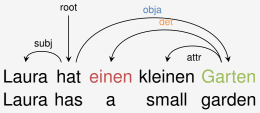
Neural Machine Translation
what's linguistics got to do with it?

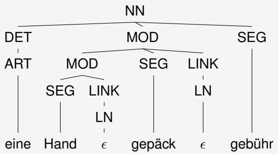
Rico Sennrich

University of Edinburgh

research trend: more linguistics for statistical machine translation



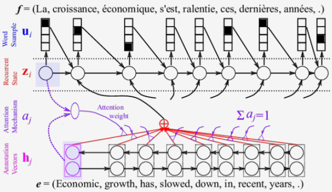
syntax-based LM
[Sennrich, TACL 2015]



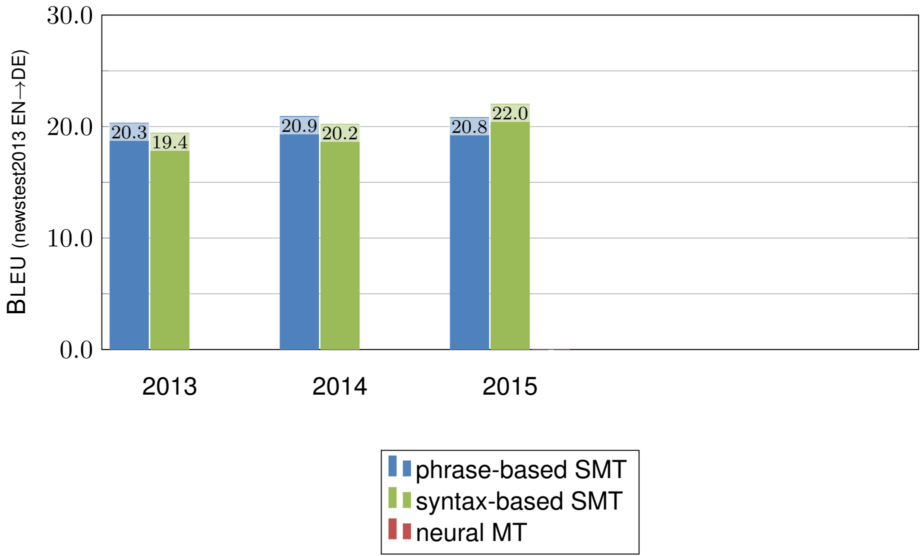
morphological structure
[Sennrich, Haddow, EMNLP 2015]

a new challenger appears: neural machine translation

- requires minimal domain knowledge
- similar models used for speech and computer vision

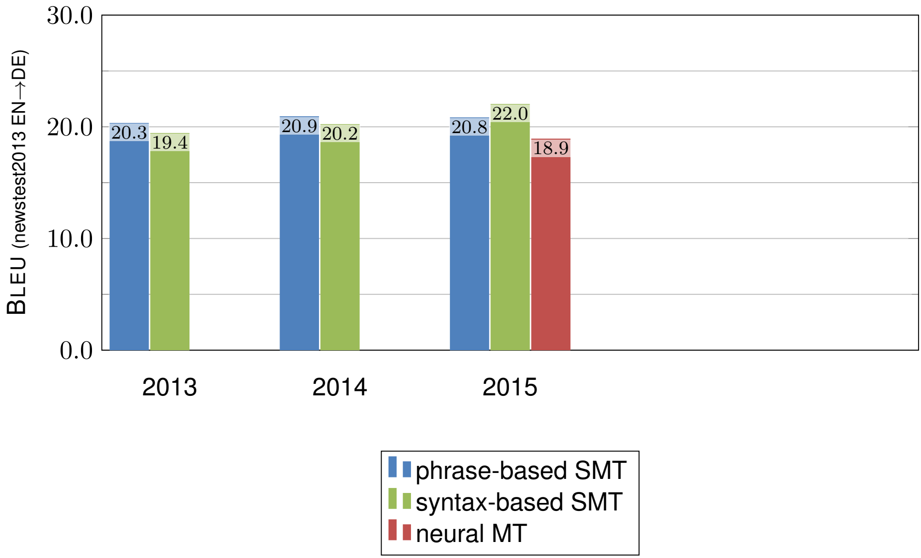


Edinburgh's* WMT Results over the Years



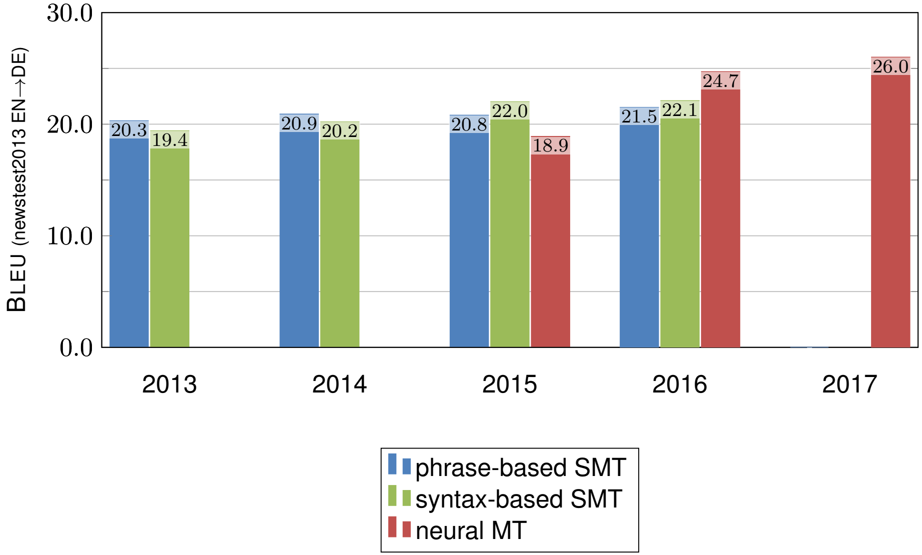
*NMT 2015 from U. Montréal: `https://sites.google.com/site/acl16nmt/`

Edinburgh's* WMT Results over the Years



*NMT 2015 from U. Montréal: `https://sites.google.com/site/acl16nmt/`

Edinburgh's* WMT Results over the Years



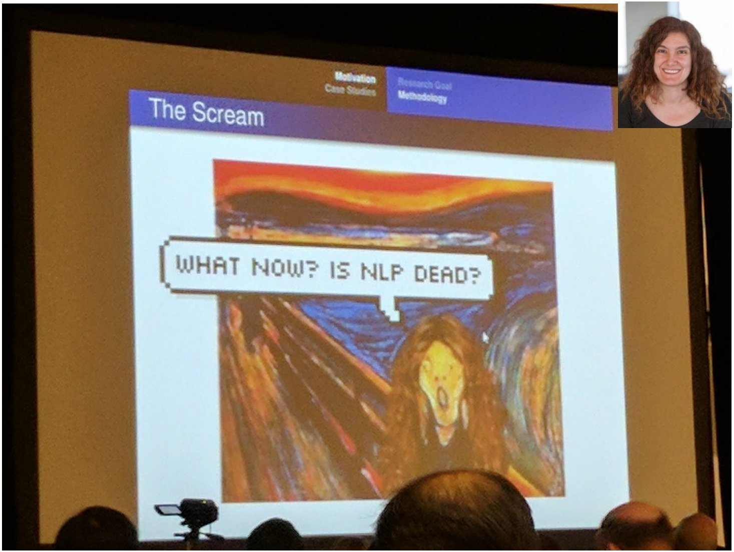
*NMT 2015 from U. Montréal: `https://sites.google.com/site/acl16nmt/`

What Now?

do we still need linguistics for MT?

