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

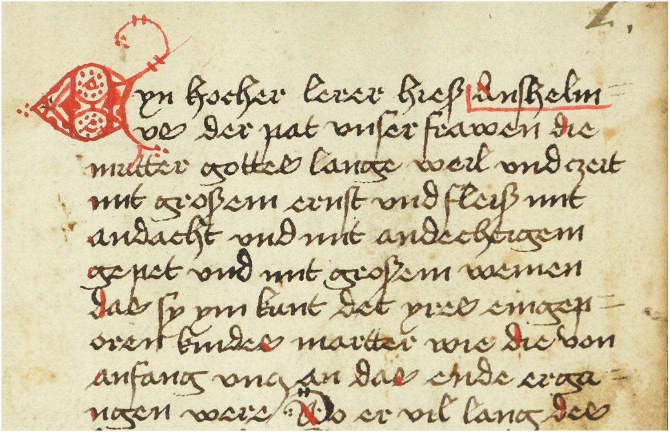
Marcel Bollmann[1] Anders Søgaard[2]

[1]Ruhr-Universität Bochum, Germany
[2]University of Copenhagen, Denmark

COLING 2016

December 13, 2016

Motivation

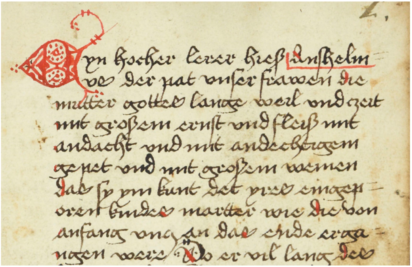


Sample of a manuscript from Early New High German

A corpus of Early New High German

- Medieval religious treatise
  "Interrogatio Sancti Anselmi de Passione Domini"

- $>$ 50 manuscripts and prints (in German)
- 14^{th}–16^{th} century
- Various dialects
  - *Bavarian*
  - *Middle German*
  - *Low German*
  - ...



Sample from an Anselm manuscript

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

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

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	vrow
Norm	frau

Our approach

- Character-based sequence labelling

Hist	v	r	o	w
Norm	f	r	a	u

Our approach

- Character-based sequence labelling

Hist	v	r	o	w
Norm	f	r	a	u

- Not all examples are so straightforward...

Our approach

Hist	vsfuret
Norm	ausführt

Our approach

Hist		v	s	f	u		r	e	t
Norm	a	u	s	f	ü	h	r		t

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

Our approach

Hist		v	s	f	u		r	e	t
Norm	a	u	s	f	ü	h	r	ε	t

- Iterated Levenshtein distance alignment (Wieling et al., 2009)
- Epsilon label for “deletions”

Our approach

Hist		v	s	f	u	r	e	t
Norm	a	u	s	f	üh	r	ε	t

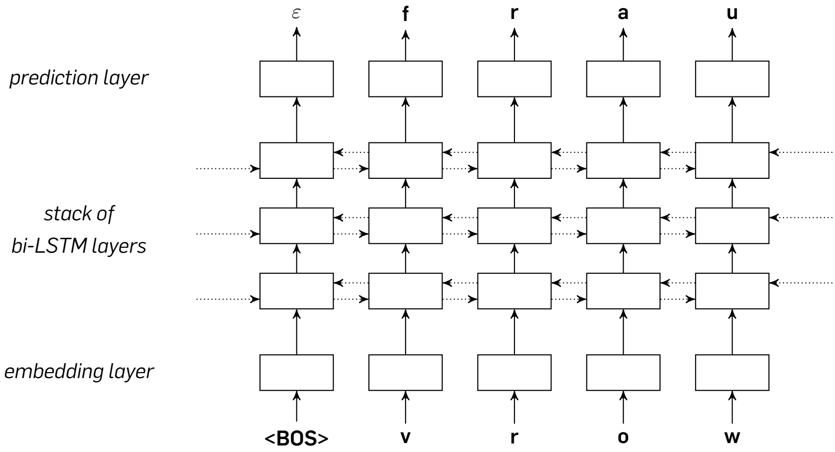
- Iterated Levenshtein distance alignment (Wieling et al., 2009)
- Epsilon label for "deletions"
- Leftward merging of "insertions"

Our approach

Hist	_	v	s	f	u	r	e	t
Norm	a	u	s	f	üh	r	ε	t

- Iterated Levenshtein distance alignment (Wieling et al., 2009)
- Epsilon label for "deletions"
- Leftward merging of "insertions"
- Special "beginning of word" symbol

