Historical text normalization with neural networks

Marcel Bollmann


marcel@di.ku.dk


mmbollmann

UNIVERSITY OF
COPENHAGEN

Stefanie Dipper

RUHR
UNIVERSITÄT
BOCHUM

RUB

Anders Søgaard

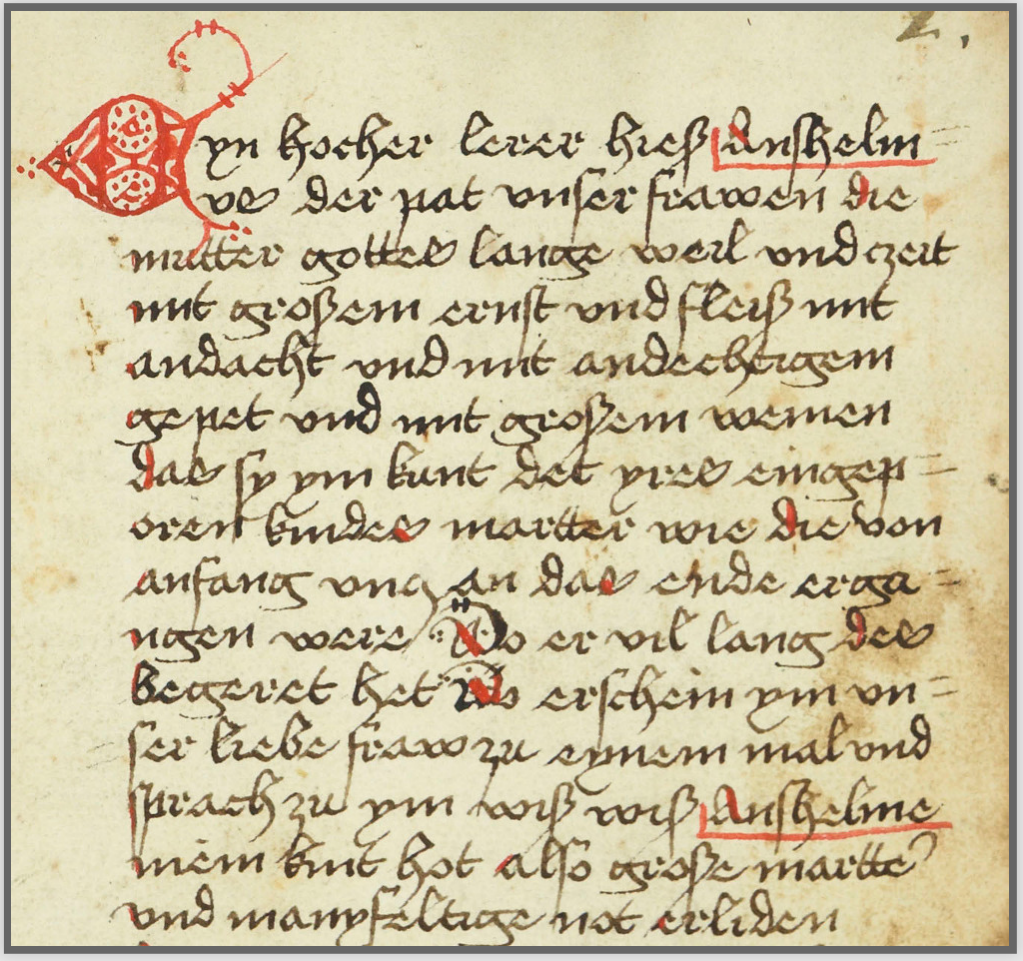
UNIVERSITY OF
COPENHAGEN

Anselm of Canterbury (1033—1109)

Motivation

The Anselm Project

- Collection of *Early New High German* texts
- 14^{th}—16^{th} centuries
- **Goal:** annotated corpus



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

All attested spellings of modern German *“Frau”* in the Anselm corpus

Normalization to the rescue!

fraw	*vrouwe*	*frauwe*	*vrowe*	*fräw*	*vorwe*	…
↓	↓	↓	↓	↓	↓	
Frau	Frau	Frau	Frau	Frau	Frau	…

Outline

1. Defining normalization
2. Automatic normalization
3. Takeaways & recommendations

The Why, What, and How of Normalization

Why normalization?

