Neural Machine Translation: Breaking through the Performance Ceiling

Rico Sennrich

University of Edinburgh; Universität Zürich

April 19 2018

Statistical Machine Translation (SMT)

given a sequence of words s in the source language,
find the most probable sequence t in the target language [Brown et al., 1993]

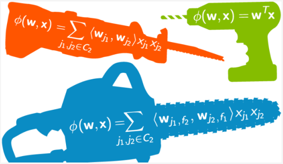
$$t^* \approx \arg\max_t \sum_{m=1}^{M} \lambda_m h_m(s, t)$$ [Och, 2003]

Statistical Machine Translation (SMT)

given a sequence of words s in the source language,
find the most probable sequence t in the target language [Brown et al., 1993]

$$t^* \approx \arg\max_t \sum_{m=1}^{M} \lambda_m h_m(s,t)$$ [Och, 2003]

main research trend



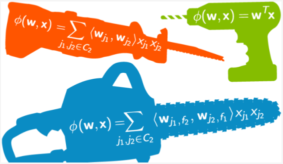
feature engineering

Statistical Machine Translation (SMT)

given a sequence of words s in the source language,
find the most probable sequence t in the target language [Brown et al., 1993]

$$t^* \approx \arg\max_t \sum_{m=1}^{M} \lambda_m h_m(s,t)$$ [Och, 2003]

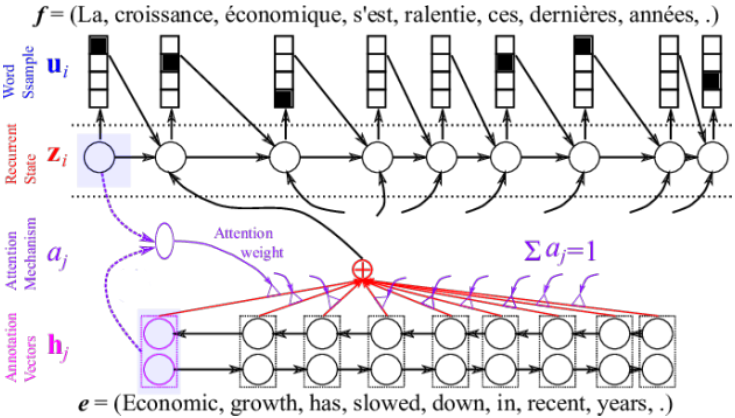
main research trend



feature engineering

$\rightarrow$

Neural Machine Translation

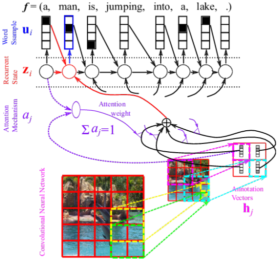


Kyunghyun Cho
http://devblogs.nvidia.com/parallelforall/introduction-neural-machine-translation-gpus-part-3/

Neural Machine Translation

very general model:

- applied to many sequence–to–sequence tasks
- variants used in computer vision



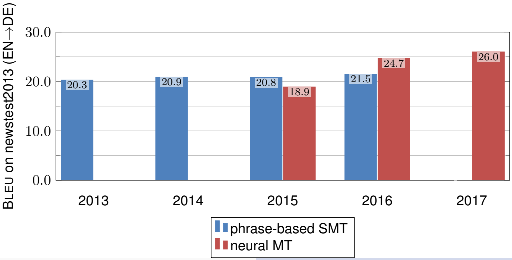
[Cho et al., 2015]

Neural Machine Translation: Timeline

1987 Early encoder-decoder, with vocabulary size 30-40 [Allen, 1987]

⋮

2013 Pure neural MT system presented [Kalchbrenner and Blunsom, 2013]
2014 RNN Encoder-Decoder with Attention [Bahdanau et al., 2015]
2015 WMT 15: Montreal NMT is competitive [Jean et al., 2015b]
2015 Subword-level NMT **[Sennrich et al., 2016c]**
2015 Monolingual Data in NMT **[Sennrich et al., 2016b]**
2016 WMT 16: Edinburgh NMT is dominant **[Sennrich et al., 2016a]**
2017 Various architectures competitive [Gehring et al., 2017, Vaswani et al., 2017]



WMT16 Shared Translation Task: Results

(tied) best constrained system for **7 out of 8** translation directions

system	BLEU	official rank
uedin-nmt	34.2	1
metamind	32.3	2
uedin-syntax	30.6	3
NYU-UMontreal	30.8	4
online-B	29.4	5-10
KIT/LIMSI	29.1	5-10
cambridge	30.6	5-10
online-A	29.9	5-10
promt-rule	23.4	5-10
KIT	29.0	6-10
jhu-syntax	26.6	11-12
jhu-pbmt	28.3	11-12
uedin-pbmt	28.4	13-14
online-F	19.3	13-15
online-G	23.8	14-15