Our model


ε
f
r
a
u
prediction layer
stack of
bi-LSTM layers
embedding layer
<BOS>
v
r
o
w

Evaluation

- 44 texts from the Anselm corpus
  - $\approx$ 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

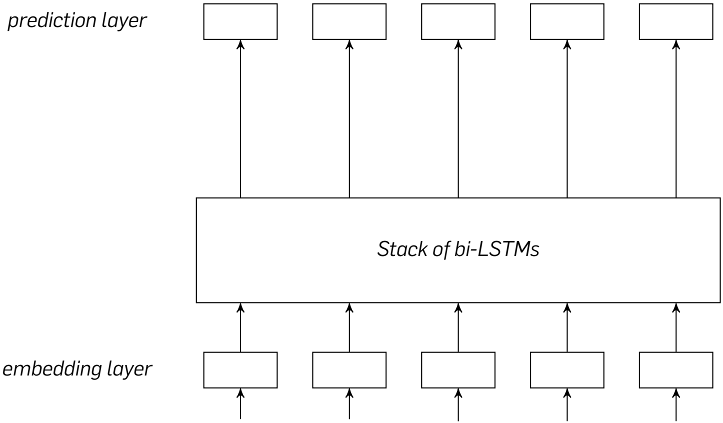
Methods for comparison

- 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

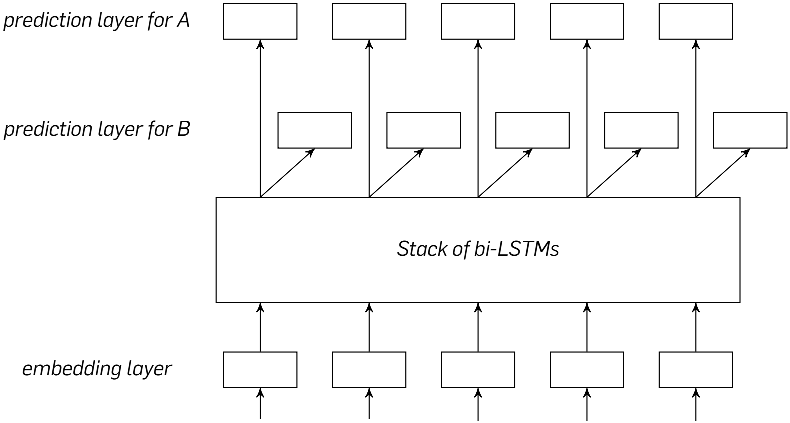
Results

ID	Region	Norma	CRF	Bi-LSTM
B2	West Central	76.10%	74.60%	82.00%
D3	East Central	80.50%	77.20%	80.10%
M	East Upper	74.30%	72.80%	83.90%
M5	East Upper	80.60%	76.40%	77.70%
St2	West Upper	73.20%	73.20%	78.20%
⋮		⋮		⋮
Average		77.83%	75.73%	79.90%

