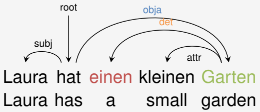
Neural Machine Translation what's linguistics got to do with it?

Rico Sennrich

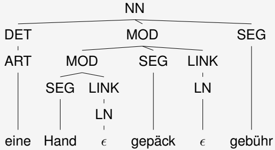
University of Edinburgh

Setting the Scene: 2014–2015

research trend: more linguistics for statistical machine translation



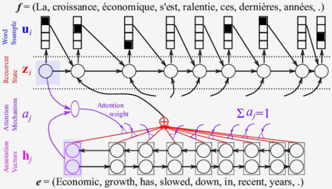
syntax-based LM
[Sennrich, TACL 2015]



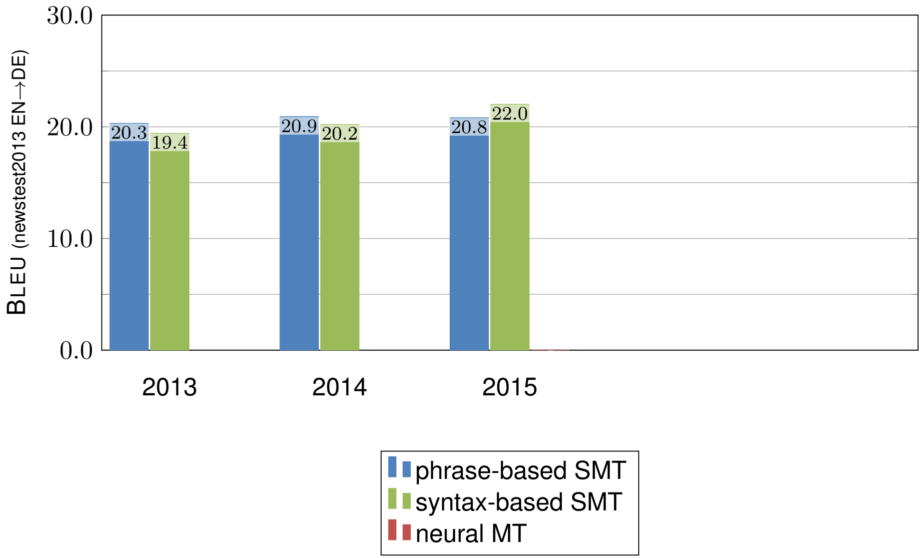
morphological structure
[Sennrich, Haddow, EMNLP 2015]