1. It reduces variance.
2. It can be useful to all users, not just NLP tools.
3. It enables re-using existing tools and resources.

What and how to normalize?

From the “Innsbruck Letter Corpus”:

þe	*quene*	*was*	*ryght*	*gretly*	*displisyd*
the	queen	was	right	greatly	displeased

...but is it always that easy?

Morphology and morphosyntax

- Spanish:

  tu me dyjistes

  dijistes ⟵ graphematically close form

  dijiste ⟵ correct modern form

  “you told me”

Morphology and morphosyntax

- German:

  trinck des waffers zum tag zweymal

  trink des Wassers zum Tag zweimal

  trink das Wasser am Tag zweimal

  “drink the water twice a day”

- Genitive vs. accusative case
- Modern German uses different preposition

Archaic lexemes

- English:

  I let hym wete

  I let him know

- Hungarian:

  ýsa *“certainly”*

  isa

  bizony

Graphematic criterion

Normalization is determined "exclusively via phonological and/or graphematic equivalence operations."

(Krasselt et al., 2015, p. 17)

Dictionary criterion

Normalization "should be a word form that is likely to be present in a modern language dictionary."

(Pettersson, 2016, p. 50)

More problems...

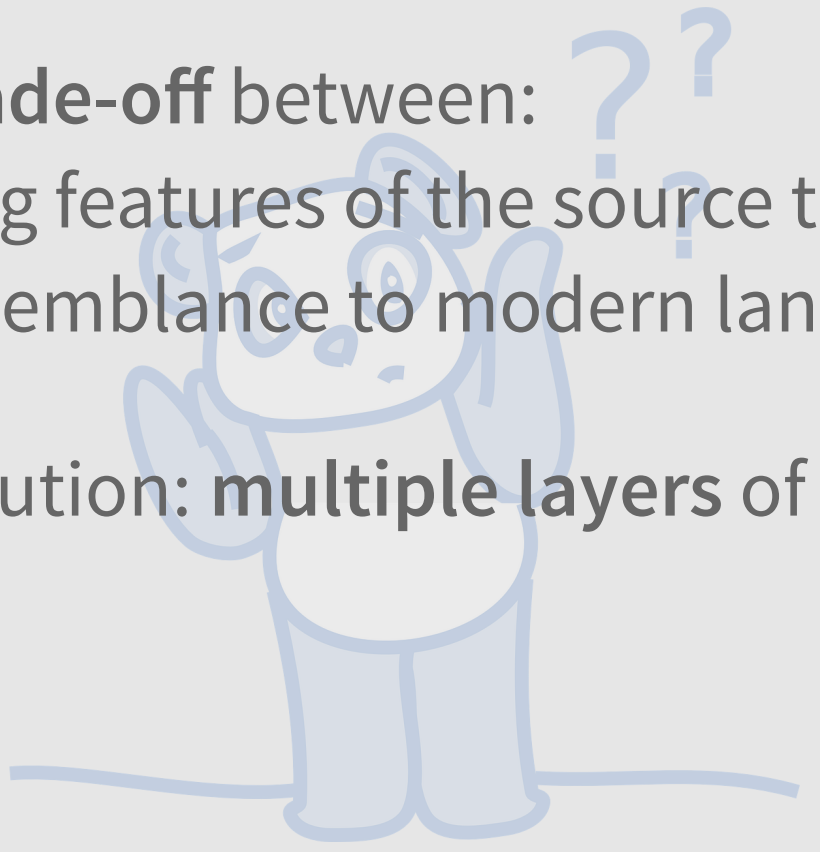
- **Tokenization:** non-1:1 mappings

folt	*u*	*foltu*
sollst	du	sollst_du

- **Abbreviations:** *vra* → *vuestra*
- **Proper nouns:** *nafzerenus* → *Nazareth*
- **Capitalization?**

How to normalize?

