Learning attention for historical text normalization by learning to pronounce

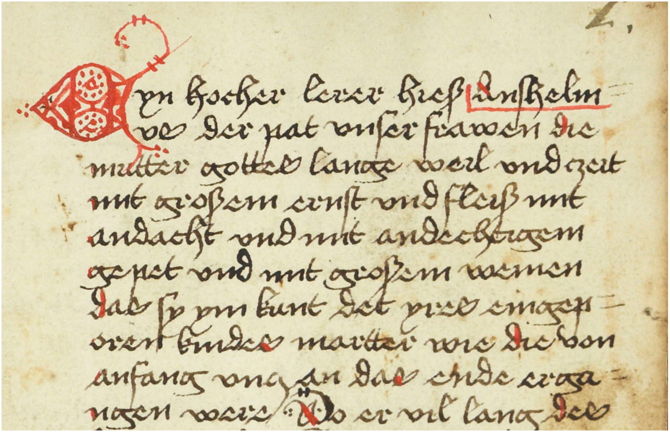
Marcel Bollmann[1] Joachim Bingel[2] Anders Søgaard[2]

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

ACL 2017, Vancouver, Canada

July 31, 2017

Motivation

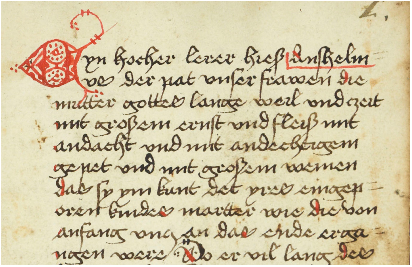

Eyn hocher lerer hieß Anshelm
ve der pat vnser frawen die
mutter gottes lange weil vnd zeit
mit großem ernst vnd fleiß mit
andacht vnd mit andechtigem
gepet vnd mit großem weinen
das sy ym kunt det yrer einigen
eren kindes marter wie die von
anfang vntz an das ende erga
ngen were Do er vil lang des

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)
- 14th–16th 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

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

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

Previous work

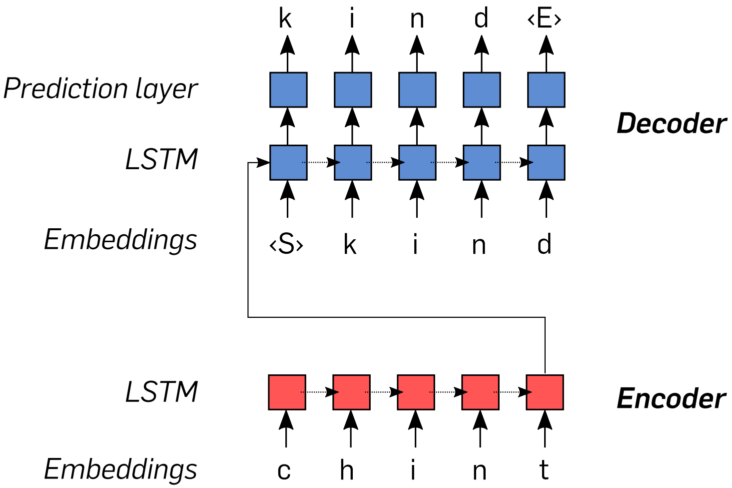
- Hand-crafted algorithms
  - VARD (Baron & Rayson, 2008)
  - Norma (Bollmann, 2012)
- Character-based statistical machine translation (CSMT)
  - Scherrer and Erjavec (2013), Pettersson et al. (2013), ...
- Sequence labelling with neural networks
  - Bollmann and Søgaard (2016)

Previous work

- Hand-crafted algorithms
  - VARD (Baron & Rayson, 2008)
  - Norma (Bollmann, 2012)
- Character-based statistical machine translation (CSMT)
  - Scherrer and Erjavec (2013), Pettersson et al. (2013), ...
- Sequence labelling with neural networks
  - Bollmann and Søgaard (2016)

- **Now:** "Character-based neural machine translation"

An encoder/decoder model



An encoder/decoder model

	Avg. Accuracy
Bi-LSTM tagger *(Bollmann & Søgaard, 2016)*	**79.91%**
Greedy	78.91%
Base model	

