Neural Text Normalization with Adapted Decoding and POS Features

Swiss German WhatsApp messages

Tatyana Ruzsics[1] Massimo Lusetti[1,2] Anne Göhring[3]
Tanja Samardžić[1] and Elisabeth Stark[2]

[1]Language and Space Lab

[2]Department of Romance Studies

[3]Institute of Computational Linguistics

Why text normalization?

Historical texts

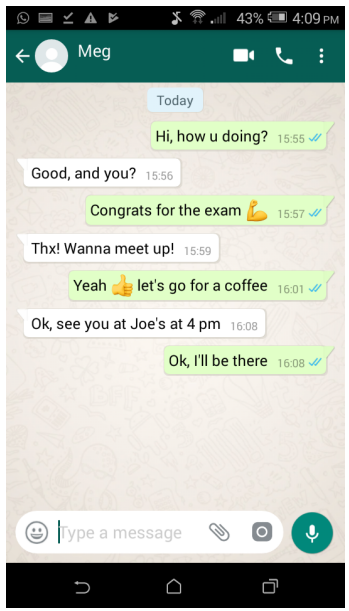
Sprooche [ändere | Quälltäxt bearbeite]

D Schwyz het luut Bündesverfassig vier Landes- ùn Amtssprooch
mit 22,8 Prozänt, Italiänisch mit 8,4 Prozänt ùn Rätoromanisch m
als Institution: Di meischte Kantön hen dergäge nùmme ei Amtss
zweisproochig (Französisch-Dütsch) ùn de Kanton Graubünda is
Au di meischte Gmeie hen nùmme ei offizielli Sprooch, au wänn
eso gmäss em Territorialitätsprinzip). D Sproochgränze sin mee
d Sproochgränz syt lengrem zuungùnschte vùm Rätoromanische

Minderheitesprooche, wo territorial nit bùnde sin, sin no s Jenisch
Bevölcherig e Sprooch, wo nit zue de Landessprooche ghört. D S
Änglisch ùn Türkisch hen alli mee Sprecher wie d Landessprooch
de Schwyz e Gebärdesprooch, je nooch Region di Dütschschwyz
Anerkannt isch d Gebärdesprooch aber nùmme i de Kantön Züri

Mit Ussnaam vo de Romandie wird d Sproochsituation in de Schw
entweder d Standardsprooch (Hochdütsch, Standarditaliänisch o
gschwätzt. Am kompliziertesche isch d Situation bi de Rätoroma
überregional Rumantsch Grischun, wo aber nit überall akzeptiert

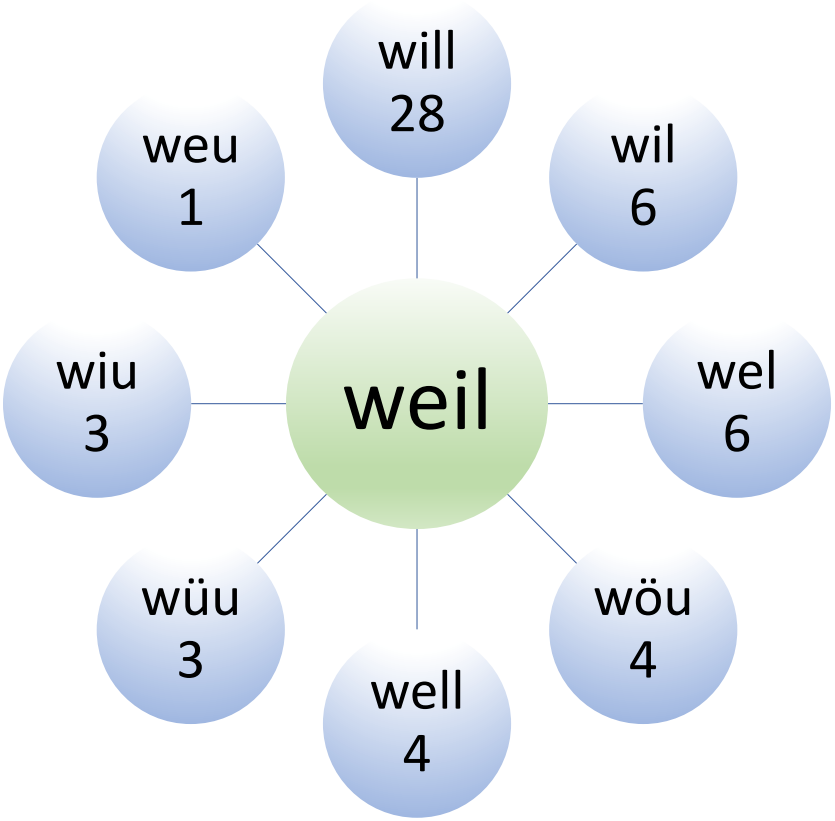
Dialects and languages with no orthographic standard