What Now?

do we still need linguistics for MT?



What Now?

do we still need linguistics for MT?

Today's Talk

case studies on how linguistics is helping neural MT research

- linguistically motivated (but non-linguistic) models
- targeted evaluation of neural MT
- linguistically informed models

1. Linguistically Motivated (but Non-Linguistic) Models
2. Targeted Evaluation of Neural MT
3. Linguistically Informed Models

Open-Vocabulary Neural MT

problem

word-level neural networks use one-hot encoding
$\rightarrow$ closed and small vocabulary

this gets you 95% of the way...
... if you only care about automatic metrics

Open-Vocabulary Neural MT

problem

word-level neural networks use one-hot encoding
$\rightarrow$ closed and small vocabulary

this gets you 95% of the way...
... if you only care about automatic metrics

why 95% is not enough

rare outcomes have high self-information

source	The **indoor temperature** is very pleasant.	
reference	Das **Raumklima** ist sehr angenehm.	
[Bahdanau et al., 2015]	Die **UNK** ist sehr angenehm.	✗

Open-Vocabulary Neural MT

problem

word-level neural networks use one-hot encoding
$\rightarrow$ closed and small vocabulary

this gets you 95% of the way...
... if you only care about automatic metrics

why 95% is not enough

rare outcomes have high self-information

source	The **indoor temperature** is very pleasant.	
reference	Das **Raumklima** ist sehr angenehm.	
[Bahdanau et al., 2015]	Die **UNK** ist sehr angenehm.	✗
[Jean et al., 2015]	Die **Innenpool** ist sehr angenehm.	✗

Open-Vocabulary Neural MT

problem

word-level neural networks use one-hot encoding
$\rightarrow$ closed and small vocabulary

this gets you 95% of the way...
... if you only care about automatic metrics

why 95% is not enough

rare outcomes have high self-information

source	The **indoor temperature** is very pleasant.	
reference	Das **Raumklima** ist sehr angenehm.	
[Bahdanau et al., 2015]	Die **UNK** ist sehr angenehm.	✗
[Jean et al., 2015]	Die **Innenpool** ist sehr angenehm.	✗
[**Sennrich**, Haddow, Birch, ACL 2016]	Die **Innen+ temperatur** ist sehr angenehm.	✓

Subword Neural MT [**Sennrich**, Haddow, Birch, ACL 2016]

linguistic motivation

- translation is open-vocabulary problem
- rare words matter
- morphological typology: 1-to-many translations are common
  $\rightarrow$ problem for backoff mechanism
- rare words are often morphologically complex and can be broken down into smaller units
  - solar system (English)
  - Sonnen|system (German)
  - Nap|rendszer (Hungarian)

Subword Neural MT

goal

subword segmentation that:

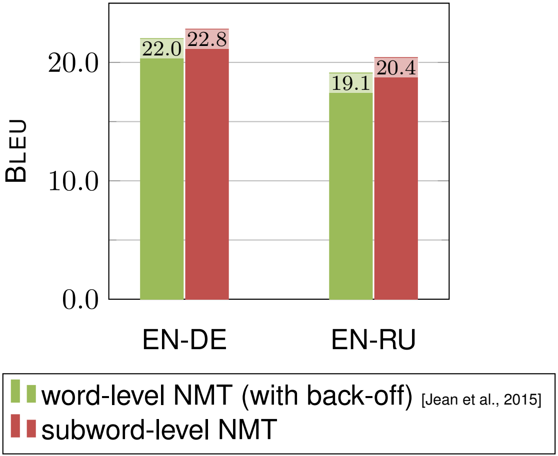
- uses a closed vocabulary of subword units
- can represent open vocabulary (including unknown words)
- minimizes the sequence length (given the vocabulary size)

solution

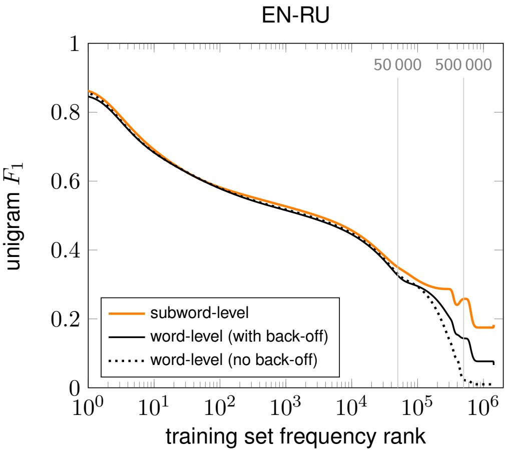
- greedy compression algorithm: byte pair encoding (BPE) [Gage, 1994]
- we adapt BPE to word segmentation
- hyperparameter: vocabulary size

vocabulary size	text
300	t+ h+ e i+ n+ d+ o+ o+ r t+ e+ m+ p+ e+ r+ a+ t+ u+ r+ e i+ s v+ e+ r+ y p+ l+ e+ a+ s+ a+ n+ t
1300	the in+ do+ or t+ em+ per+ at+ ure is very p+ le+ as+ ant
10300	the in+ door temper+ ature is very pleasant
50300	the indoor temperature is very pleasant

Subword NMT: Translation Quality

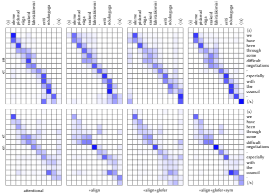


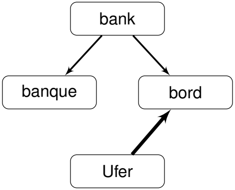
Subword NMT: Translation Quality



Linguistically Motivated Models

氵	(water)
河	river
湖	lake
海	sea





logographic input

[Costa-jussà et al., 2017]

[Cai and Dai, 2017]

structural alignment biases

[Cohn et al., 2016]

multi-source translation

[Zoph and Knight, 2016]

NMT: what's linguistics got to do with it?

1. Linguistically Motivated (but Non-Linguistic) Models
2. Targeted Evaluation of Neural MT
3. Linguistically Informed Models

What Hypotheses Do We Test?

hypothesis:	model A obtains higher BLEU than model B on data set X

What Hypotheses Do We Test?

hypothesis: | model A obtains higher BLEU than model B on data set X

What Hypotheses Do We Test?

hypothesis:	model A is better model of translation than model B
evidence:	model A obtains higher BLEU than model B on data set X

What Hypotheses Do We Test?

hypothesis:	model A is better model of translation than model B
evidence:	model A obtains higher BLEU than model B on data set X

What Hypotheses Do We Test?

hypothesis:	many languages have long-distance interactions. model A produces disfluent output because it models these interactions poorly. model B can better model long-distance interactions, and produces more fluent output.