Evaluation on 43 texts from the Anselm corpus
(≈ 4,000–13,000 tokens each)

An encoder/decoder model

		Avg. Accuracy
Bi-LSTM tagger *(Bollmann & Søgaard, 2016)*		**79.91%**
Base model	Greedy	78.91%
	Beam	79.27%

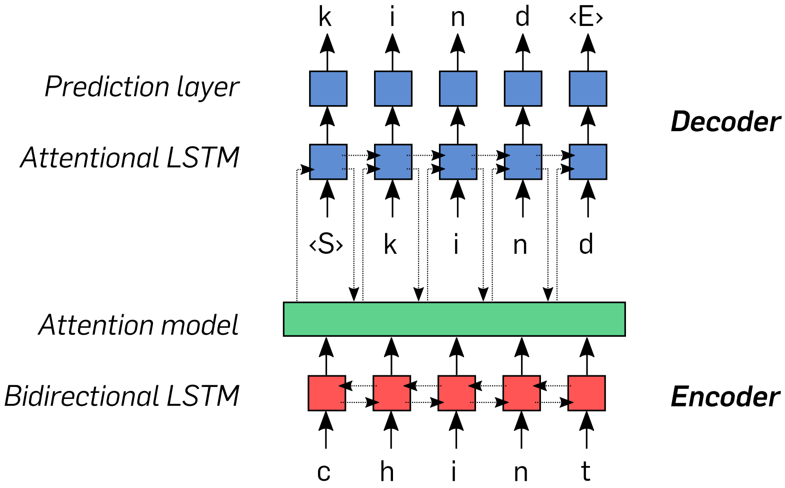
Evaluation on 43 texts from the Anselm corpus
(≈ 4,000–13,000 tokens each)

An encoder/decoder model

		Avg. Accuracy
Bi-LSTM tagger *(Bollmann & Søgaard, 2016)*		79.91%
Base model	Greedy	78.91%
	Beam	79.27%
	Beam + Filter	**80.46%**

Evaluation on 43 texts from the Anselm corpus
(≈ 4,000–13,000 tokens each)

Attentional model



Attentional model

		Avg. Accuracy
Bi-LSTM tagger *(Bollmann & Søgaard, 2016)*		79.91%
Base model	Greedy	78.91%
	Beam	79.27%
	Beam + Filter	80.46%
	Beam + Filter + Attention	**82.72%**

Evaluation on 43 texts from the Anselm corpus
(≈ 4,000–13,000 tokens each)

Learning to pronounce

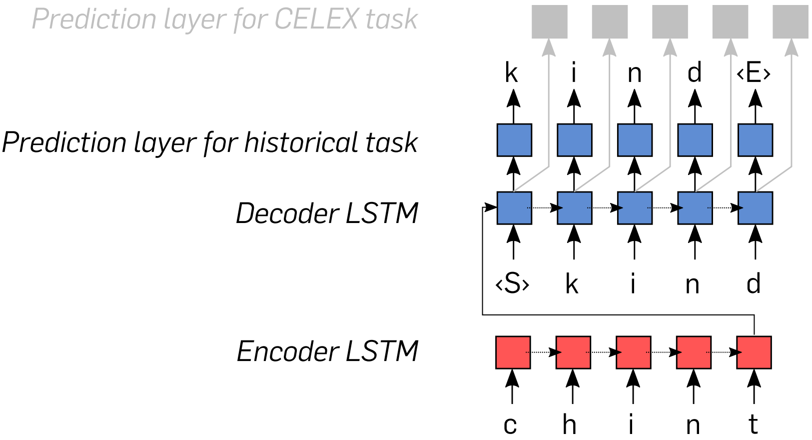
Can we improve results with multi-task learning?