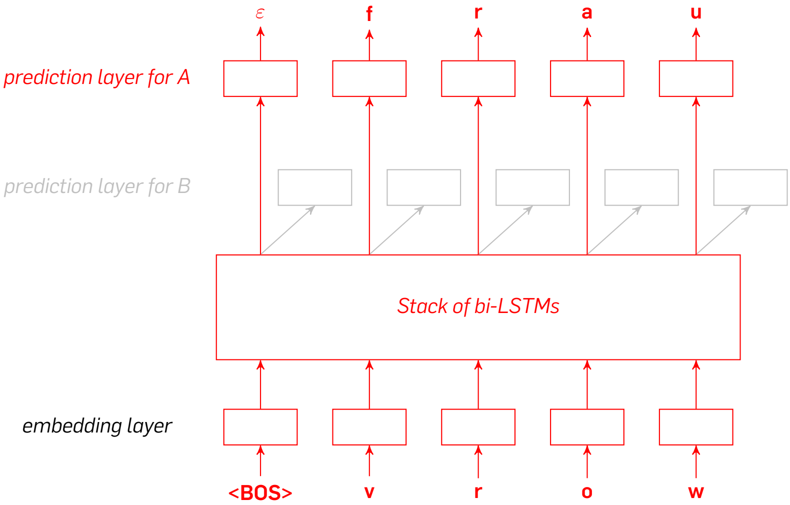
Multi-task learning



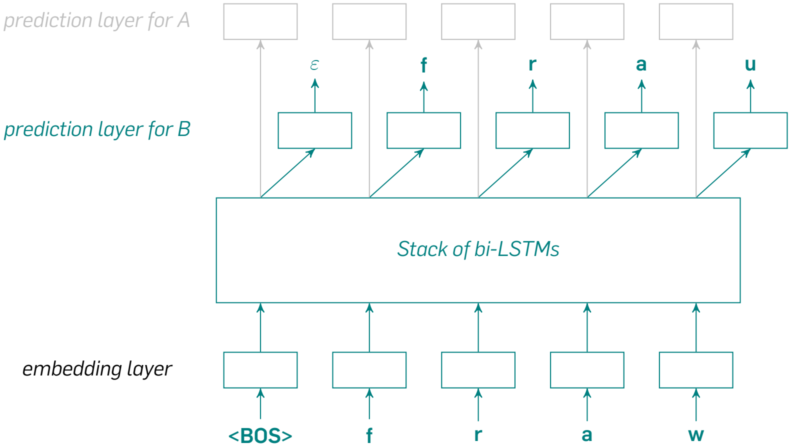
Multi-task learning



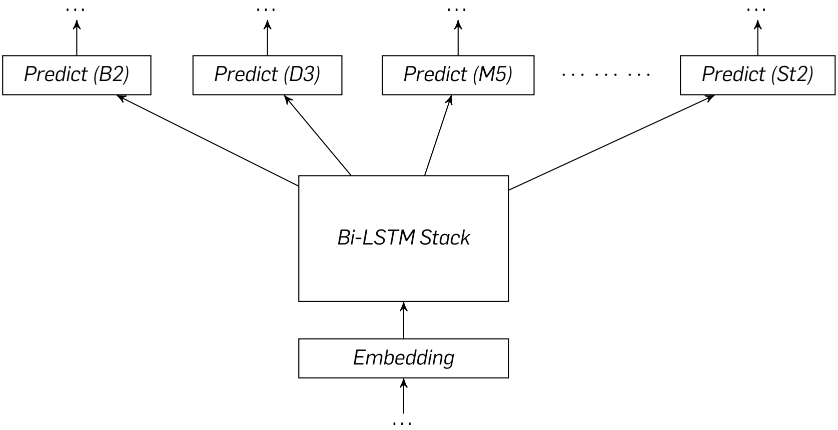
Multi-task learning



Multi-task learning



One prediction layer for each text



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

ID	Region	Norma		Bi-LSTM	
		Plain	*Aug.*	*Plain*	*MTL*
B2	West Central	76.10%	77.60%	82.00%	79.60%
D3	East Central	80.50%	80.20%	80.10%	81.20%
M	East Upper	74.30%	74.40%	83.90%	80.90%
M5	East Upper	80.60%	80.70%	77.70%	82.90%
St2	West Upper	73.20%	73.40%	78.20%	79.90%
⋮		⋮		⋮	
Average		77.83%	77.48%	79.90%	80.55%

Results

ID	**Region**	**Norma**		**Bi-LSTM**	
		Plain	*Aug.*	*Plain*	*MTL*
B2	West Central	76.10%	77.60%	82.00%	79.60%
D3	East Central	80.50%	80.20%	80.10%	81.20%
M	East Upper	74.30%	74.40%	83.90%	80.90%
M5	East Upper	80.60%	80.70%	77.70%	82.90%
St2	West Upper	73.20%	73.40%	78.20%	79.90%
⋮		⋮		⋮	
Average		77.83%	77.48%	79.90%	80.55%

Conclusion

- Deep learning works for historical spelling normalization
  - …despite small datasets ($\approx$ 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

Bollmann, M. (2012). (Semi-)automatic normalization of historical texts using distance measures and the Norma tool. In *Proceedings of the Second Workshop on Annotation of Corpora for Research in the Humanities (ACRH-2).* Lisbon, Portugal.

Okazaki, N. (2007). *CRFsuite: a fast implementation of conditional random fields (CRFs).* http://www.chokkan.org/software/crfsuite/. Retrieved from `http://www.chokkan.org/software/crfsuite/`

Wieling, M., Prokić, J., & Nerbonne, J. (2009). Evaluating the pairwise string alignment of pronunciations. In *Proceedings of the EACL 2009 Workshop on Language Technology and Resources for Cultural Heritage, Social Sciences, Humanities, and Education (LaTeCH – SHELT&R 2009)* (pp. 26–34). Athens, Greece.