- Always a **trade-off** between:
  - preserving features of the source text
  - closer resemblance to modern language
- Possible solution: **multiple layers** of normalization



Automatic Normalization

(supervised)

Many possible techniques

1. Wordlist mapping
2. Rule-based methods
3. Distance-based methods
4. **Norma**

https://github.com/comphist/norma

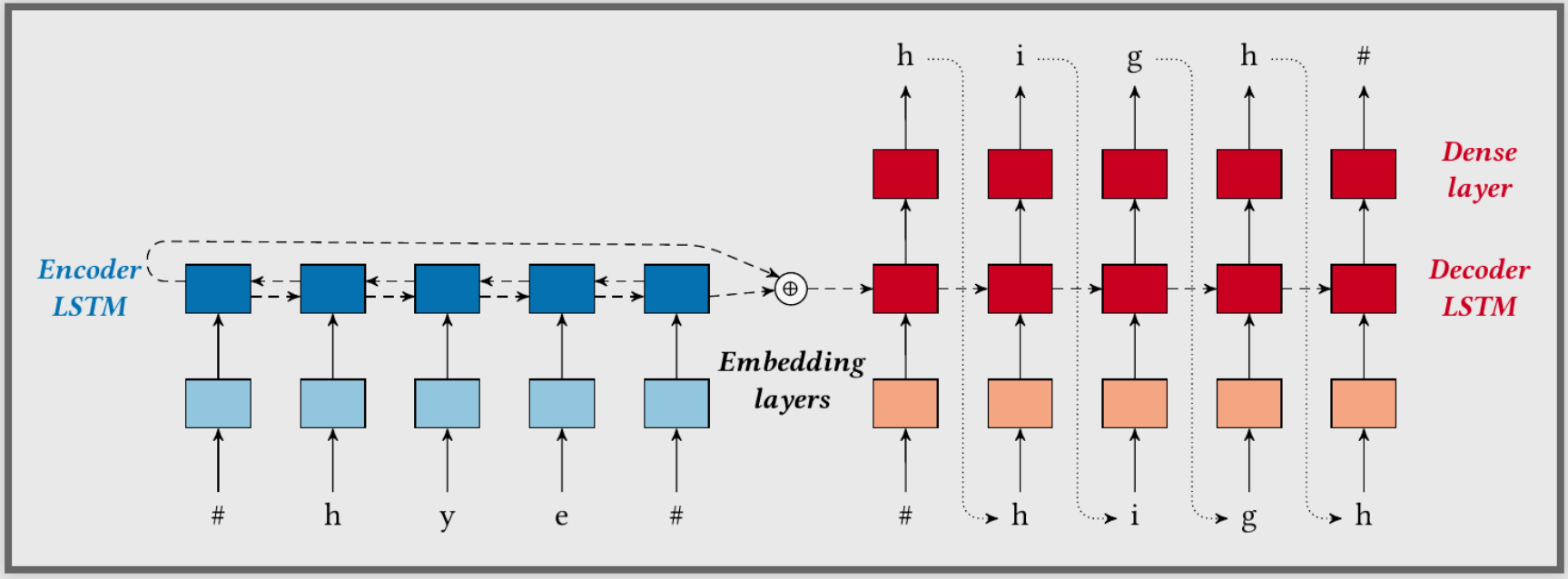
Character-based statistical machine translation

- **Source:** ÷ h y e ÷
- **Target:** ÷ h i g h ÷
- Sánchez-Martínez et al. (2013), Scherrer & Erjavec (2013), Pettersson et al. (2014), Ljubešić et al. (2016), Pettersson (2016), Domingo & Casacuberta (2018), ...
- **cSMTiser**

  https://github.com/clarinsi/csmtiser

Neural networks

(encoder—decoder, seq2seq, ...)