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

- **Emojis and emoticons**
- **Vowel reduplication**
  aaaaber → *aber* 'but'
- **Uncoventional abbreviations**
  eig → *eigentlich* 'actually'
- **Typos**
  wicbtig → *wichtig* 'important'
- **Unconventional casing**
  sach → *Sache* 'thing'

Methods

- Plain character-level neural architecture (NMT)
- Enhancements:
  1. Integrated word-level language model (NMT + LM)
  2. POS features as input (NMT + POS)

The corpora: composition

Corpus		Messages	Alignment units
WUS	What's up, Switzerland?[1]	5,345	54,202
SMS	Swiss SMS corpus[2]	10,674	262,494

[1] Corpus: Stark et al., 2014; Documentation: Ueberwasser and Stark, 2017
[2] Corpus: Stark et al., 2015; Documentation: Ueberwasser, 2015

POS tagging

Source	Target
Mir	Wir
sind	sind
ja	ja
nöd	nicht
dehei	daheim
,	,
wie	wie
cha	kann
da	da
de	der
schlüssel	Schlüssel
stecke	stecken
?	?
Etz	Jetzt
simmer	sind wir
kaputt	kaputt
Ja	Ja
scho	schon
ziemli	ziemlich

POS tagging

Source	Target	
Mir	Wir	PPER
sind	sind	VAFIN
ja	ja	ADV
nöd	nicht	PTKNEG
dehei	daheim	ADV
,	,	$,
wie	wie	KOUS
cha	kann	VMFIN
da	da	ADV
de	der	ART
schlüssel	Schlüssel	NN
stecke	stecken	VVINF
?	?	$.
Etz	Jetzt	ADV
simmer	sind wir	VAFIN+PPER
kaputt	kaputt	ADJD
Ja	Ja	ADV
scho	schon	ADV
ziemli	ziemlich	ADV

Creation of a silver standard through POS tagging of the normalized forms with the TreeTagger (Schmid, 1994) trained on the SMS corpus.

POS tagging

Source		Target	
Mir	PPER	Wir	PPER
sind	VAFIN	sind	VAFIN
ja	ADV	ja	ADV
nöd	PTKNEG	nicht	PTKNEG
dehei	ADV	daheim	ADV
,	$,	,	$,
wie	KOUS	wie	KOUS
cha	VMFIN	kann	VMFIN
da	ADV	da	ADV
de	ART	der	ART
schlüssel	NN	Schlüssel	NN
stecke	VVINF	stecken	VVINF
?	$.	?	$.
Etz	ADV	Jetzt	ADV
simmer	VAFIN+PPER	sind wir	VAFIN+PPER
kaputt	ADJD	kaputt	ADJD
Ja	ADV	Ja	ADV
scho	ADV	schon	ADV
ziemli	ADV	ziemlich	ADV

Projection of POS tags from normalized forms (target side) to Swiss German forms (source side)

POS tagging

Source

Mir	PPER
sind	VAFIN
ja	ADV
nöd	PTKNEG
dehei	ADV
,	$,
wie	KOUS
cha	VMFIN
da	ADV
de	ART
schlüssel	NN
stecke	VVINF
?	$.

Etz
simmer
kaputt

Ja
scho
ziemli

Split the corpus into:

- Train set
- Development set
- Test set

Train the BTagger
(Gesmundo and Samardžić, 2012)
on the train set

POS tagging

Source

Mir	PPER
sind	VAFIN
ja	ADV
nöd	PTKNEG
dehei	ADV
,	$,
wie	KOUS
cha	VMFIN
da	ADV
de	ART
schlüssel	NN
stecke	VVINF
?	$.

Etz	ADV
simmer	VAFIN+PPER
kaputt	ADJD

Ja	ADV
scho	ADJD
ziemli	ADV

Tag the development and test set with the trained BTagger

Evaluate POS tagging performance by comparison with the silver standard:

- Dev set: 90.67%
- Test set: 90.30%

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.