What Hypotheses Do We Test?

hypothesis:	many languages have long-distance interactions. model A produces disfluent output because it models these interactions poorly. model B can better model long-distance interactions, and produces more fluent output.

What Hypotheses Do We Test?

	many languages have long-distance interactions.
hypothesis:	model A produces disfluent output because it models these interactions poorly.
	model B can better model long-distance interactions, and produces more fluent output.
evidence:	model A obtains higher BLEU than model B on data set X

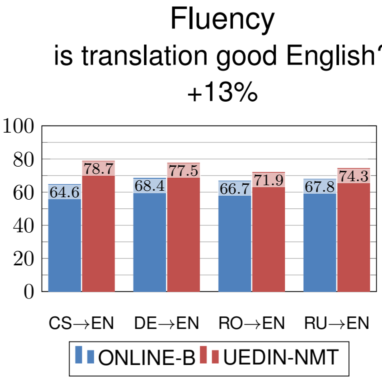
What Hypotheses Do We Test?

	many languages have long-distance interactions.
hypothesis:	model A produces disfluent output because it models these interactions poorly.
	model B can better model long-distance interactions, and produces more fluent output.
evidence:	model A obtains higher BLEU than model B on data set X

What Hypotheses Do We Test?

- being able to test our hypotheses is beauty of empirical NLP
- complex, interesting hypotheses need targeted evaluation
- I want to see more interesting hypotheses
  $\rightarrow$ we need more targeted evaluation

Human Evaluation of Neural MT [Bojar et al., 2016]



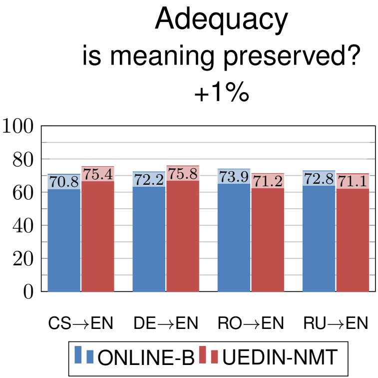


Figure: WMT16 direct assessment results

Human Evaluation in TraMOOC

[Castilho, Moorkens, Gaspari, Sennrich, Sosoni, Georgakopoulou, Lohar, Way, Miceli Barone, Gialama, MT Summit XVI, 2017]

- direct assessment of NMT (vs. PBSMT):
  - fluency: +10%
  - adequacy: +1%

Error Annotation

category	SMT	NMT	difference
inflectional morphology	2274	1799	**-21%**
word order	1098	691	**-37%**
omission	421	362	-14%
addition	314	265	-16%
mistranslation	1593	1552	-3%
"no issue"	449	788	**+75%**

Human Evaluation of Neural MT

Neural Machine Translation is very fluent.

Human Evaluation of Neural MT

~~Neural Machine Translation~~ is very fluent.

Attentional encoder-decoder with BPE segmentation and recurrent GRU decoder

Human Evaluation of Neural MT

~~Neural Machine Translation~~ is very fluent.

Attentional encoder-decoder with BPE segmentation and recurrent GRU decoder

what about...?

- character-level models [Lee et al., 2016]
- convolutional models [Gehring et al., 2017]
- models with self-attention [Vaswani et al., 2017]

Human Evaluation of Neural MT

~~Neural Machine Translation~~ is very fluent.

Attentional encoder-decoder with BPE segmentation and recurrent GRU decoder

what about...?

- character-level models [Lee et al., 2016]
- convolutional models [Gehring et al., 2017]
- models with self-attention [Vaswani et al., 2017]

Adequacy remains a major problem in Neural Machine Translation

Human Evaluation of Neural MT

~~Neural Machine Translation~~ is very fluent.

Attentional encoder-decoder with BPE segmentation and recurrent GRU decoder

what about...?

- character-level models [Lee et al., 2016]
- convolutional models [Gehring et al., 2017]
- models with self-attention [Vaswani et al., 2017]

Adequacy remains a major problem in Neural Machine Translation

...using a shallow NMT model at WMT 2016

Human Evaluation of Neural MT

~~Neural Machine Translation~~ is very fluent.

Attentional encoder-decoder with BPE segmentation and recurrent GRU decoder

what about...?

- character-level models [Lee et al., 2016]
- convolutional models [Gehring et al., 2017]
- models with self-attention [Vaswani et al., 2017]

Adequacy remains a major problem in Neural Machine Translation

...using a shallow NMT model at WMT 2016

how...?

- do we compare different architectures?
- do we measure improvement over time?

How to Assess Specific Aspects in MT?

- human evaluation
  - × costly; hard to compare to previous work
- automatic metrics (BLEU)
  - × too coarse; blind towards specific aspects

How to Assess Specific Aspects in MT?

- human evaluation
  $\times$ costly; hard to compare to previous work
- automatic metrics (BLEU)
  $\times$ too coarse; blind towards specific aspects

contrastive translation pairs

- NMT models assign probability to any translation
- binary classification task: which translation is better?
- choice between reference translation and contrastive variant
  $\rightarrow$ corrupted with single error of specific type
- $\approx$ minimal pairs in linguistics

Assessment with Contrastive Translation Pairs

workflow

- researcher wants to analyse difficult translation problem
- researcher predicts what errors NMT system might make
- researcher creates test set with correct translations and corrupted variants
- test set allows automatic, quantitative, and reproducible analysis of NMT model

example

Assessment with Contrastive Translation Pairs

workflow

- researcher wants to analyse difficult translation problem
- researcher predicts what errors NMT system might make
- researcher creates test set with correct translations and corrupted variants
- test set allows automatic, quantitative, and reproducible analysis of NMT model

example

- subject–verb agreement

Assessment with Contrastive Translation Pairs

workflow

- researcher wants to analyse difficult translation problem
- researcher predicts what errors NMT system might make
- researcher creates test set with correct translations and corrupted variants
- test set allows automatic, quantitative, and reproducible analysis of NMT model

example

- subject–verb agreement
- change grammatical number of verb to introduce agreement error

Assessment with Contrastive Translation Pairs

workflow

- researcher wants to analyse difficult translation problem
- researcher predicts what errors NMT system might make
- researcher creates test set with correct translations and corrupted variants
- test set allows automatic, quantitative, and reproducible analysis of NMT model

example

- subject–verb agreement
- change grammatical number of verb to introduce agreement error
- 35000 contrastive pairs created with simple linguistic rules

Contrastive Translation Pairs

	sentence	prob.
English	[...] that the **plan will** be approved	
German (correct)	[...], dass der **Plan** verabschiedet **wird**	0.1 ✓
German (contrastive)	* [...], dass der **Plan** verabschiedet **werden**	0.01

subject-verb agreement

LingEval97: A Test Set of Contrastive Translation Pairs

LingEval97

- 97 000 contrastive translation pairs
- based on English→German WMT test sets
- rule-based, automatic creation of errors
- 7 error types
- metadata for in-depth analysis:
  - error type
  - distance between words
  - word frequency in WMT15 training set

Case Study: Some Open Questions in Neural MT

Case Study: Some Open Questions in Neural MT

Kyunghyun Cho @kchonyc — Following

Fully char-level NMT! It works well on all four language pairs we've considered ({Cs, De, Ru, Fi}->En), and we... fb.me/1oRwyQvZD

RETWEETS 32 LIKES 83

9:12 AM - 11 Oct 2016

2 32 83

Emiel van Miltenburg @evanmiltenburg — Follow

@kchonyc Are there any benefits to using these models for longer dependencies?