a new challenger appears: neural machine translation

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

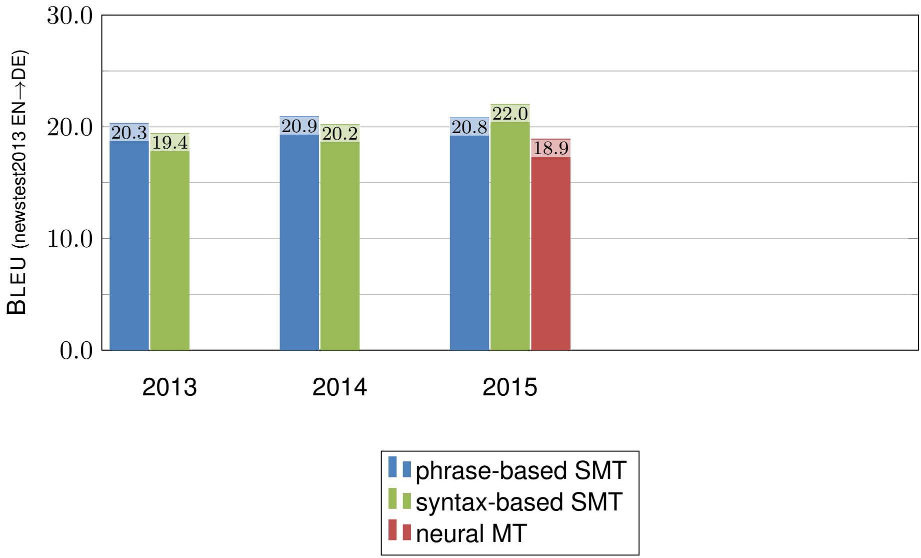


Edinburgh's* WMT Results over the Years



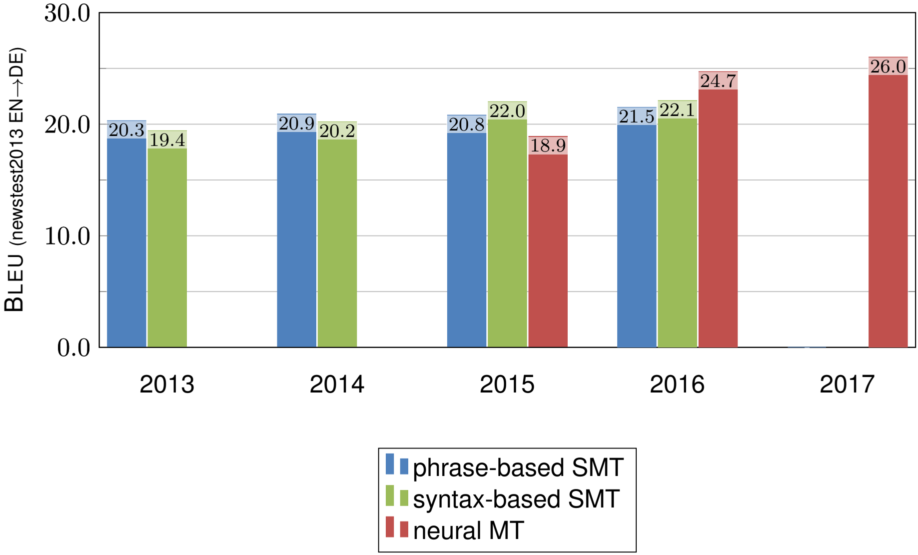
*NMT 2015 from U. Montréal: `https://sites.google.com/site/acl16nmt/`

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



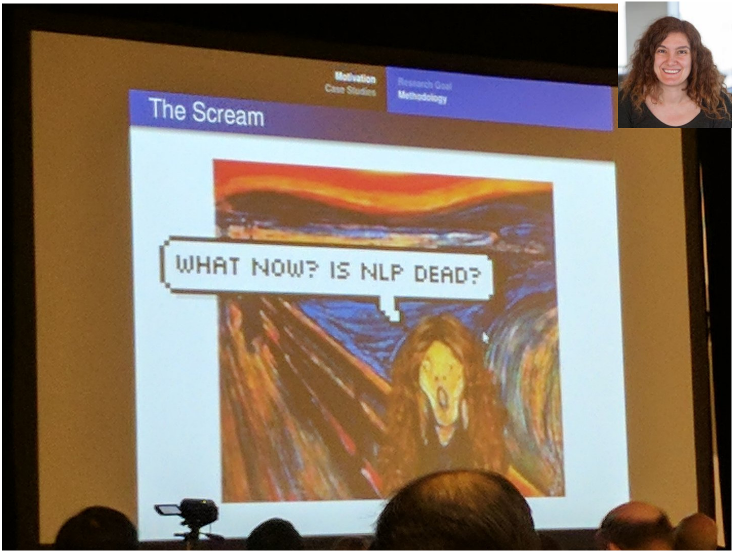
*NMT 2015 from U. Montréal: `https://sites.google.com/site/acl16nmt/`

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

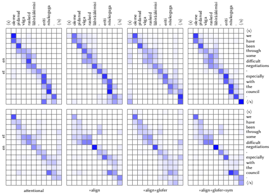
areas in which linguistics is helping neural MT research

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

Linguistically Motivated Models

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

氵	(water)
河	river
湖	lake
海	sea



subword segmentation

[Sennrich et al., 2016b]

logographic input

[Costa-jussà et al., 2017]

[Cai and Dai, 2017]

structural alignment biases

[Cohn et al., 2016]

Linguistic Structure is Coming Back to (Neural) MT

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

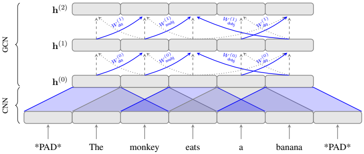
Morphology

[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]

Targeted Evaluation of Neural MT

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?