	Baseline	
Word category	**Proportion**	**Accuracy**
Unique	53.82	98.16
Ambiguous	34.09	80.40
New	12.09	28.85
Total	100	83.72

Methods: baseline + POS

	Baseline + POS	
Word category	**Proportion**	**Accuracy**
Unique	53.82	98.16
Ambiguous:	34.09	86.04
- POS unambiguous	21.24	91.22
- POS ambiguous	12.85	77.47
New	12.09	28.85
Total	100	85.64

NMT

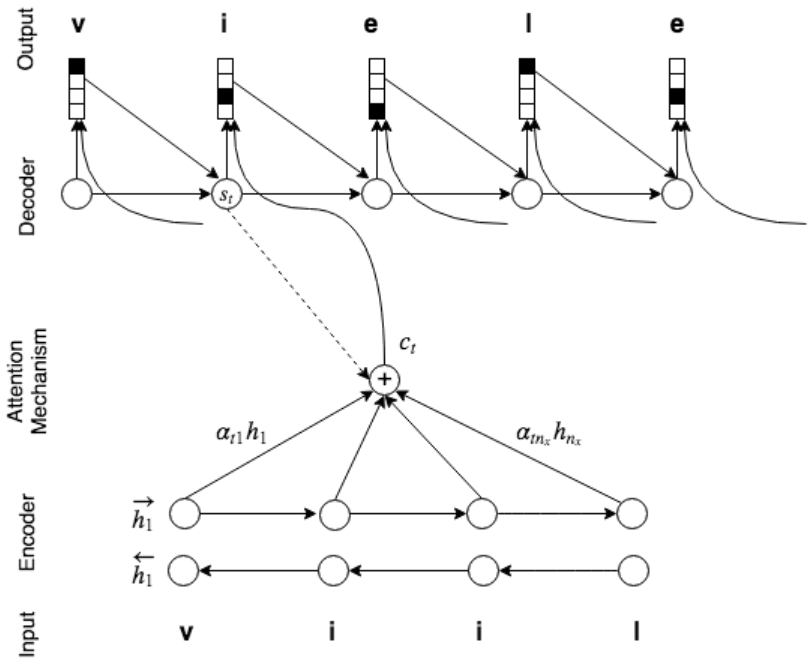


Figure: Basic NMT architecture.

- LSTM decoder & bidirectional encoder (Hochreiter and Schmidhuber, 1997, Sutskever et al., 2014)
- Soft Attention (Bahdanau et al., 2014, Luong et al., 2015)

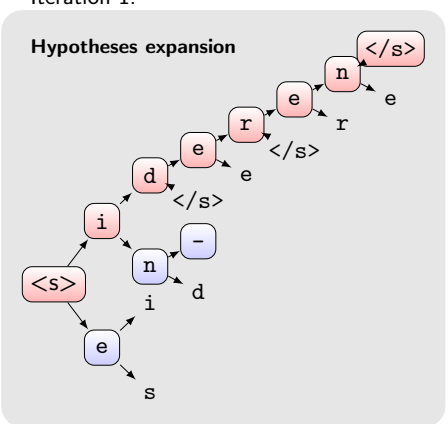
$$p(u_t) = f(s_t, c_t)$$
$$c_t = \sum_i a_t^i h_i$$
$$a_t = \phi(s_t, h_{1:N})$$

NMT + LM: synchronized decoding

Ruzsics and Samardžić, 2017

idere → in dieser

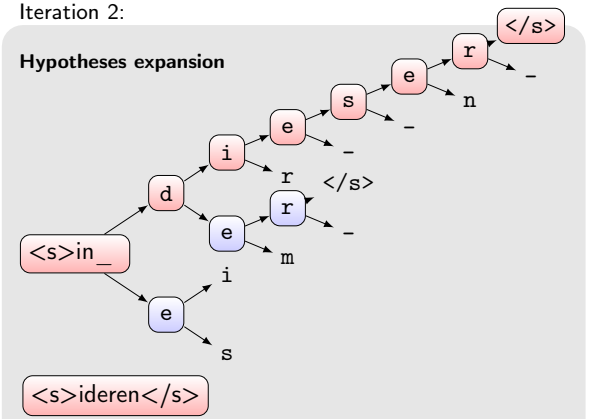
Iteration 1:



Synchronization scores

Hypothesis	NMT score	NMT+LM score
ideren</s>	**-4**	-5
in_	-8	**-2**

Iteration 2:



Synchronization scores

Hypothesis	NMT score	NMT+LM score
in_dieser</s>	-30	**-3**
in_der</s>	-35	-9
ideren</s>	**-4**	-5

