



# Encoder-decoder methods for text normalization

## Swiss German WhatsApp messages



Massimo Lusetti<sup>1,3</sup> Tatyana Ruzsics<sup>2</sup>  
Tanja Samardžić<sup>2</sup> and Elisabeth Stark<sup>1</sup>

Anne Göhring<sup>1,3</sup>

<sup>1</sup>Department of Romance Studies

<sup>2</sup>Language and Space Lab

<sup>3</sup>Institute of Computational Linguistics



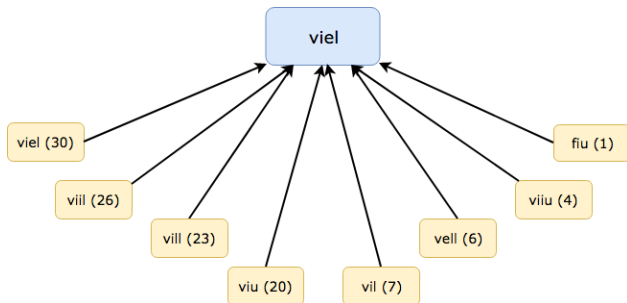
## Why text normalization?

- Historical texts
- Dialects and languages with no orthographic standard
- Computer-mediated communication (CMC)



## Variation in the source text: the dialect factor

Many different spelling variants for the same source word.





## Variation in the source text: the CMC factor

- **Vowel reduplication**  
*gaaaanz* → *ganz*      'all, whole'
- **Omitted vowels**  
*vllt* → *vielleicht*      'maybe'
- **Typos, unconventional casing and abbreviations**



## State of the art

Text normalization as a machine translation task.

Statistical Machine Translation (SMT)

vs.

Neural Machine Translation (NMT)



## The corpora: composition

	<b>Corpus</b>	<b>Messages</b>	<b>Alignment units</b>
<b>WUS</b>	What's up, Switzerland? <sup>1</sup>	5,345	54,229
<b>SMS</b>	Swiss SMS corpus <sup>2</sup>	10,674	262,494

<sup>1</sup>Corpus: [Stark et al., 2014](#); Documentation: [Ueberwasser and Stark, 2017](#)

<sup>2</sup>Corpus: [Stark et al., 2015](#); Documentation: [Ueberwasser, 2015](#)



## The corpus: alignment units

- **one-to-one**  
*hüt* → *heute*      'today'
- **one-to-many**  
*hämmers* → *haben wir es*      'have we ... it'
- **many-to-one**  
*aweg riise* → *wegreissen*      'tear away'
- **many-to-many**  
*über chunts* → *überkommt es*      'receives it'



## Methods: baseline

### Baseline

For each word in the test set we choose its most frequent normalization form in the training set. Unseen words are copied.

	<b>Baseline</b>	
<b>Word category</b>	<b>Proportion</b>	<b>Accuracy</b>
Unique	55.05	98.43
Ambiguous	32.70	81.22
New	12.25	30.27
Total	100	84.45





## Methods: character-level approach

1. Statistical Machine Translation (cSMT)
2. Neural Machine Translation (cNMT)
  - Formally similar word pairs  
*Sunne* → *Sonne*    'sun'
  - Regular transformation patterns (e.g., *ii* → *ei*)  
*Ziit* → *Zeit*    'time'  
*wiiter* → *weiter*    'further'  
*Priis* → *Preis*    'price'



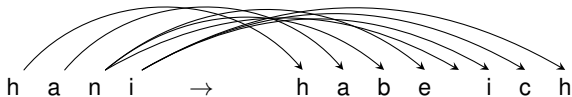
## Methods: character-level approach

1. Statistical Machine Translation (cSMT)
  2. Neural Machine Translation (cNMT)
- **Formally similar word pairs**  
*Sunne* → *Sonne*     'sun'
  - **Regular transformation patterns** (e.g., *ii* → *ei*)  
*Ziit* → *Zeit*     'time'  
*wiiter* → *weiter*     'further'  
*Priis* → *Preis*     'price'

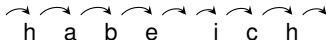


## cSMT

Translation model:  $p(\textit{normalised}|\textit{source})$



Language model:  $p(\textit{normalised}_i|\textit{normalised}_{i-1})$

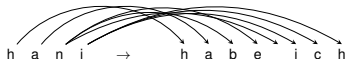




## cSMT vs cNMT

### cSMT:

TM:  $p(\textit{normalised}|\textit{source})$

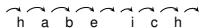


### cNMT:



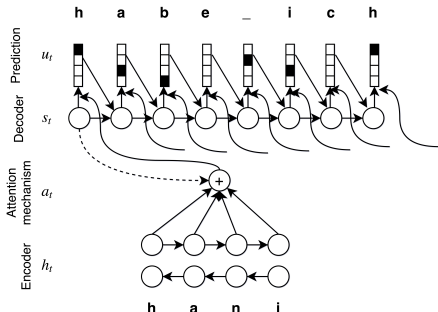
LM:

$p(\textit{normalised}_i|\textit{normalised}_{i-1})$





## cNMT



- LSTM decoder & bidirectional encoder  
Hochreiter and Schmidhuber, 1997,  
Sutskever et al., 2014
- Soft Attention  
Bahdanau et al., 2014

$$p(u_t) = f(s_t, c_t)$$

$$c_t = \sum_i a_t^i h_i$$

$$a_t = \phi(s_t, h_{1:N})$$



## cNMT mistakes

cNMT search for *'idere'*:

||| ideren ||| nmt= -4.341753 - *NMT prediction*  
||| in dieser ||| nmt= -29.94 - *correct form*  
||| ider einer ||| nmt= -30.29

- Word *'ideren'* does not appear in the target side of the corpus while *'in'* and *'dieser'* do

→ give more signal about word frequencies and rerank?