Learning to pronounce

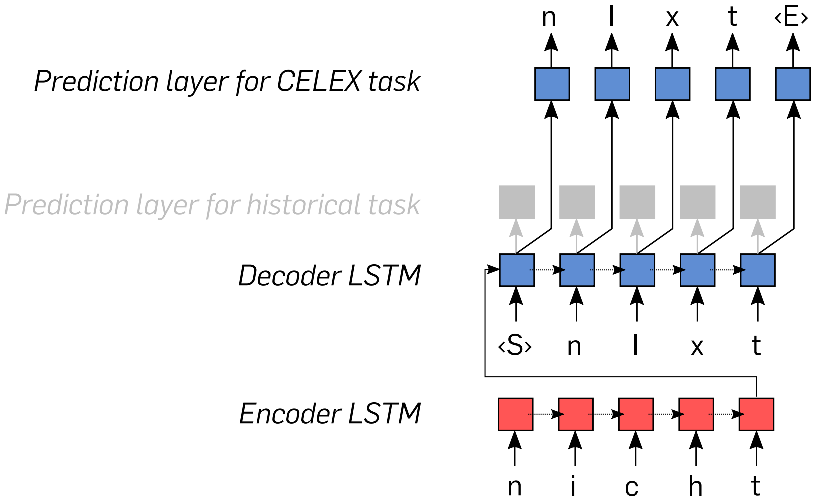
- **Idea:** grapheme-to-phoneme mapping as auxiliary task
- CELEX 2 lexical database (Baayen et al., 1995)
- Sample mappings for German:

Jungfrau	$\rightarrow$	*jUN-frB*
Abend	$\rightarrow$	*ab@nt*
nicht	$\rightarrow$	*nIxt*

Multi-task learning



Multi-task learning



Multi-task learning

		Avg. Accuracy
Bi-LSTM tagger *(Bollmann & Søgaard, 2016)*		79.91%
Base model	Greedy	78.91%
	Beam	79.27%
	Beam + Filter	80.46%
	Beam + Filter + Attention	**82.72%**
MTL model	Greedy	80.64%
	Beam	81.13%
	Beam + Filter	**82.76%**
	Beam + Filter + Attention	82.02%

Why does MTL not improve with attention?

Hypothesis

Attention and MTL learn similar functions of the input data.

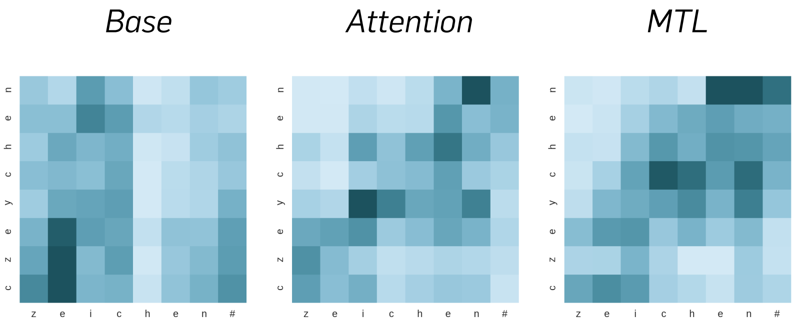
> *"MTL can be used to coerce the learner to attend to patterns in the input it would otherwise ignore. This is done by forcing it to learn internal representations to support related tasks that depend on such patterns."*
>
> – Caruana (1998), p. 112 f.

Comparing the model outputs

		gewarnet	uberhübe	scholt
Base model	G	prandet	überbroch	sollt
	B	prandert	überbräche	sollt
	B+F	pranget	über	sollt
	B+F+A	gewarnt	übergebe	sollte
MTL model	G	gewarntet	überbeh	sollte
	B	gewarntet	übereube	sollte
	B+F	gewarnt	übergebe	sollte
	B+F+A	gewand	über	sollte
Target		gewarnt	überhob	sollte

Saliency plots

Li, Chen, Hovy, and Jurafsky (2016)



→ for words ≥ 7 characters, Attention/MTL correlate most

Conclusion

- Encoder/decoder models for historical text normalization are competitive
  - Despite small datasets ($\approx$ 4,200 – 13,200 tokens per text)
  - Beam search & attention improve results further
- MTL with grapheme-to-phoneme task helps
- Attention and MTL have a similar effect
  - Can this be reproduced on other tasks?
  - What factors affect this (choice of attention mechanism/auxiliary task/…)?

Thank you for listening!

Code `https://bitbucket.org/mbollmann/acl2017`

Further Qs? `bollmann@linguistics.rub.de` @mmbollmann

References I

Baayen, R. H., Piepenbrock, R., & Gulikers, L. (1995). *The CELEX lexical database (Release 2) (CD-ROM).* Linguistic Data Consortium, University of Pennsylvania, Philadelphia, PA.