Datasets

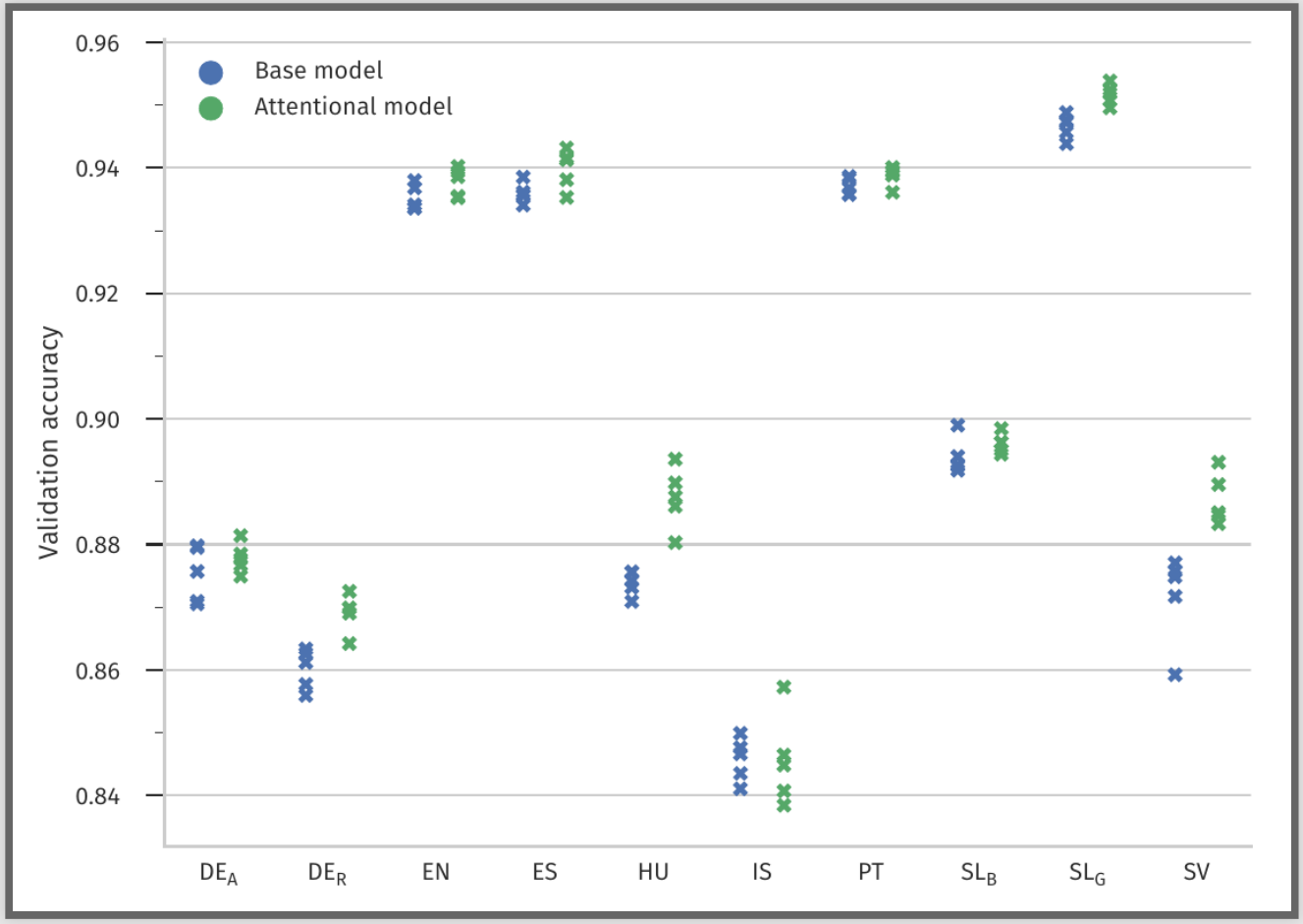
	Language	Corpus	Time Period	Genre	Size (Tokens)
DE_A	**German**	Anselm	14^{th}—16^{th} c.	Religion	326,000
DE_R	**German**	RIDGES	1482—1652	Science	61,000
EN	**English**	ICAMET	1386—1698	Letters	182,000
ES	**Spanish**	Post Scriptum	15^{th}—19^{th} c.	Letters	121,000
HU	**Hungarian**	HGDS	1440—1541	Religion	167,000
IS	**Icelandic**	IcePaHC	15^{th} c.	Religion	62,000
PT	**Portuguese**	Post Scriptum	15^{th}—19^{th} c.	Letters	276,000
SL_B	**Slovene**	goo300k	1750—1840s	Mixed	62,000
SL_G	**Slovene**	goo300k	1840s—1899	Mixed	204,000
SV	**Swedish**	GaW	1527—1812	Mixed	56,000

