UNIVERSITY OF COPENHAGEN
DEPARTMENT OF COMPUTER SCIENCE

Multi-task learning for historical text normalization: Size matters

Marcel Bollmann, Anders Søgaard, Joachim Bingel

@mmbollmann
marcel@di.ku.dk

What is normalization?

- Mapping words in historical text to modern equivalent
- Example:

erthely
↓
earthly

Why normalization?

- Reducing spelling variation
- Typical use cases:
  - Preprocessing step for NLP tools
  - Preprocessing for linguistic analyses
  - Improve search queries
- Increasingly relevant as more historical texts are digitized!

Dataset/Language		Time Period	Tokens	
			Train	Test
DE_A	German (Anselm)	14th–16th c.	233,947	45,999
DE_R	German (RIDGES)	1482–1652	41,857	9,587
EN	English	1386–1698	147,826	17,644
ES	Spanish	15th–19th c.	97,320	12,479
HU	Hungarian	1440–1541	134,028	16,779
IS	Icelandic	15th c.	49,633	6,037
PT	Portuguese	15th–19th c.	222,525	27,078
SL_B	Slovene (Bohorič)	1750–1840s	50,023	5,969
SL_G	Slovene (Gaj)	1840s–1899	161,211	21,493
SV	Swedish	1527–1812	24,458	29,184

Datasets used in our experiments

Some normalization examples

DE_A	deſe wort ſpricht vnſer liber here iheſus criſtus czu eyme iczlychen menſchen diese wort spricht unser lieber herr jesus christus zu einem ieteslichen menschen
DE_R	ſeind ſÿ doch alle auſz den vier elementen gemiſchet vnd eins feüchter deñ das ander sind sie doch alle aus den vier elementen gemischt und eins feuchter denn das andere
EN	whan your graciouse erthely persoune from your inward spirit ys dessolued when your gracious earthly person from your inward spirit is dissolved
ES	anque tomeys mui mucho travajo tengola guardada pa quando dios sea servido aunque toméis muy mucho trabajo téngola guardada para cuando dios sea servido
HU	ǫ́ zauoc ėſmėǵ felèmèluē kèzdėnc̣ ſirńoc èlmēnèc èzèkèt tolga ez a noemi azeꝛt iǫ́uǫ́ ő szavuk ismét felemelvén kezdének sírniuk elmenjek ezeket toldja ez a noémi azért jöve
IS	þá sem hanz gödverk voru i og þá vrdu hanns gödverk miklu þýngre enn ill þá sem hans góðverk voru í og þá urðu hans góðverk miklu þyngri en ill
PT	cõ a poenetencia que lhe derão pera avisar aos snres do sancto oficio com a penitência que lhe deram para avisar aos senhores do santo ofício
SL_B	ter ne bodi nevéren zhe ſe zherna perſt premozhi tezhe od nje rjav mòk ter ne bodi neveren če se črna prst premoči teče od nje rjav mok
SL_G	in privéže na vsak konec niti drobtino kruha in verže vse kokóšem breskevno vkuhanje lovre in priveže na vsak konec niti drobtino kruha in vrže vse kokošim breskvino vkuhanje lovre
SV	blef av rätten afsagdt det en syyn och rådhgångh nu nästkommande wårdagh hållas blev av rätten avsagt det en syn och rådgång nu nästkommande vårdag hållas

Neural sequence-to-sequence model

- **Input:** Single word form, represented as sequence of characters
- **Encoder:** bi-directional LSTM
- **Decoder:** uni-directional LSTM with attention mechanism
- Generate output sequence character by character, using greedy decoding

Main research questions

★ Can we improve normalization with cross-lingual learning?

★ Do some dataset pairings work better than others?

★ What dataset properties are most predictive of this?

Multi-task learning

- Training on **pairs of datasets** A and B
- Share all model components except for the final prediction layer
- Mini-batch training with 50 samples from both A and B
- Validation on dev sets after every 50,000 samples, keeping only the best models for A and B

Full data scenario

Train on full training sets for both datasets

	DE_A	DE_R	EN	ES	HU	IS	PT	SL_B	SL_G	SV
DE_A		+1.8	+5.4	+6.5	+4.8	+4.5	+4.7	+8.3	+5.6	+6.5
DE_R	-6.4		-18.5	-14.4	-10.9	-14.0	-14.1	-14.4	-17.1	-14.2
EN	+7.5	+23.6		+0.7	-5.1	+1.3	-0.5	+0.0	+1.4	+6.4
ES	+10.2	+8.9	+17.5		+13.2	+17.8	+1.7	+5.3	+3.1	+5.2
HU	+17.6	+42.1	+20.8	+5.2		+2.2	+0.4	+5.3	+7.6	+15.6
IS	-12.0	-9.7	-6.3	-10.9	-5.9		-13.8	-10.2	-10.9	-7.4
PT	+20.7	+29.2	+16.6	+1.4	+12.3	+13.8		+5.3	+6.8	+17.3
SL_B	-16.4	-26.1	-13.2	-15.0	-17.5	-20.9	-20.7		-5.0	-33.2
SL_G	+4.8	+11.2	+13.1	+16.7	+8.1	+16.2	+24.0	+0.8		+11.5
SV	-16.5	-12.7	-14.0	-16.4	-16.7	-13.5	+1.5	-16.0	-16.3	

Sparse data scenario

Train on 5,000 tokens for target dataset (*rows*), jointly with full auxiliary dataset (*columns*)

	DE_A	DE_R	EN	ES	HU	IS	PT	SL_B	SL_G	SV
DE_A		-21.1	-10.8	-4.9	-11.3	-5.6	-10.1	-3.0	-1.3	-6.7
DE_R	-12.9		-10.4	-6.5	-11.6	-10.2	-6.2	-10.8	-13.8	-16.3
EN	-18.5	-23.6		-22.9	-27.0	-19.7	-28.6	-25.1	-29.5	-23.2
ES	-10.2	-13.8	-18.2		-22.0	-17.0	-30.5	-18.0	-19.1	-16.8
HU	-12.1	-12.5	-4.8	-8.0		-6.7	-5.2	+0.7	-2.4	-8.4
IS	-6.5	-7.7	-12.1	-10.2	-11.5		-7.7	-7.7	-8.5	-11.8
PT	-8.8	-7.7	-7.5	-18.6	-13.3	-10.8		-10.6	-9.8	-5.6
SL_B	-11.9	-17.0	-13.3	-13.3	-16.9	-19.8	-14.7		-29.0	-13.4
SL_G	+2.1	-8.0	-8.3	-7.1	-4.8	-5.6	-6.3	-16.6		-4.1
SV	+7.7	-9.8	-16.2	-15.0	-8.3	-14.1	-2.9	-12.4	-7.3	

Percentage change of error of multi-task learning over single-task models
(*blue* = improvements, *red* = error increases)
***Rows* are target datasets, *columns* are auxiliary datasets**

Results

- Multi-task learning helps most **when target dataset is small**
- Multi-task learning can even be **detrimental** when target datasets are already (sufficiently) large
- Size of target training set is more important than choice of auxiliary dataset

Takeaways

★ Multi-task learning can help a lot when you don't have a lot of training data!

★ Always consider the size of the training set when evaluating multi-task learning approaches!

This research was supported by ERC Starting Grant LOWLANDS No. 313695 as well as by Trygfonden.
Parts of this research were carried out by the first author at the Ruhr-Universität Bochum, Germany, supported by Deutsche Forschungsgemeinschaft (DFG), Grant DI 1558/4.

Deutsche Forschungsgemeinschaft
German Research Foundation

European Research Council
Established by the European Commission