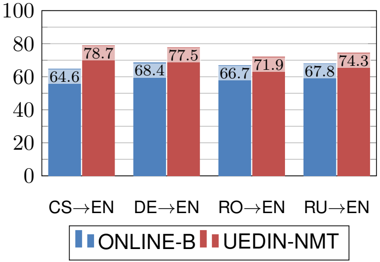
	many languages have long-distance interactions.
hypothesis:	model A produces disfluent output because it models these interactions poorly.
	model B can better model long-distance interactions, and produces more fluent output.
evidence:	model A obtains higher BLEU than model B on data set X

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]

Fluency
is translation good English?
+13%



Adequacy
is meaning preserved?
+1%

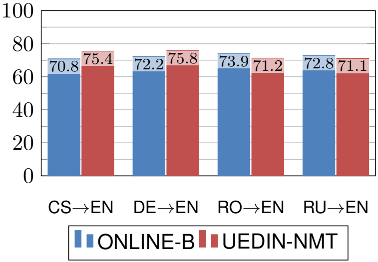


Figure: WMT16 direct assessment results

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

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

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

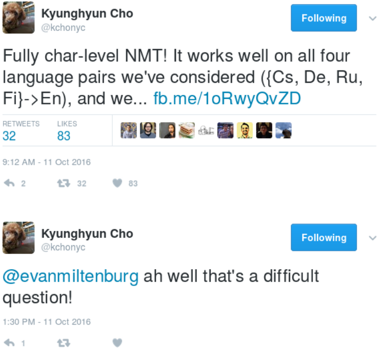
1

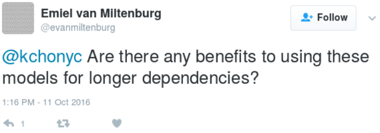
Case Study: Some Open Questions in Neural MT

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





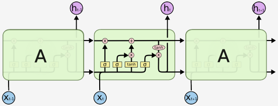
text representation

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



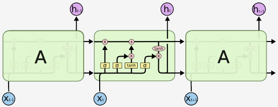
RNN vs. GRU vs. LSTM

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

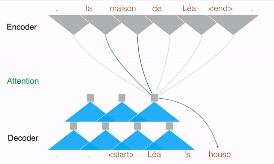
Case Study: Some Open Questions in Neural MT

does network architecture affect learning of long-distance dependencies?

architectures

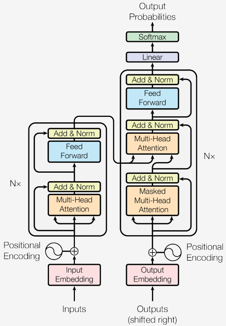


RNN vs. GRU vs. LSTM



(convolution)

[Gehring et al., 2017]

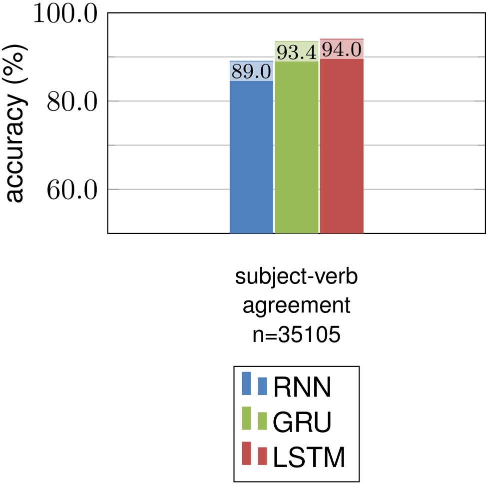


(self-attention)

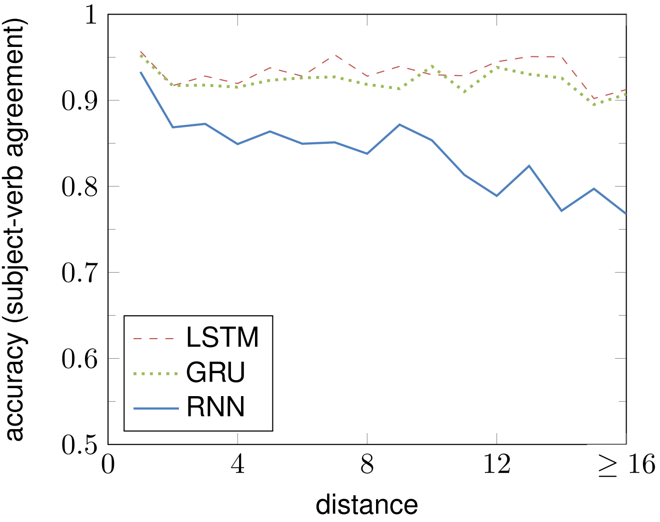
[Vaswani et al., 2017]