Hyperparameter tuning

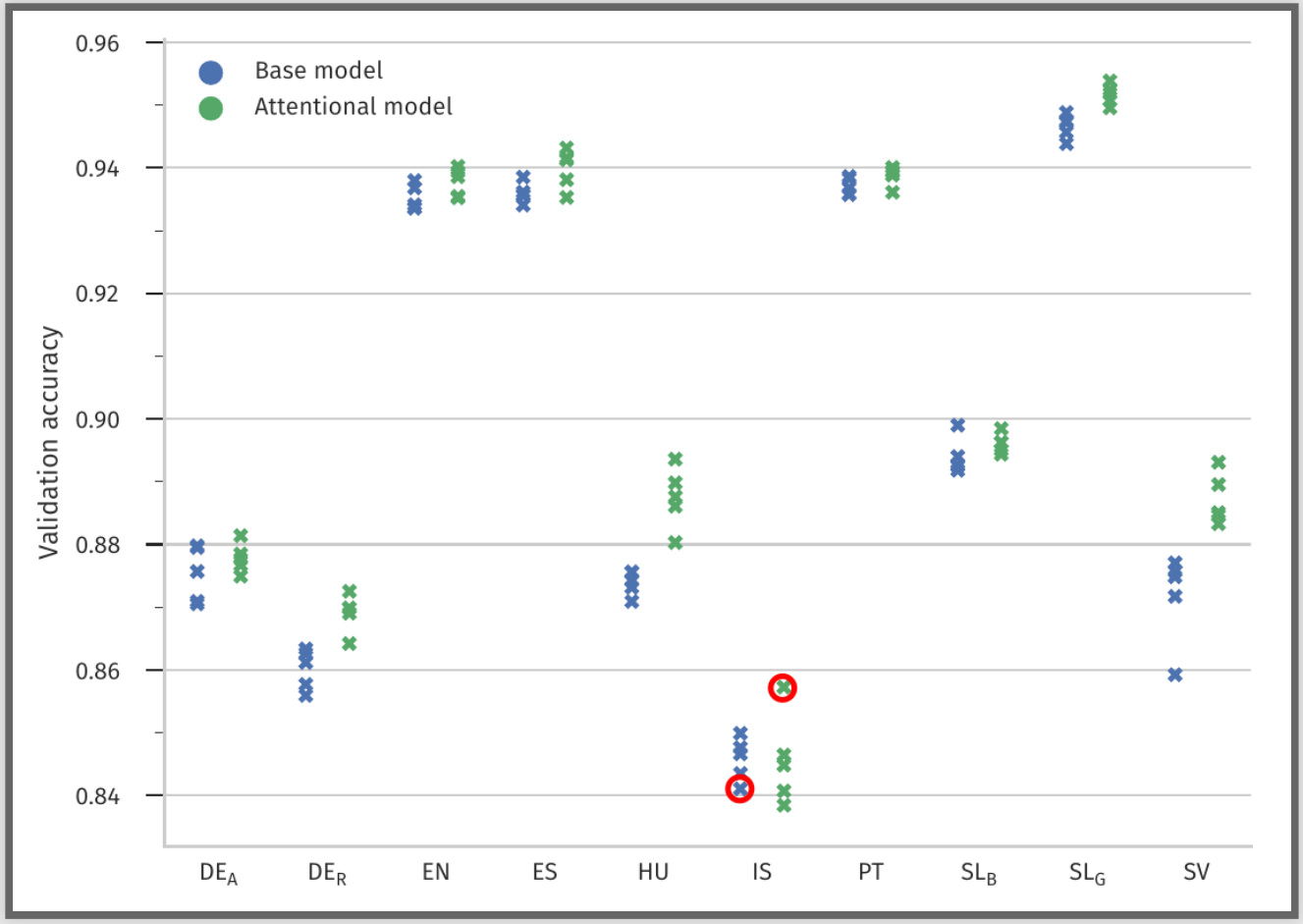
- On 5 datasets, fixed training set size