1:16 PM - 11 Oct 2016

1

Case Study: Some Open Questions in Neural MT

Kyunghyun Cho
@kchonyc
Following

Fully char-level NMT! It works well on all four language pairs we've considered ({Cs, De, Ru, Fi}->En), and we... fb.me/1oRwyQvZD

RETWEETS 32 LIKES 83

9:12 AM - 11 Oct 2016

2 32 83

Emiel van Miltenburg
@evanmiltenburg
Follow

@kchonyc Are there any benefits to using these models for longer dependencies?

1:16 PM - 11 Oct 2016

1

Kyunghyun Cho
@kchonyc
Following

@evanmiltenburg ah well that's a difficult question!

1:30 PM - 11 Oct 2016

1

Case Study: Some Open Questions in Neural MT

Kyunghyun Cho @kchonyc — Following

Fully char-level NMT! It works well on all four language pairs we've considered ({Cs, De, Ru, Fi}->En), and we... fb.me/1oRwyQvZD

RETWEETS 32 LIKES 83

9:12 AM - 11 Oct 2016

2 32 83

Emiel van Miltenburg @evanmiltenburg — Follow

@kchonyc Are there any benefits to using these models for longer dependencies?

1:16 PM - 11 Oct 2016

1

Kyunghyun Cho @kchonyc — Following

@evanmiltenburg ah well that's a difficult question!

1:30 PM - 11 Oct 2016

1

text representation

word-level	but as the **example** of Mobilking in Poland **shows** |——— 5 steps ———|
subword-level (byte-pair encoding)	but as the **example** of Mobil+ king in Poland **shows** |——— 6 steps ———|
character-level	b u t _ a s _ t h e _ **e x a m p l e** _ o f _ M o b i l k i n g _ i n _ P o l a n d _ **s h o w s** |——— 29 steps ———|

Case Study: Some Open Questions in Neural MT

Kyunghyun Cho @kchonyc · Following

Fully char-level NMT! It works well on all four language pairs we've considered ({Cs, De, Ru, Fi}->En), and we... fb.me/1oRwyQvZD

RETWEETS 32 · LIKES 83

9:12 AM - 11 Oct 2016

2 · 32 · 83

Emiel van Miltenburg @evanmiltenburg · Follow

@kchonyc Are there any benefits to using these models for longer dependencies?

1:16 PM - 11 Oct 2016

1

Kyunghyun Cho @kchonyc · Following

@evanmiltenburg ah well that's a difficult question!

1:30 PM - 11 Oct 2016

1

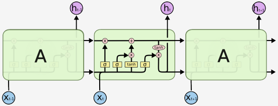
text representation

word-level	but as the **example** of UNK in Poland **shows** |——— 5 steps ———|
subword-level (byte-pair encoding)	but as the **example** of Mobil+ king in Poland **shows** |——— 6 steps ———|
character-level	b u t _ a s _ t h e _ **e x a m p l e** _ o f _ M o b i l k i n g _ i n _ P o l a n d _ **s h o w s** |——— 29 steps ———|

Case Study: Some Open Questions in Neural MT

does network architecture affect learning of long-distance dependencies?

architectures



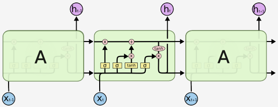
RNN vs. GRU vs. LSTM

Christopher Olah http://colah.github.io/posts/2015-08-Understanding-LSTMs/

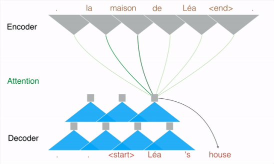
Case Study: Some Open Questions in Neural MT

does network architecture affect learning of long-distance dependencies?

architectures

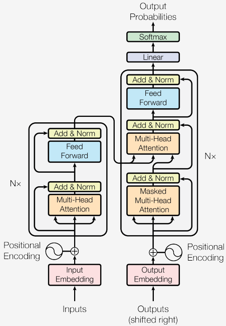


RNN vs. GRU vs. LSTM



(convolution)

[Gehring et al., 2017]

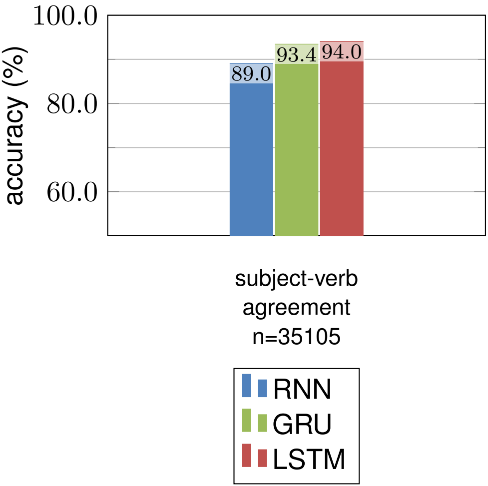


(self-attention)

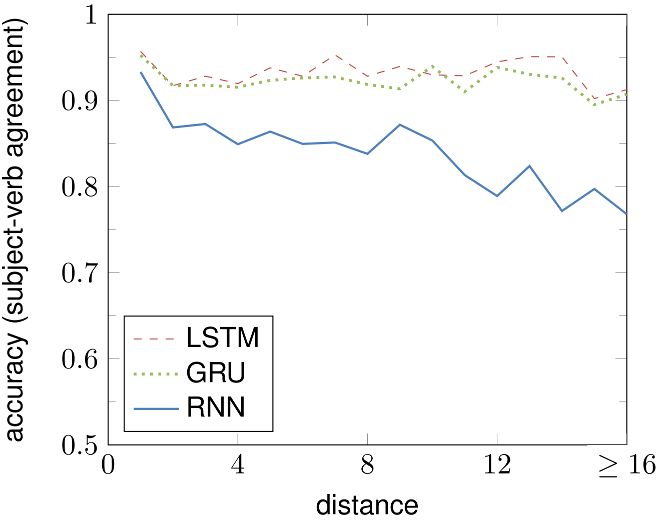
[Vaswani et al., 2017]