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

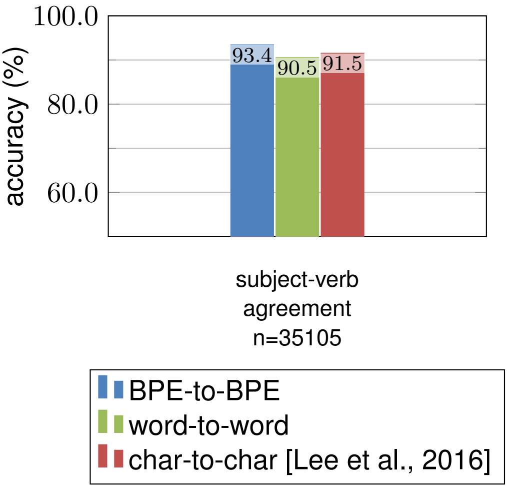
Results: Architecture



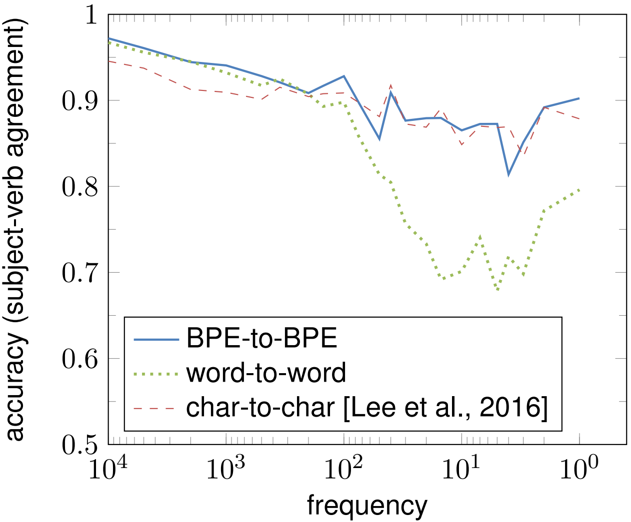
Results: Architecture



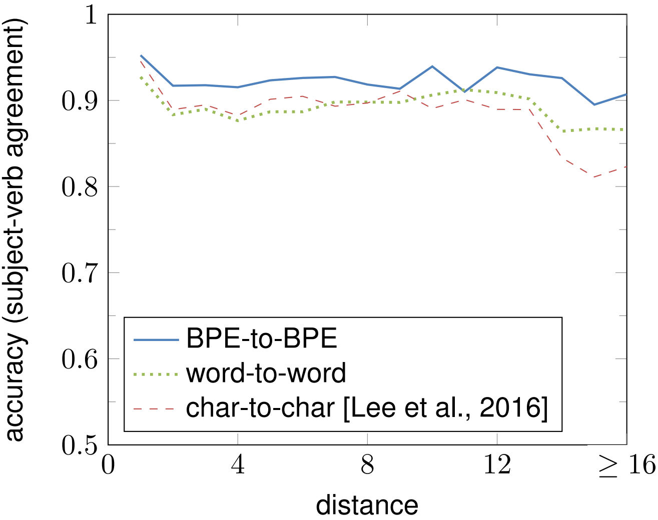
Results: Text Representation



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.