- *Dimensionality of embeddings:* **60**
- *Number of LSTMs:* **1**
- *Dimensionality of LSTMs:* **300**
- *Dropout rate:* **0.2**
- *Learning rate (Adam):* **0.001** (default)

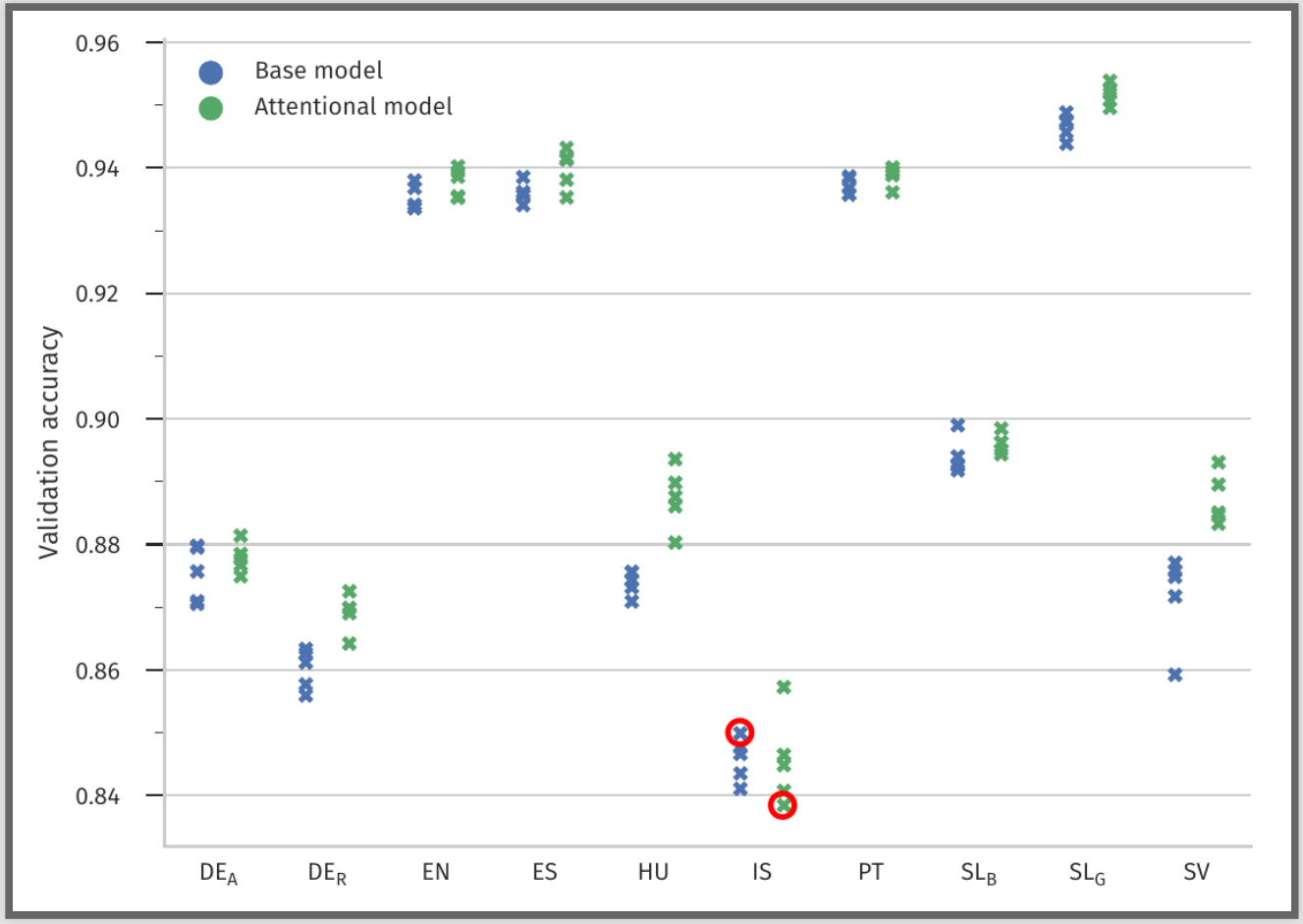
Be careful of variance!