Christopher Olah http://colah.github.io/posts/2015-08-Understanding-LSTMs/

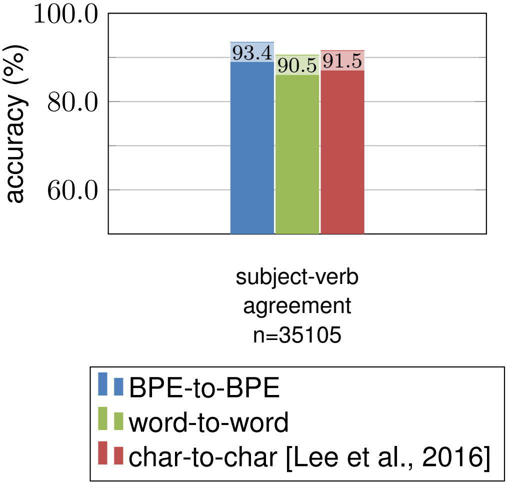
Results: Architecture



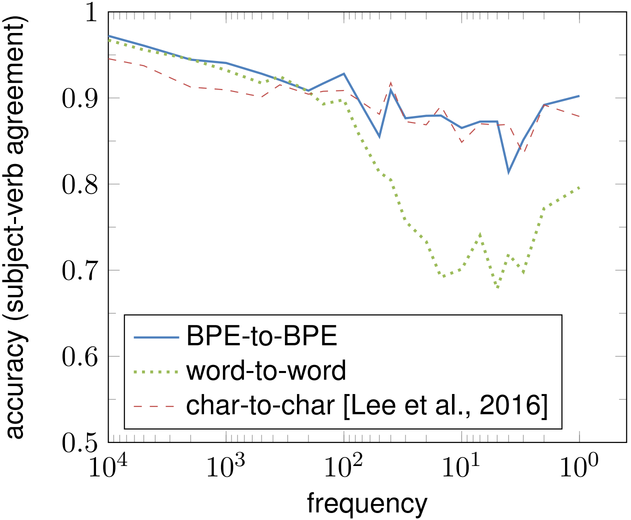
Results: Architecture



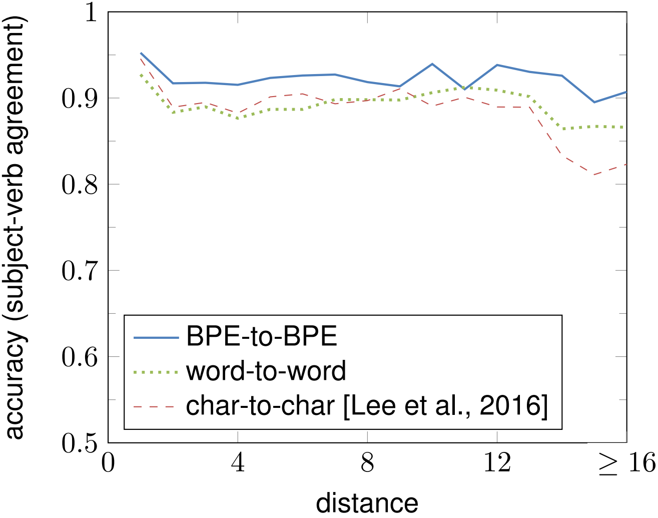
Results: Text Representation



Results: Text Representation



Results: Text Representation



What Did We Learn?

- method verifies strength of LSTM and GRU
  $\rightarrow$ future work: test of convolutional model and self-attention
- word-level model is poor for rare words
- character-level model is poor for long distances
- BPE subword segmentation is good compromise

Targeted Analysis: Adequacy

adequacy is open problem

system	sentence
source	Dort wurde er von dem **Schläger** und einer weiteren männl. Person erneut angegriffen.
reference	There he was attacked again by his **original attacker** and another male.
our NMT	There he was attacked again by the **racket** and another male person.
Google	There he was again attacked by the **bat** and another male person.

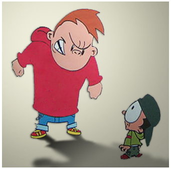
Schläger

Targeted Analysis: Adequacy

adequacy is open problem

system	sentence
source	Dort wurde er von dem **Schläger** und einer weiteren männl. Person erneut angegriffen.
reference	There he was attacked again by his **original attacker** and another male.
our NMT	There he was attacked again by the **racket** and another male person.
Google	There he was again attacked by the **bat** and another male person.

Schläger



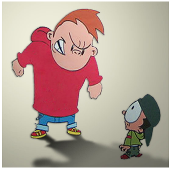
attacker

Targeted Analysis: Adequacy

adequacy is open problem

system	sentence
source	Dort wurde er von dem **Schläger** und einer weiteren männl. Person erneut angegriffen.
reference	There he was attacked again by his **original attacker** and another male.
our NMT	There he was attacked again by the **racket** and another male person.
Google	There he was again attacked by the **bat** and another male person.

Schläger



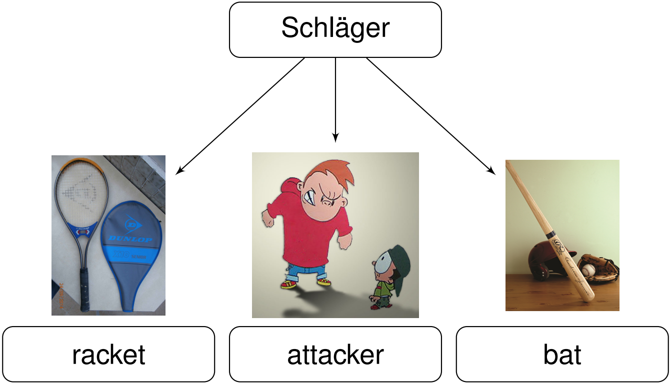
racket

attacker

Targeted Analysis: Adequacy

adequacy is open problem

system	sentence
source	Dort wurde er von dem **Schläger** und einer weiteren männl. Person erneut angegriffen.
reference	There he was attacked again by his **original attacker** and another male.
our NMT	There he was attacked again by the **racket** and another male person.
Google	There he was again attacked by the **bat** and another male person.

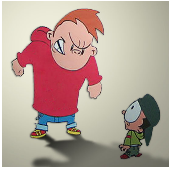


Targeted Analysis: Adequacy

adequacy is open problem

system	sentence
source	Dort wurde er von dem **Schläger** und einer weiteren männl. Person erneut angegriffen.
reference	There he was attacked again by his **original attacker** and another male.
our NMT	There he was attacked again by the **racket** and another male person.
Google	There he was again attacked by the **bat** and another male person.

Schläger



racket attacker bat

Targeted Analysis: Adequacy

focus on two types of adequacy errors:

- lexical word sense disambiguation:
  translate ambiguous word with wrong word sense
- polarity:
  deletion or insertion of negation marker ("not", "no", "un-")

Polarity

manual error analysis [Fancellu and Webber, 2015]

translation errors (Chinese→English hierarchical PBSMT):

- insertion of negation (1–2%)
- deletion of negation (10–20%)
- reordering errors (1–20%)

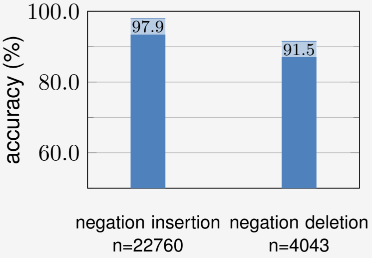
Polarity

manual error analysis [Fancellu and Webber, 2015]

translation errors (Chinese→English hierarchical PBSMT):

- insertion of negation (1–2%)
- deletion of negation (10–20%)
- reordering errors (1–20%)

automatic analysis (Lingeval97; NMT)