Baron, A., & Rayson, P. (2008). VARD 2: A tool for dealing with spelling variation in historical corpora. In *Proceedings of the Postgraduate Conference in Corpus Linguistics.*

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.

Bollmann, M., & Søgaard, A. (2016). Improving historical spelling normalization with bi-directional lstms and multi-task learning. In *Proceedings of coling 2016* (pp. 131–139). Osaka, Japan.

Caruana, R. (1998). Multitask learning. In *Learning to learn* (pp. 95–133). Springer. Retrieved from `http://dl.acm.org/citation.cfm?id=296635.296645`

References II

Li, J., Chen, X., Hovy, E., & Jurafsky, D. (2016). Visualizing and understanding neural models in NLP. In *Proceedings of the 2016 conference of the north american chapter of the association for computational linguistics: Human language technologies* (pp. 681–691). Association for Computational Linguistics. Retrieved from `http://aclweb.org/anthology/N16-1082` doi: 10.18653/v1/N16-1082

Pettersson, E., Megyesi, B., & Tiedemann, J. (2013). An SMT approach to automatic annotation of historical text. In *Proceedings of the nodalida workshop on computational historical linguistics.* Oslo, Norway.

Scherrer, Y., & Erjavec, T. (2013). Modernizing historical Slovene words with character-based SMT. In *Proceedings of the 4th biennial workshop on balto-slavic natural language processing.* Sofia, Bulgaria.

References III

Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., … Bengio, Y. (2015). Show, attend and tell: Neural image caption generation with visual attention. In *Jmlr workshop and conference proceedings: Proceedings of the 32nd international conference on machine learning* (Vol. 37, pp. 2048–2057). Lille, France. Retrieved from `http://proceedings.mlr.press/v37/xuc15.pdf`

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.

Attention mechanism: details

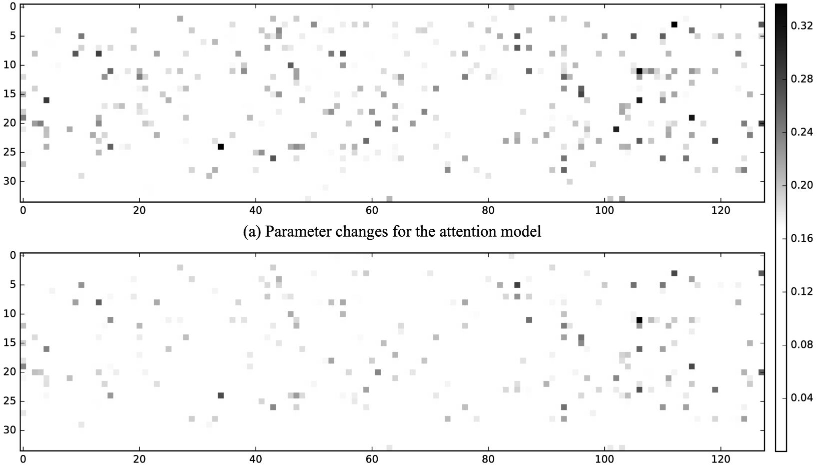
- Attention mechanism follows Xu et al. (2015)

$$\hat{z}_t = \sum_{i=1}^{n} \alpha_i a_i \tag{1}$$

$$\alpha = softmax(f_{\text{att}}(a, h_{t-1})) \tag{2}$$

$$\begin{aligned}
i_t &= \sigma(W_i[h_{t-1}, y_{t-1}, \hat{z}_t] + b_i) \\
f_t &= \sigma(W_f[h_{t-1}, y_{t-1}, \hat{z}_t] + b_f) \\
o_t &= \sigma(W_o[h_{t-1}, y_{t-1}, \hat{z}_t] + b_o) \\
g_t &= \tanh(W_g[h_{t-1}, y_{t-1}, \hat{z}_t] + b_g) \\
c_t &= f_t \odot c_{t-1} + i_t \odot g_t \\
h_t &= o_t \odot \tanh(c_t)
\end{aligned} \tag{3}$$

Differences of learned parameters



(a) Parameter changes for the attention model

(b) Parameter changes for the multi-task model