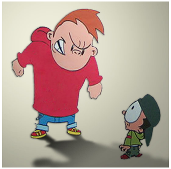
Schläger

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

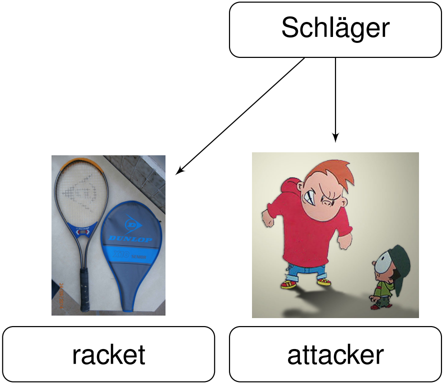


attacker

Targeted Analysis: Adequacy

adequacy is open problem

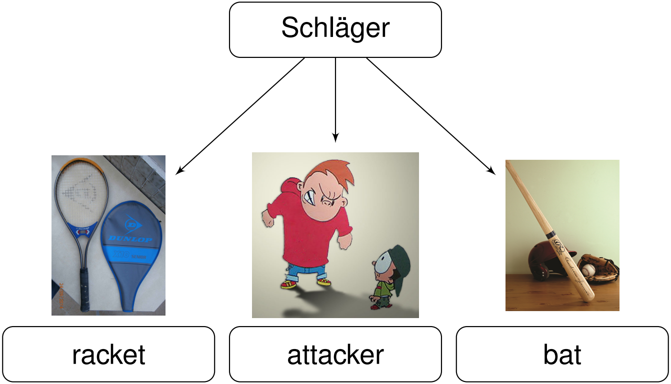
system	sentence
source	Dort wurde er von dem **Schläger** und einer weiteren männl. Person erneut angegriffen.
reference	There he was attacked again by his **original attacker** and another male.
our NMT	There he was attacked again by the **racket** and another male person.
Google	There he was again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



Targeted Analysis: Adequacy

adequacy is open problem

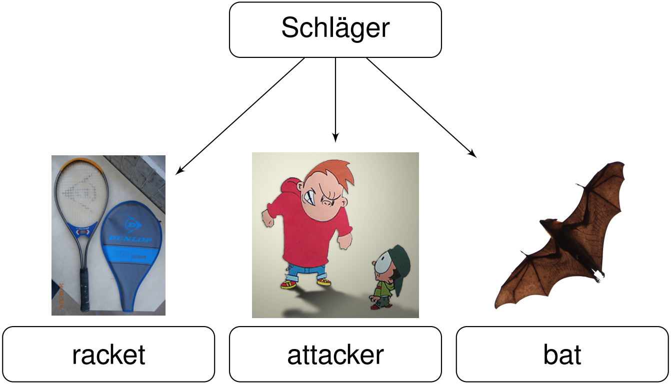
system	sentence
source	Dort wurde er von dem **Schläger** und einer weiteren männl. Person erneut angegriffen.
reference	There he was attacked again by his **original attacker** and another male.
our NMT	There he was attacked again by the **racket** and another male person.
Google	There he was again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



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%)

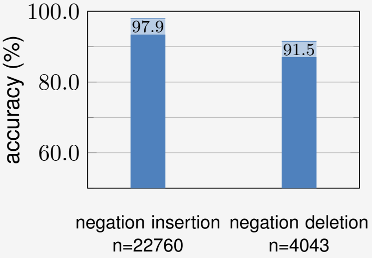
Polarity

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)



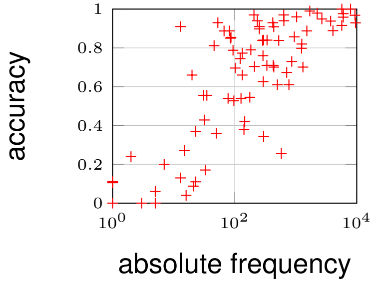
Word Sense Disambiguation [Rios, Mascarell, Sennrich, WMT 2017]

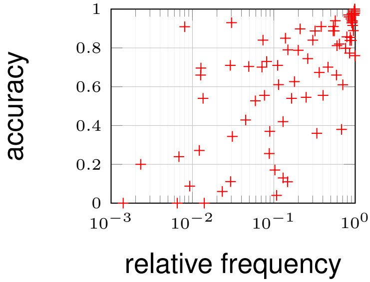
test set (ContraWSD)

- 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

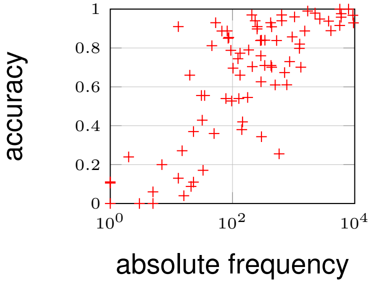
source:	*Also nahm ich meinen amerikanischen Reisepass und stellte mich in die **Schlange** für Extranjeros.*
reference:	*So I took my U.S. passport and got in the **line** for Extranjeros.*
contrastive:	*So I took my U.S. passport and got in the **snake** for Extranjeros.*
contrastive:	*So I took my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