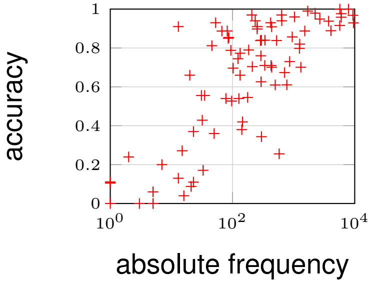
Word Sense Disambiguation [Rios, Mascarell, Sennrich, WMT 2017]

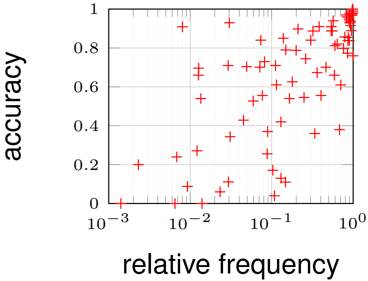
test set (ContraWSD)

- 35 ambiguous German nouns
- 2–4 senses per source noun
- contrastive translation sets (1 or more contrastive translations)
- $\approx$ 100 test instances per sense
  $\rightarrow \approx$ 7000 test instances

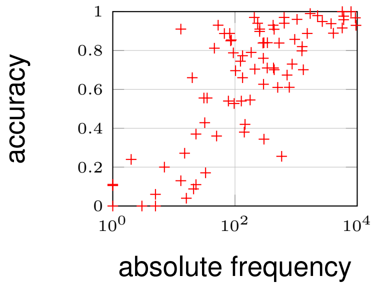
source:	*Also nahm ich meinen amerikanischen Reisepass und stellte mich in die **Schlange** für Extranjeros.*
reference:	*So I took my U.S. passport and got in the **line** for Extranjeros.*
contrastive:	*So I took my U.S. passport and got in the **snake** for Extranjeros.*
contrastive:	*So I took my U.S. passport and got in the **serpent** for Extranjeros.*

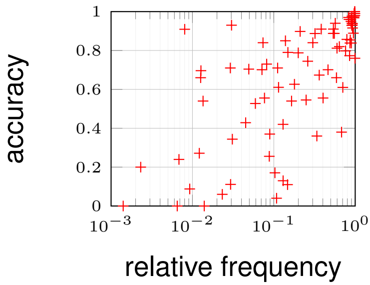
Word Sense Disambiguation





Word Sense Disambiguation





WSD is challenging, especially for rare word senses

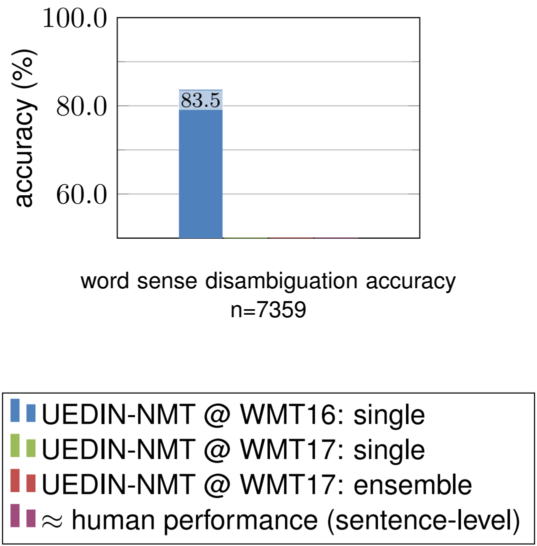
Word Sense Disambiguation: Measuring Progress

UEDIN-NMT at WMT (German→English)

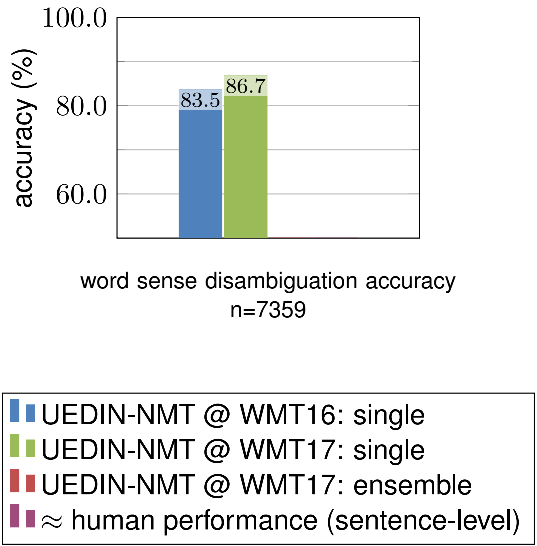
[Sennrich, Birch, Currey, Germann, Haddow, Heafield, Miceli Barone, Williams, WMT 2017]

- at WMT16, UEDIN-NMT was top-ranked
- large lead in fluency; small lead in adequacy
- for WMT17, we improved our MT system in several ways:
    - deep transition networks
    - layer normalization
    - better hyperparameters
    - better ensembles
    - (slightly) more training data
- are we getting better at word sense disambiguation?

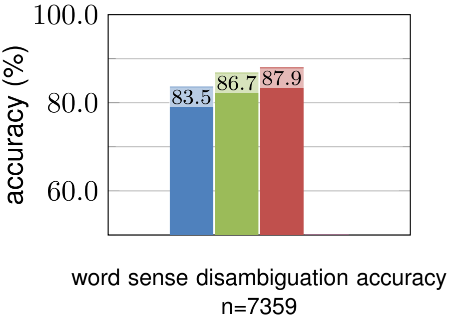
Results: Word Sense Disambiguation



Results: Word Sense Disambiguation

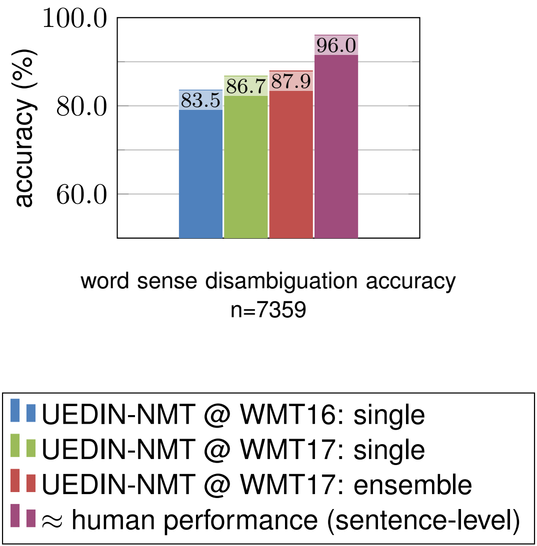


Results: Word Sense Disambiguation



UEDIN-NMT @ WMT16: single
UEDIN-NMT @ WMT17: single
UEDIN-NMT @ WMT17: ensemble
≈ human performance (sentence-level)

Results: Word Sense Disambiguation



What Did We Learn?

- word sense disambiguation remains challenging problem in MT, but measurable progress in last year
- On sentence-level, even humans may find it challenging

German	*Sehen Sie die **Muster**?*
reference	*Do you see the **patterns**?*
contrastive	*Do you see the **examples**?*

$\rightarrow$ new possibility for targeted evaluation of document-level modelling

NMT: what's linguistics got to do with it?