Be careful of variance!



Be careful of variance!



More improvements...

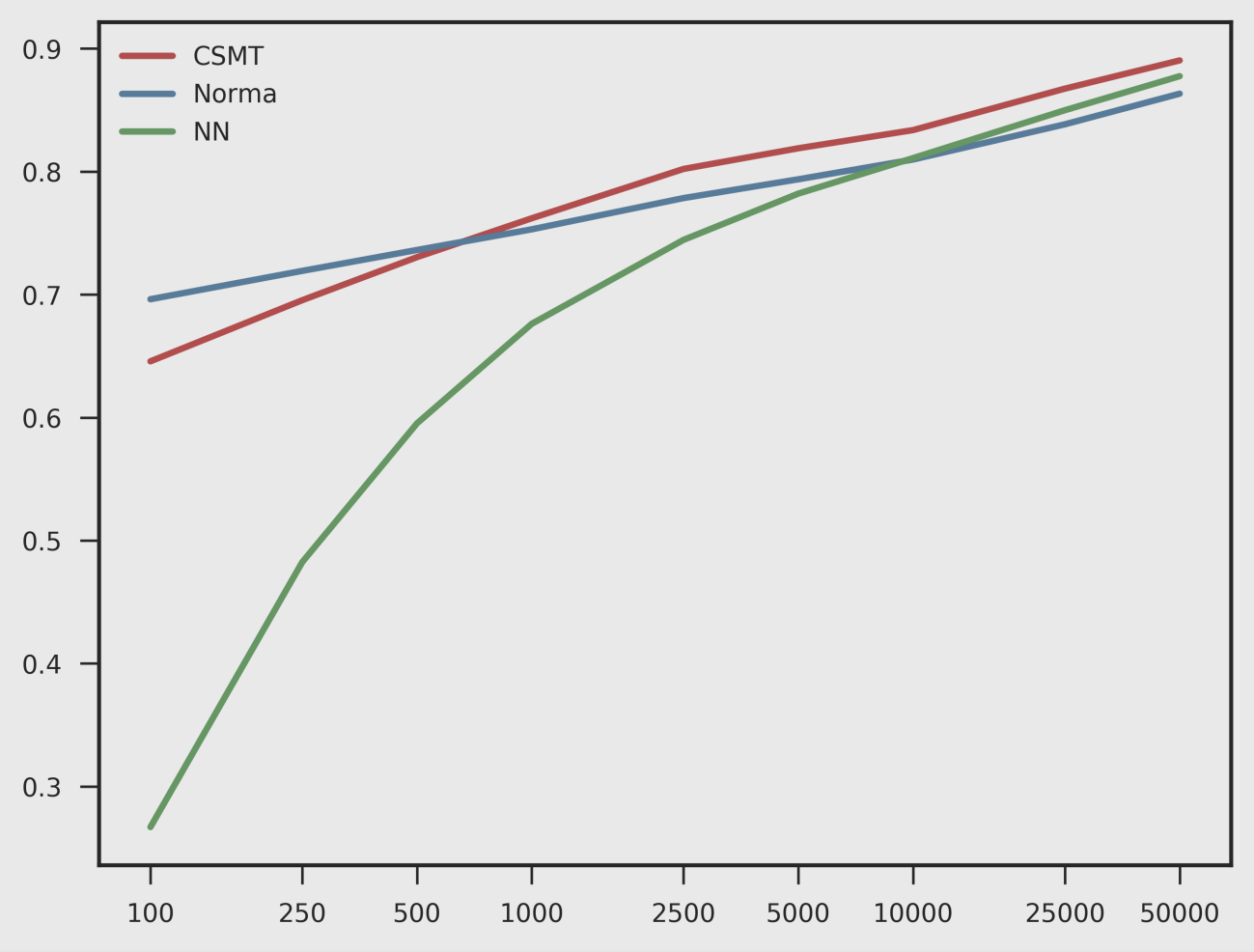
- **Attention** mechanism — *helps often* ✔
- **Beam search** — *helps almost always* ✔
- **Ensemble** of five models — *helps a lot* ✔✔
- **Dictionary filtering** — *helps only sometimes* ✖