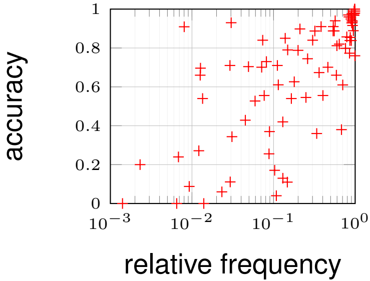
Word Sense Accuracy





Word Sense Accuracy





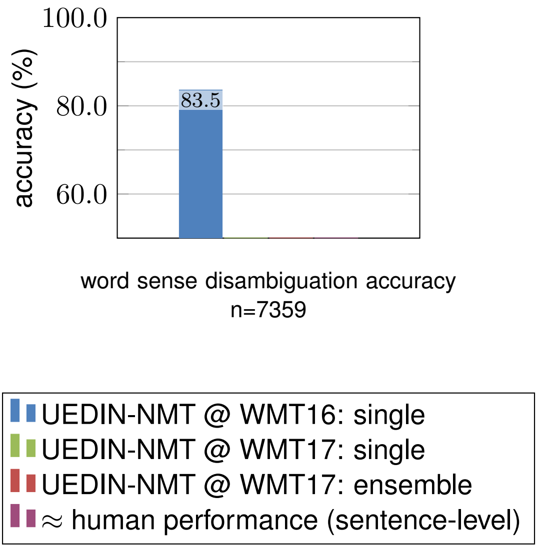
WSD is challenging, especially for rare word senses

Word Sense Disambiguation: Measuring Progress

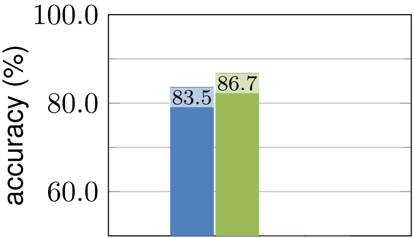
UEDIN-NMT at WMT (German→English)
[Sennrich, Birch, Currey, Germann, Haddow, Heafield, Miceli Barone, Williams, WMT 2017]

- 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?

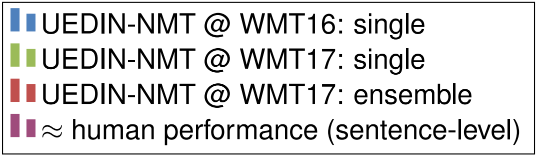
Results: Word Sense Disambiguation



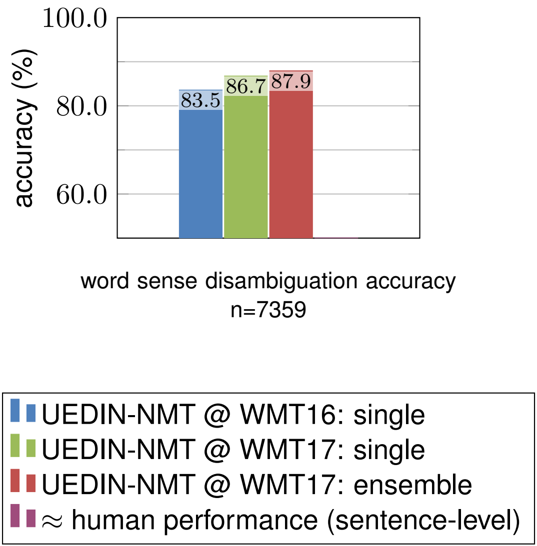
Results: Word Sense Disambiguation



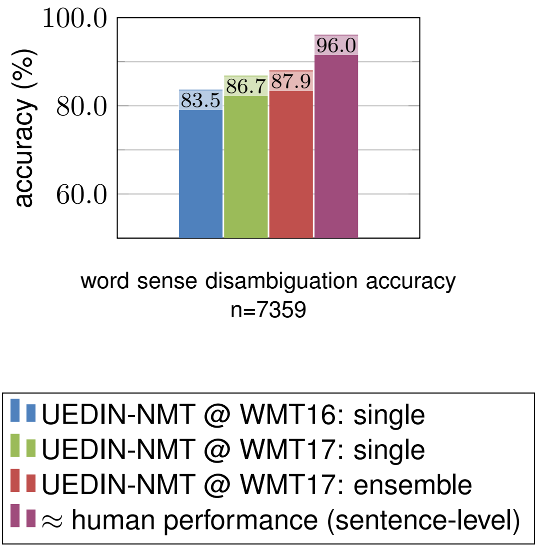
word sense disambiguation accuracy
n=7359



Results: Word Sense Disambiguation



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

Targeted Analysis: Coreference

background

antecedent agreement can often not be predicted based on source sentence, but requires extra-sentential context:

English	I made a decision.	Please respect it.
French	J’ai pris une décision.	Respectez-la s’il vous plaît.
French	J’ai fait un choix.	Respectez-le s’il vous plaît.