1. Linguistically Motivated (but Non-Linguistic) Models
2. Targeted Evaluation of Neural MT
3. Linguistically Informed Models

Linguistic Structure is Coming Back to (Neural) MT

segmentation	word
None	perusasian
BPE	perusasi: an
Omorfi	perus: asia: n

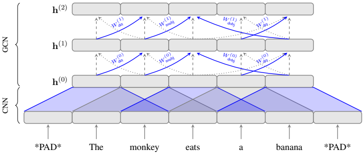
Morphology

[Sánchez-Cartagena and Toral, 2016]

[Tamchyna et al., 2017]

[Huck et al., 2017]

[Pinnis et al., 2017]



Syntax

[Sennrich and Haddow, 2016]

[Eriguchi et al., 2016]

[Bastings et al., 2017]

[Aharoni and Goldberg, 2017]

[Nadejde et al., 2017]

Linguistic Input Features: Motivation [Sennrich, Haddow, WMT 2016]

disambiguate words by POS

English	German
$\text{close}_{\text{verb}}$	schließen
$\text{close}_{\text{adj}}$	nah
$\text{close}_{\text{noun}}$	Ende

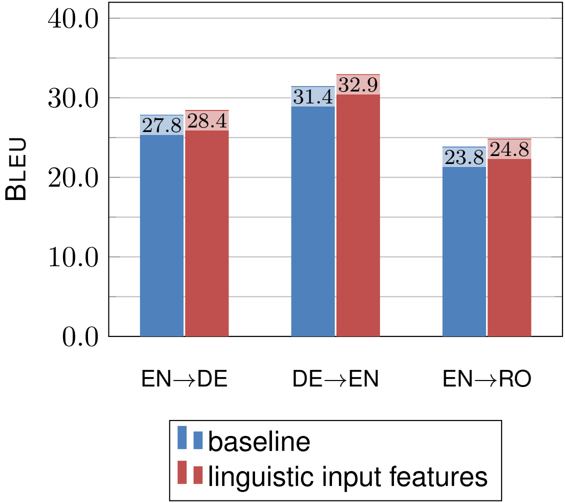
source	*We thought a win like this might be* $\textit{close}_{\textit{adj}}$.
reference	*Wir dachten, dass ein solcher Sieg nah sein könnte.*
baseline NMT	**Wir dachten, ein Sieg wie dieser könnte schließen.*

Neural Machine Translation: Multiple Input Features

use separate embeddings for each feature, then concatenate

$$E_1(close) = \begin{bmatrix} 0.4 \\ 0.1 \\ 0.2 \end{bmatrix} \qquad E_2(adj) = \begin{bmatrix} 0.1 \end{bmatrix} \qquad E_1(close) \parallel E_2(adj) = \begin{bmatrix} 0.4 \\ 0.1 \\ 0.2 \\ 0.1 \end{bmatrix}$$

Results



BLEU
40.0
30.0
20.0
10.0
0.0
27.8
28.4
31.4
32.9
23.8
24.8
EN→DE
DE→EN
EN→RO
baseline
linguistic input features

Predicting Target-Side Syntax (CCG)

[Nadejde, Reddy, Sennrich, Dwojak, Junczys-Dowmunt, Koehn, Birch, WMT 2017]

Core Idea

- CCG supertags carry information about type/direction of arguments
- predict supertags to help model produce good grammatical structure
- we associate words with their supertag by *interleaving*

words:	**Obama**	**receives**	**Netanyahu**	**in**	**the**	**capital**	**of**	**USA**
CCG:	NP	S\NP/PP/NP	NP	PP/NP	NP/N	N	NP\NP/NP	NP

interleaved: NP **Obama** S\NP/PP/NP **receives** NP **Netanyahu** PP/NP **in** NP/N **the** N **capital** NP\NP/NP **of** NP **USA**

similar idea: serialized dependency tree [Aharoni and Goldberg, 2017]

Jane hatte eine Katze .

$\rightarrow$

$(_{ROOT}$ $(_{S}$ $(_{NP}$ **Jane** $)_{NP}$ $(_{VP}$ **had** $(_{NP}$ **a cat** $)_{NP}$ $)_{VP}$ **.** $)_{S}$ $)_{ROOT}$

Results

[Nadejde et al., 2017]

system	DE→EN	RO→EN
baseline	32.1	28.4
interleaved CCG	32.7	29.3

[Aharoni and Goldberg, 2017]

system	DE→EN
baseline	32.4
serialized dependencies	33.2

Results

[Nadejde et al., 2017]

system	DE→EN	RO→EN
baseline	32.1	28.4
interleaved CCG	32.7	29.3

[Aharoni and Goldberg, 2017]

system	DE→EN
baseline	32.4
serialized dependencies	33.2

Results

[Nadejde et al., 2017]

system	DE→EN	RO→EN
baseline	32.1	28.4
interleaved CCG	32.7	29.3

[Aharoni and Goldberg, 2017]

system	DE→EN
baseline	32.4
serialized dependencies	33.2

...but more analysis in the papers

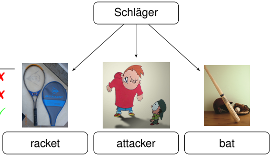
Conclusions

- neural machine translation does not *need* linguistic knowledge...
- ...but linguistics *should* play an important role for

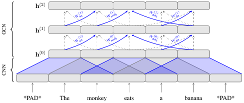
inspiring research

source	**indoor temperature**	
reference	**Raumklima**	
[Bahdanau et al., 2015]	**UNK**	✗
[Jean et al., 2015]	**Innenpool**	✗
[**Sennrich**, Haddow, Birch, ACL 2016a]	**Innen+ temperatur**	✓

targeted evaluation



informing models



Collaborators

Alexandra Birch

Barry Haddow

Antonio Valerio Miceli Barone

Kenneth Heafield

Maria Nadejde

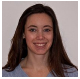
Phil Williams

Ulrich Germann

Tomasz Dwojak

Philipp Koehn

Siva Reddy



Anna Currey

Marcin Junczys-Dowmunt

Annette Rios

Laura Mascarell

Martin Volk

Open Positions

PhD positions

I have two PhD positions available at the University of Edinburgh.

postdoc

open position for post-doctoral researcher.

Contact me if you're interested.

Thanks

Acknowledgments

Some of the research presented was conducted in cooperation with Samsung Electronics Polska.

This work has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreements 645452 (QT21), TraMOOC (644333), HimL (644402), and SUMMA (688139).

This work has received funding from the Swiss National Science Foundation (SNF) in the project CoNTra (grant number 105212_169888).

TSD 2017 Sponsors

Springer

Thanks

Thank you for your attention

Resources

- LingEval97: https://github.com/rsennrich/lingeval97
- ContraWSD: https://github.com/a-rios/ContraWSD
- pre-trained models:
  - WMT16: http://data.statmt.org/wmt16_systems/
  - WMT17: http://data.statmt.org/wmt17_systems/

Bibliography I

Aharoni, R. and Goldberg, Y. (2017).
Towards String-To-Tree Neural Machine Translation.
In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 132–140, Vancouver, Canada. Association for Computational Linguistics.

