

▶ Norma (Bollmann, 2012)
  ▶ Developed on the same corpus
  ▶ Methods
    ▶ Automatically learned “replacement rules”
    ▶ Weighted Levenshtein distance
  ▶ Requires lexical resource

▶ CRFsuite (Okazaki, 2007)
  ▶ Same input as the bi-LSTM model
  ▶ Features: two surrounding characters
### Results

<table>
<thead>
<tr>
<th>ID</th>
<th>Region</th>
<th>Norma</th>
<th>CRF</th>
<th>Bi-LSTM</th>
</tr>
</thead>
<tbody>
<tr>
<td>B2</td>
<td>West Central</td>
<td>76.10%</td>
<td>74.60%</td>
<td>82.00%</td>
</tr>
<tr>
<td>D3</td>
<td>East Central</td>
<td>80.50%</td>
<td>77.20%</td>
<td>80.10%</td>
</tr>
<tr>
<td>M</td>
<td>East Upper</td>
<td>74.30%</td>
<td>72.80%</td>
<td>83.90%</td>
</tr>
<tr>
<td>M5</td>
<td>East Upper</td>
<td>80.60%</td>
<td>76.40%</td>
<td>77.70%</td>
</tr>
<tr>
<td>St2</td>
<td>West Upper</td>
<td>73.20%</td>
<td>73.20%</td>
<td>78.20%</td>
</tr>
</tbody>
</table>

<p>| | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td><strong>Average</strong></td>
<td>77.83%</td>
<td>75.73%</td>
<td>79.90%</td>
</tr>
</tbody>
</table>
Multi-task learning

Stack of bi-LSTMs

embedding layer

prediction layer

Marcel Bollmann, Anders Søgaard
Historical spelling normalization with bi-LSTMs and MTL
Multi-task learning

Stack of bi-LSTMs

prediction layer for A

prediction layer for B

embedding layer
Multi-task learning

Stack of bi-LSTMs

Prediction layer for A

Embedding layer

Prediction layer for B

Marcel Bollmann, Anders Søgaard

Historical spelling normalization with bi-LSTMs and MTL
Multi-task learning

- Problem definition
- Neural network approach
- Multi-task learning

Learning a joint model
- Evaluation
- Conclusion

Marcel Bollmann, Anders Søgaard

Historical spelling normalization with bi-LSTMs and MTL
One prediction layer for each text

Embedding

Bi-LSTM Stack

Predict (B2) → ... → Predict (B2)

Predict (D3) → ... → Predict (D3)

Predict (M5) → ... → Predict (M5)

... ··· ··· ··· ...

Predict (St2)
Evaluation

- Each of the 44 texts as a separate task
  - **Training**: Randomly sample from all texts
  - **Evaluation**: Use the prediction layer for the current task

- For comparison: Norma/CRF
  - **Augment** training set with 10,000 randomly sampled instances
## Results

<table>
<thead>
<tr>
<th>ID</th>
<th>Region</th>
<th>Norma</th>
<th>Bi-LSTM</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Plain</td>
<td>Aug.</td>
</tr>
<tr>
<td>B2</td>
<td>West Central</td>
<td>76.10%</td>
<td>77.60%</td>
</tr>
<tr>
<td>D3</td>
<td>East Central</td>
<td>80.50%</td>
<td>80.20%</td>
</tr>
<tr>
<td>M</td>
<td>East Upper</td>
<td>74.30%</td>
<td>74.40%</td>
</tr>
<tr>
<td>M5</td>
<td>East Upper</td>
<td>80.60%</td>
<td>80.70%</td>
</tr>
<tr>
<td>St2</td>
<td>West Upper</td>
<td>73.20%</td>
<td>73.40%</td>
</tr>
<tr>
<td></td>
<td><strong>Average</strong></td>
<td><strong>77.83%</strong></td>
<td><strong>77.48%</strong></td>
</tr>
</tbody>
</table>
### Results

<table>
<thead>
<tr>
<th>ID</th>
<th>Region</th>
<th>Norma</th>
<th>Bi-LSTM</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Plain</td>
<td>Aug.</td>
</tr>
<tr>
<td>B2</td>
<td>West Central</td>
<td>76.10%</td>
<td>77.60%</td>
</tr>
<tr>
<td>D3</td>
<td>East Central</td>
<td>80.50%</td>
<td>80.20%</td>
</tr>
<tr>
<td>M</td>
<td>East Upper</td>
<td>74.30%</td>
<td>74.40%</td>
</tr>
<tr>
<td>M5</td>
<td>East Upper</td>
<td>80.60%</td>
<td>80.70%</td>
</tr>
<tr>
<td>St2</td>
<td>West Upper</td>
<td>73.20%</td>
<td>73.40%</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Average</td>
<td></td>
<td>77.83%</td>
<td>77.48%</td>
</tr>
</tbody>
</table>
Conclusion

- Deep learning works for historical spelling normalization
  - …despite small datasets (≈ 4,200 – 13,200 tokens per text)

- Outperforms Norma & CRF baseline
  - …despite not using a lexical resource (like Norma)

- Multi-task learning setup improves results
  - Way to deal with data sparsity problem
  - Many improvements conceivable
Thank you for listening!
References