Evaluation

	Method	Accuracy	Best on...
	Norma	89.97%	
⭐	CSMT	91.97%	DE_R, EN, ES, HU, IS, PT, SL_B, SL_G, SV
☹	NN	91.42%	DE_A

(Accuracy on dev sets, macro-averaged over all datasets)

Learning curves



Error classification

	Original	Correct	Prediction
Valid	kingis	king’s	king's
Good	wyse	ways	wise
Fair	meynteigne	maintain	meintain
Bad	t'acertaine	to ascertain	trace taint

Most errors fall into Good category!

More “good” cases

	Original	Correct	Prediction
DE_A	chu̇ff	küsse	kuss
EN	recomaundehyde	recommended	recommend
ES	enbie	envíe	envié
HU	yduewzewlendewk	üdvözülendőek	üdvözülendők

Character error rate (CER)

- CER is a very crude measure
  - *envíe — envié:* 2/5
  - *envíe — envxq:* 2/5
- CER correlates strongly with accuracy

Character error rate (CER)

- CER_I: evaluate CER on *incorrect* predictions only

Method	Accuracy	CER_I
Norma	89.97%	0.415
CSMT	**91.97%**	0.399
NN	91.42%	**0.384**

Stemming

	Correct	Prediction	Stem	
DE_A	küsse	kuss	kuss	✓
EN	recommended	recommend	recommend	✓
ES	envíe	envié	envi	✓
ES	envíe	envqx	envqx	✗
HU	üdvözülendőek	üdvözülendők	üdvözülendő	✓

Takeaways

Normalization ≠ Normalization

- Are there **normalization guidelines?**
- How do they handle...
  - ...inflectional mismatches?
  - ...morphosyntactic changes?
  - ...archaic/extinct lexemes?
  - ...tokenization?
  - ...proper nouns?
- **Be aware** of the data!

Practical recommendations

- **Norma** — *with little annotated data*

  https://github.com/comphist/norma

  - Annotated data always helps, even if it's 100 tokens!

- **cSMTiser**

  https://github.com/clarinsi/csmtiser

Are neural networks really inferior?

- Tang et al. (2018) report the **opposite result!**
- Many conceivable **improvements**
  - Multi-task learning (Bollmann et al., 2017; 2018)
  - Hard monotonic attention (Aharoni & Goldberg, 2017)
- Little work on **contextual** normalization

i thanke yow as hertly as i can

i thank you as heartily as i can

marcel@di.ku.dk

mmbollmann

https://marcel.bollmann.me/