Targeted Analysis: Coreference

previous work: shared task on pronoun prediction

[Hardmeier et al., 2015, Guillou et al., 2016, Loáiciga et al., 2017]

- focus on correctness of pronouns, which are often coreferent.
- pronoun errors impact meaning, but only have small effect on BLEU.
- limitations of shared task:
  - many pronouns do not require extra-sentential context; sentence-level system still best at DiscoMT17 [Loáiciga et al., 2017].
  - we want to analyze NMT systems' ability to model coreference, without training specifically for this task, but:
    - task gives lemmatized target side
    - long tail of possible pronouns handled via OTHER category

Contrastive Pairs for Analysis of Coreference in MT

[Bawden, Sennrich, Birch, Haddow, in preparation]

Source:

context:	Oh, I hate **flies**. Look, there's another one!
current sent.:	Don't worry, I'll kill **it** for you.

Target:

1	context:	Ô je déteste les **mouches**. Regarde, il y en a une autre !
	correct:	T'inquiète, je **la** tuerai pour toi.
	incorrect:	T'inquiète, je **le** tuerai pour toi.
2	context:	Ô je déteste les **moucherons**. Regarde, il y en a un autre !
	correct:	T'inquiète, je **le** tuerai pour toi.
	incorrect:	T'inquiète, je **la** tuerai pour toi.

design of test set

- hand-crafted set of 200 contrastive pairs
- previous sentence required for correct prediction
- balanced so that sentence-level system scores 50%

Coreference Models

baseline setup

- training on OpenSubtitles EN-FR [Tiedemann, 2012]
- attentional encoder-decoder (Nematus) with BPE

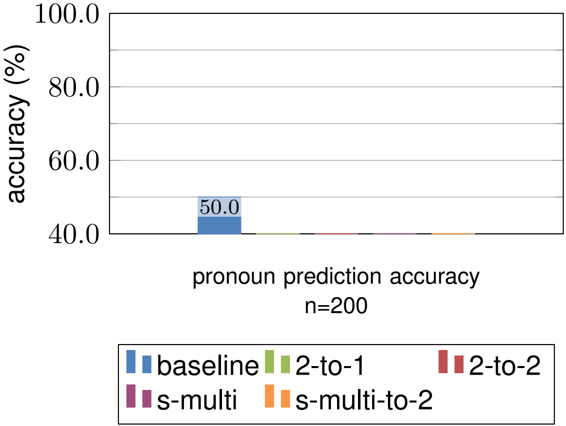
architectures

- sentence-level baseline
- 2-TO-1: concatentation of previous source sentence
- 2-TO-2: concatentation of previous source and target sentence
- S-MULTI: separate encoder for previous source; hierarchical attention
- S-MULTI-TO-2: separate encoder for previous source; previous target sentence concatenated

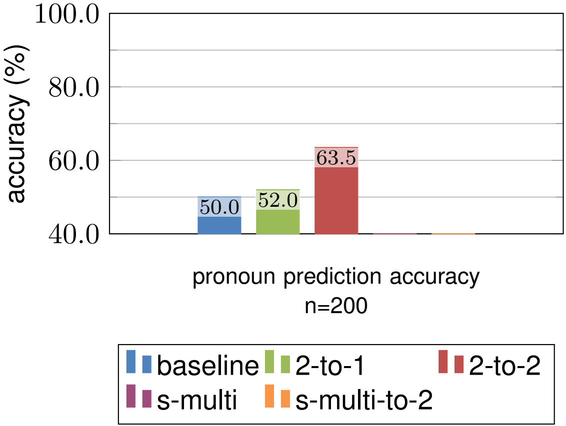
related work

- [Tiedemann and Scherrer, 2017] (2-TO-*)
- [Zoph and Knight, 2016, Libovický and Helcl, 2017] (S-MULTI)