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

NMT with POS

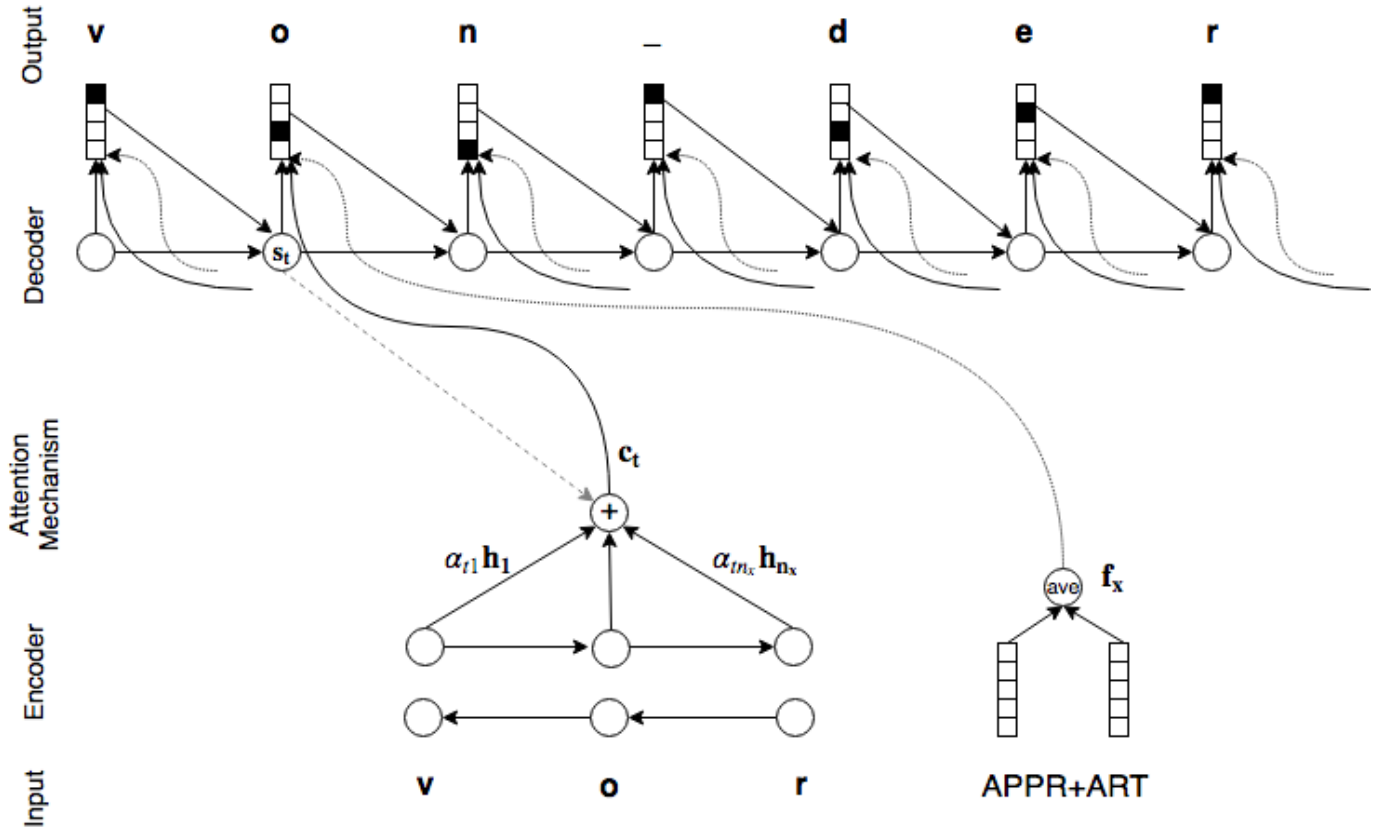


Figure: NMT with POS tags as features.

Results

	Setting	Accuracy (%)
Without POS	Baseline	83.72
	CSMT + LMwus+sms:char[3]	86.46
	NMT	86.81
	NMT + LMwus+sms:word[4]	87.09
With POS	Baseline + POS	85.64
	NMT + POS	89.13
	NMT + POS + LMwus+sms:word	**89.53**

Table: Text normalization accuracy scores.

[3] LMwus+sms:char : language model trained over characters on the target side of the WUS corpus extended with the target side of the SMS corpus.

[4] LMwus+sms:word : language model trained over words on the target side of the WUS corpus extended with the target side of the SMS corpus.