Bahdanau, D., Cho, K., and Bengio, Y. (2015).
Neural Machine Translation by Jointly Learning to Align and Translate.
In Proceedings of the International Conference on Learning Representations (ICLR).

Bastings, J., Titov, I., Aziz, W., Marcheggiani, D., and Sima'an, K. (2017).
Graph Convolutional Encoders for Syntax-aware Neural Machine Translation.
Proceedings of EMNLP.

Bojar, O., Chatterjee, R., Federmann, C., Graham, Y., Haddow, B., Huck, M., Jimeno Yepes, A., Koehn, P., Logacheva, V., Monz, C., Negri, M., Neveol, A., Neves, M., Popel, M., Post, M., Rubino, R., Scarton, C., Specia, L., Turchi, M., Verspoor, K., and Zampieri, M. (2016).
Findings of the 2016 Conference on Machine Translation (WMT16).
In Proceedings of the First Conference on Machine Translation, Volume 2: Shared Task Papers, pages 131–198, Berlin, Germany.

Cai, Z. and Dai, Z. (2017).
Glyph-aware Embedding of Chinese Characters.
In 1st Workshop on Subword and Character level models in NLP (SCLeM), Copenhagen, Denmark.

Castilho, S., Moorkens, J., Gaspari, F., Sennrich, R., Sosoni, V., Georgakopoulou, Y., Lohar, P., Way, A., Barone, A. V. M., and Gialama, M. (2017).
A Comparative Quality Evaluation of PBSMT and NMT using Professional Translators.
In Proceedings of Machine Translation Summit XVI, Nagoya, Japan.

Bibliography II

Cohn, T., Hoang, C. D. V., Vymolova, E., Yao, K., Dyer, C., and Haffari, G. (2016).
Incorporating Structural Alignment Biases into an Attentional Neural Translation Model.
In
Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Langua
pages 876–885, San Diego, California.

Costa-jussà, M. R., Aldón, D., and Fonollosa, J. A. R. (2017).
Chinese–Spanish neural machine translation enhanced with character and word bitmap fonts.
Machine Translation, 31(1):35–47.

Eriguchi, A., Hashimoto, K., and Tsuruoka, Y. (2016).
Tree-to-Sequence Attentional Neural Machine Translation.
In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 823–833, Berlin, Germany.

Fancellu, F. and Webber, B. (2015).
Translating Negation: A Manual Error Analysis.
In
Proceedings of the Second Workshop on Extra-Propositional Aspects of Meaning in Computational Semantics (ExProM 2015),
pages 2–11, Denver, Colorado. Association for Computational Linguistics.

Gage, P. (1994).
A New Algorithm for Data Compression.
C Users J., 12(2):23–38.

Gehring, J., Auli, M., Grangier, D., Yarats, D., and Dauphin, Y. N. (2017).
Convolutional Sequence to Sequence Learning.
CoRR, abs/1705.03122.

Bibliography III

Huck, M., Riess, S., and Fraser, A. (2017).
Target-side Word Segmentation Strategies for Neural Machine Translation.
In Proceedings of the Second Conference on Machine Translation, Volume 1: Research Papers, Copenhagen, Denmark. Association for Computational Linguistics.

Jean, S., Cho, K., Memisevic, R., and Bengio, Y. (2015).
On Using Very Large Target Vocabulary for Neural Machine Translation.
In
Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference o
pages 1–10, Beijing, China. Association for Computational Linguistics.

Lee, J., Cho, K., and Hofmann, T. (2016).
Fully Character-Level Neural Machine Translation without Explicit Segmentation.
ArXiv e-prints.

Nadejde, M., Reddy, S., Sennrich, R., Dwojak, T., Junczys-Dowmunt, M., Koehn, P., and Birch, A. (2017).
Syntax-aware Neural Machine Translation Using CCG.
ArXiv e-prints.

Pinnis, M., Krislauks, R., Deksne, D., and Miks, T. (2017).
Neural Machine Translation for Morphologically Rich Languages with Improved Sub-word Units and Synthetic Data.
In Text, Speech, and Dialogue - 20th International Conference, TSD 2017, pages 237–245, Prague, Czech Republic.

Rios, A., Mascarell, L., and Sennrich, R. (2017).
Improving Word Sense Disambiguation in Neural Machine Translation with Sense Embeddings.
In Proceedings of the Second Conference on Machine Translation, Volume 1: Research Papers, Copenhagen, Denmark.

Bibliography IV

Sánchez-Cartagena, V. M. and Toral, A. (2016).
Abu-MaTran at WMT 2016 Translation Task: Deep Learning, Morphological Segmentation and Tuning on Character Sequences.
In Proceedings of the First Conference on Machine Translation, pages 362–370, Berlin, Germany. Association for Computational Linguistics.

Sennrich, R. (2015).
Modelling and Optimizing on Syntactic N-Grams for Statistical Machine Translation.
Transactions of the Association for Computational Linguistics, 3:169–182.

Sennrich, R. (2017).
How Grammatical is Character-level Neural Machine Translation? Assessing MT Quality with Contrastive Translation Pairs.
In Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL), Valencia, Spain.

Sennrich, R., Birch, A., Currey, A., Germann, U., Haddow, B., Heafield, K., Miceli Barone, A. V., and Williams, P. (2017).
The University of Edinburgh's Neural MT Systems for WMT17.
In Proceedings of the Second Conference on Machine Translation, Volume 2: Shared Task Papers, Copenhagen, Denmark.

Sennrich, R. and Haddow, B. (2015).
A Joint Dependency Model of Morphological and Syntactic Structure for Statistical Machine Translation.
In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pages 2081–2087, Lisbon, Portugal.

Sennrich, R. and Haddow, B. (2016).
Linguistic Input Features Improve Neural Machine Translation.
In Proceedings of the First Conference on Machine Translation, Volume 1: Research Papers, pages 83–91, Berlin, Germany.

Bibliography V

Sennrich, R., Haddow, B., and Birch, A. (2016a).
Edinburgh Neural Machine Translation Systems for WMT 16.
In Proceedings of the First Conference on Machine Translation, Volume 2: Shared Task Papers, pages 368–373, Berlin, Germany.

Sennrich, R., Haddow, B., and Birch, A. (2016b).
Neural Machine Translation of Rare Words with Subword Units.
In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1715–1725, Berlin, Germany.

Tamchyna, A., Weller-Di Marco, M., and Fraser, A. (2017).
Modeling Target-Side Inflection in Neural Machine Translation.
In Second Conference on Machine Translation (WMT17).

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017).
Attention Is All You Need.
CoRR, abs/1706.03762.

Williams, P., Sennrich, R., Post, M., and Koehn, P. (2016).
Syntax-based Statistical Machine Translation, volume 9 of Synthesis Lectures on Human Language Technologies.
Morgan & Claypool Publishers.

Zoph, B. and Knight, K. (2016).
Multi-Source Neural Translation.
In NAACL HLT 2016.