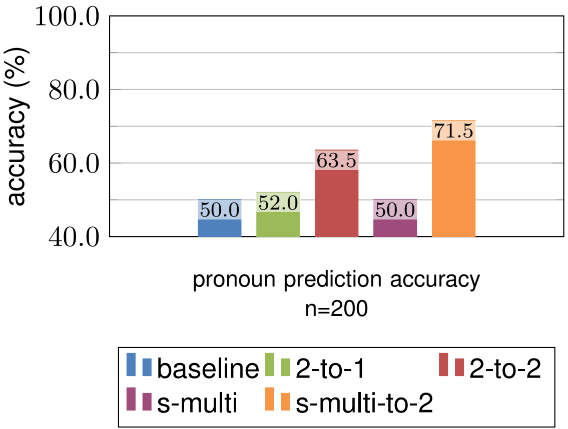
Targeted Analysis: Coreference: Results



Targeted Analysis: Coreference: Results



Targeted Analysis: Coreference: Results



Coreference Models: BLEU Results

System	BLEU ↑			
	Comedy	Crime	Fantasy	Horror
Single-encoder, non-contexual model				
BASELINE	19.52	22.07	26.30	33.05
Single-encoder with concatenated input				
2-TO-2	20.09	**22.93**	26.60	33.59
2-TO-1	19.51	21.81	26.78	**34.37**
Multi-encoder, multi-attention models (+previous source sentence)				
S-MULTI	**20.22**	21.90	26.81	34.04
Multi-encoder, multi-attention models with concatenated output				
S-MULTI-TO-2	**20.85**	**22.81**	**27.17**	**34.62**

What Did We Learn?

- target context is crucial for prediction of correct pronoun (partially due to test set, in which source words are ambiguous)
- targeted evaluation can guide our exploration of architectures $\rightarrow$ multi-encoder architecture only works in some conditions (*-to-2)

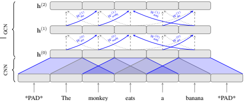
Conclusions

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

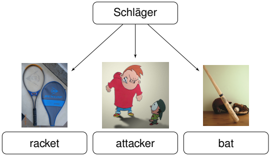
inspiring research

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

informing models



targeted evaluation



Collaborators

Joint work with:


Alexandra Birch


Barry Haddow


Rachel Bawden


Annette Rios


Laura Mascarell

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).

Thanks

Thank you for your attention

Resources

- LingEval97: https://github.com/rsennrich/lingeval97
- ContraWSD: https://github.com/a-rios/ContraWSD
- Discourse test set: https://diamt.limsi.fr/eval.html
- 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.

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

Guillou, L., Hardmeier, C., Nakov, P., Stymne, S., Tiedemann, J., Versley, Y., Cettolo, M., Webber, B., and Popescu-Belis, A. (2016).
Findings of the 2016 WMT Shared Task on Cross-lingual Pronoun Prediction.
In Proceedings of the First Conference on Machine Translation, pages 525–542, Berlin, Germany.

Bibliography III

Hardmeier, C., Nakov, P., Stymne, S., Tiedemann, J., Versley, Y., and Cettolo, M. (2015).
Pronoun-Focused MT and Cross-Lingual Pronoun Prediction: Findings of the 2015 DiscoMT Shared Task on Pronoun Translation.
In Proceedings of the Second Workshop on Discourse in Machine Translation, pages 1–16, Lisbon, Portugal. Association for Computational Linguistics.

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.

Libovický, J. and Helcl, J. (2017).
Attention Strategies for Multi-Source Sequence-to-Sequence Learning.
In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL'17, pages 196–202, Vancouver, Canada.

Bibliography IV

Loáiciga, S., Stymne, S., Nakov, P., Hardmeier, C., Tiedemann, J., Cettolo, M., and Versley, Y. (2017).
Findings of the 2017 DiscoMT Shared Task on Cross-lingual Pronoun Prediction.
In Proceedings of the Third Workshop on Discourse in Machine Translation, pages 1–16, Copenhagen, Denmark. Association for Computational Linguistics.

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.

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.

Bibliography V

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.

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.

Bibliography VI

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).

Tiedemann, J. (2012).
Parallel Data, Tools and Interfaces in OPUS.
In Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12), Istanbul, Turkey.

Tiedemann, J. and Scherrer, Y. (2017).
Neural Machine Translation with Extended Context.
In Proceedings of the Third Workshop on Discourse in Machine Translation, pages 82–92, Copenhagen, Denmark. Association for Computational Linguistics.

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.