Error analysis: integrated LM

	Input word	Normalization		English gloss	Gold seen	
		NMT	NMT+LM and Gold		in WUS?	in SMS?
a)	schwizer	schwizer	schweizer	Swiss	yes	-
	schi	schei	schon	already	yes	-
	aver	aver	aber	but	yes	-
b)	kömmer	kommer	können wir	we can	yes	-
	hanie	habeie	habe ich	I have	yes	-
	hanise	habe ise	habe ich sie	I have her	yes	-
c)	trurig	trurig	traurig	sad	no	yes
	usfüerige	ausfürigen	ausführungen	executions	no	yes
	gschune	geschune	geschienen	has seemed	no	yes

Table: Errors of NMT in the NEW words category corrected by NMT+LM, when the normalized form has been seen in the target side of the train set.

Error analysis: POS features

#[5]	source	NMT+LM	NMT+POS and Gold	English gloss	POS pred.[6]	POS alt[7]
14	bi	bin	bei	by	APPR	VAFIN
1	cho	gekommen	kommen	come	VVINF	VVPP
6	dass	dass	das	that	PDS	KOUS
19	de	dann	der	the	ART	ADV
6	di	die	dich	you (obj.)	PPER	ART
12	es	ein	es	it	PPER	ART
4	i	ich	in	in	APPR	PPER
7	s	das	es	it	PPER	ART
5	si	sie	sind	are	VAFIN	PPER
2	wenn	wenn	wann	when	PWAV	KOUS

Table: Disambiguation due to POS tag features.

[5] Number of occurrences in the test set.

[6] The POS tag that has been assigned to the test set instance by the BTagger, which in these examples is the same as the POS in the silver standard.

[7] An alternative POS tag for the test set instance, which in the silver standard is associated to the wrong NMT+LM prediction.

Conclusion

The two approaches we implement are complementary and result in the improvement of the neural models:

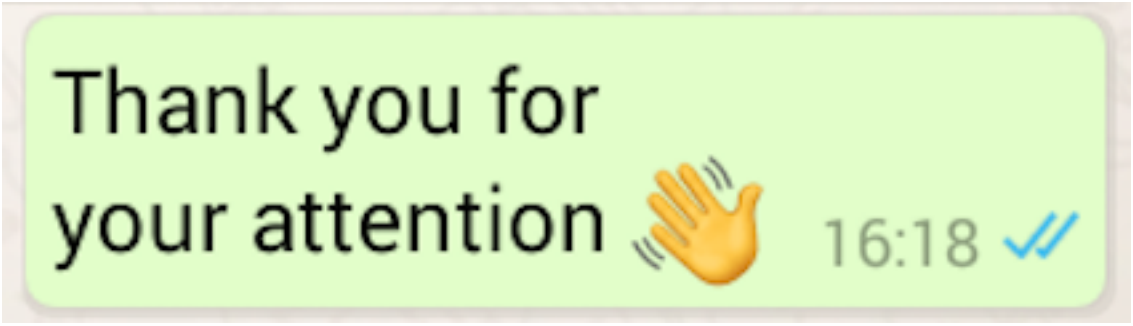
1. The **integrated word-level LM** helps to improve performance:
   - By introducing word awareness in the character-level framework
   - On unseen input words, by using additional language models
2. **POS tag** information helps to improve performance on ambiguous words, i.e. words with different possible normalization forms.

Future work

1. Use our method for the normalization of other languages characterized by dialect variation.

2. Increase the amount of target data in order to address unseen input words.

3. Tackle the problem of ambiguity:
   - Use a more fine-grained tagset.
   - Use context information more directly by including the neighboring words within the sentence boundaries in the neural system.


Thank you for
your attention
16:18

References I

- Bahdanau, D., Cho, K., and Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *CoRR*, abs/1409.0473.
- Gesmundo, A. and Samardžić, T. (2012). Lemmatisation as a tagging task. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 368–372, Jeju Island, Korea. Association for Computational Linguistics.
- Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. 9:1735–80.
- Luong, T., Pham, H., and Manning, C. D. (2015). Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1421, Lisbon, Portugal. Association for Computational Linguistics.
- 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.
- Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. In *International Conference on New Methods in Language Processing*, pages 44–49, Manchester, UK.
- Stark, E., Ueberwasser, S., and Göhring, A. (2014). Corpus "What's up, Switzerland?". University of Zurich. www.whatsup-switzerland.ch.
- Stark, E., Ueberwasser, S., and Ruef, B. (2009–2015). Swiss SMS Corpus. University of Zurich. https://sms.linguistik.uzh.ch.

References II

- 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).