→ combine cNMT scores over chars with LM scores over words



## cNMT mistakes

cNMT search for *'idere'*:

||| ideren ||| nmt= -4.341753 - *NMT prediction*  
||| in dieser ||| nmt= -29.94 - *correct form*  
||| ider einer ||| nmt= -30.29

- Word *'ideren'* does not appear in the target side of the corpus while *'in'* and *'dieser'* do
- give more signal about word frequencies and rerank?
- combine cNMT scores over chars with LM scores over words



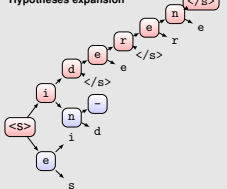
# cNMT + LM: synchronized decoding

Ruzsics and Samardžić, 2017

idere → in dieser

Iteration 1:

Hypotheses expansion

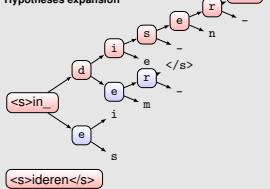


Scores Synchronization

Hypothesis	NMT score	Combined score
idere</s>	-4	-5
in_	-8	-2

Iteration 2:

Hypotheses expansion



Scores Synchronization

Hypothesis	cED score	Combined score
in_dieser</s>	-30	-3
in_der</s>	-35	-9
idere</s>	-4	-5

- Hypos generation on char level
- Scores are combined at word boundaries (synchronization)
- char-level NMT scores
- word-level LM scores
- MERT optimization for weights





## CNMT + LM: experimental setup

### LM training settings:

- REUSE:

cNMT + LM\_wus :word

- ADD:

cSMT + LM\_sms:char

cNMT + LM\_sms:char

- REUSE+ADD:

cNMT + LM\_wus+sms:word

### Corpus:

- Original

emojiQfaceWithTearsOfJoy → emojiQfaceWithTearsOfJoy

- Modified





## Results

System		Corpus	
		Original	Modified
REUSE+ADD	cNMT + LM_wus+sms:word	87.61	87.07
ADD	cNMT + LM_sms:char	87.27	86.55
REUSE	cNMT + LM_wus :word	87.14	86.55
	cNMT (ensemble 5)	87.03	86.50
ADD	cSMT + LM_sms:char	86.35	86.43
	cSMT	85.30	85.85
	Baseline	84.45	84.45

Table: Text normalization accuracy scores.



## Results: error analysis

Group	Source	cSMT Error	cNMT and Reference	English gloss
dissimilar reference	hb	hb	hauptbahnhof	'train station'
	òmu	ömel	jedenfalls	'anyway'
	gäbig	gebung	praktisch	'useful'
	eig	eig	eigentlich	'actually'
foreign word	feedback	feichtback	feedback	'feedback'
	cream	kream	cream	'cream'
> 1 target word	hanise	hanise	habe ich sie	'have I ... her'
	meini	meine	meine ich	'I mean'
	nonig	nonung	noch nicht	'not yet'
	söuis	sollte es	soll ich es	'shall I ... it'

Table: Improvements of the cNMT over cSMT.



## Conclusion

- Neural frameworks can achieve state-of-the-art performance in the task of text normalization, when the training corpus is of limited size.
- This can be achieved by exploiting their flexibility in integrating additional components, such as language models trained on different tokenization levels.



## Future work

- include context
- learn the predictor weights as a part of NN
- introduce copy for unseen characters



Thank you for your attention!



## References I

- ▶ Bahdanau, D., Cho, K., and Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *CoRR*, abs/1409.0473.
- ▶ Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. 9:1735–80.
- ▶ Ruzsics, T. and Samardžić, T. (2017). Neural sequence-to-sequence learning of internal word structure. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 184–194, Vancouver, Canada. Association for Computational Linguistics.
- ▶ Stark, E., Ueberwasser, S., and Göhring, A. (2014). Corpus "What's up, Switzerland?". University of Zurich. [www.whatsup-switzerland.ch](http://www.whatsup-switzerland.ch).
- ▶ Stark, E., Ueberwasser, S., and Ruef, B. (2009–2015). Swiss SMS Corpus. University of Zurich. <https://sms.linguistik.uzh.ch>.
- ▶ Sutskever, I., Vinyals, O., and Le, Q. V. (2014). Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 3104–3112.
- ▶ Ueberwasser, S. (2015). The Swiss SMS Corpus. Documentation, facts and figures. <https://sms.linguistik.uzh.ch>.
- ▶ Ueberwasser, S. and Stark, E. (2017). What's up, Switzerland? A corpus-based research project in a multilingual country. *Linguistik Online*, 84(5).