WMT16 EN→DE

system	BLEU	official rank
uedin-nmt	38.6	1
online-B	35.0	2-5
online-A	32.8	2-5
uedin-syntax	34.4	2-5
KIT	33.9	2-6
uedin-pbmt	35.1	5-7
jhu-pbmt	34.5	6-7
online-G	30.1	8
jhu-syntax	31.0	9
online-F	20.2	10

WMT16 DE→EN

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., 2015a]	Die **Innenpool** ist sehr angenehm.	✗

Open-Vocabulary Neural MT

problem

word-level neural networks use one-hot encoding
→ 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., 2015a]	Die **Innenpool** ist sehr angenehm.	✗
[Sennrich, Haddow, Birch, ACL 2016a]	Die **Innen+ temperatur** ist sehr angenehm.	✓

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

goal

subword segmentation that:

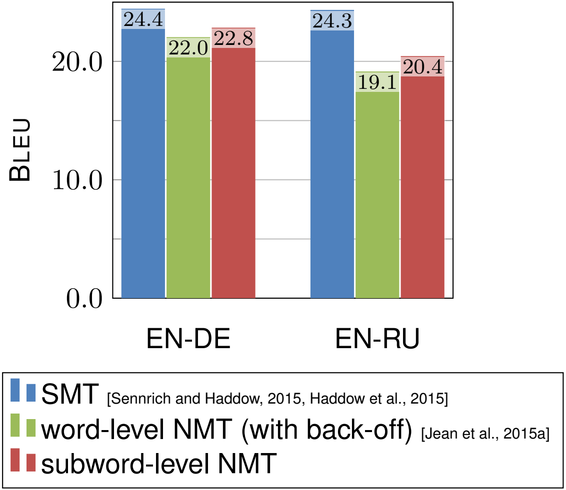
- uses a closed vocabulary of subword units
- can represent open vocabulary (including unknown words)
- minimizes the sequence length (given the vocabulary size)

solution

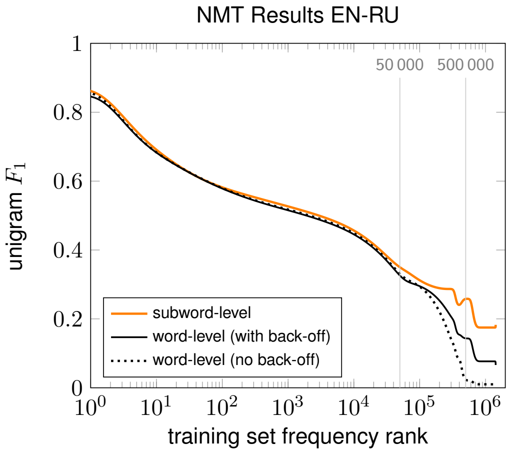
- greedy compression algorithm: byte pair encoding (BPE) [Gage, 1994]
- we adapt BPE to word segmentation
- hyperparameter: vocabulary size

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



Semi-Supervised Training for NMT [Sennrich, Haddow, Birch, ACL 2016b]

why?

monolingual data

- is much less sparse than parallel data
- is useful for structured prediction
- may be used for domain adaptation

why is this hard?

- standard in SMT: monolingual LM as feature in linear model
- linear combination of NMT and LM barely effective [Gülçehre et al., 2015]

our solution

end-to-end training of NMT model with parallel and monolingual data

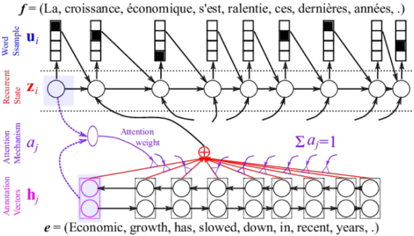
Monolingual Data in NMT

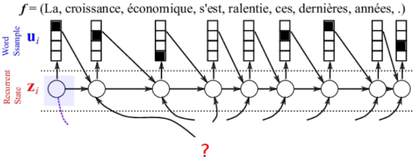
NMT is a conditional language model

$$p(u_i) = f(z_i, u_{i-1}, c_i)$$

Problem

for monolingual training instances, source context c_i is missing

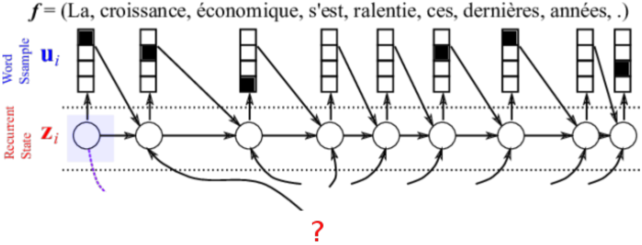




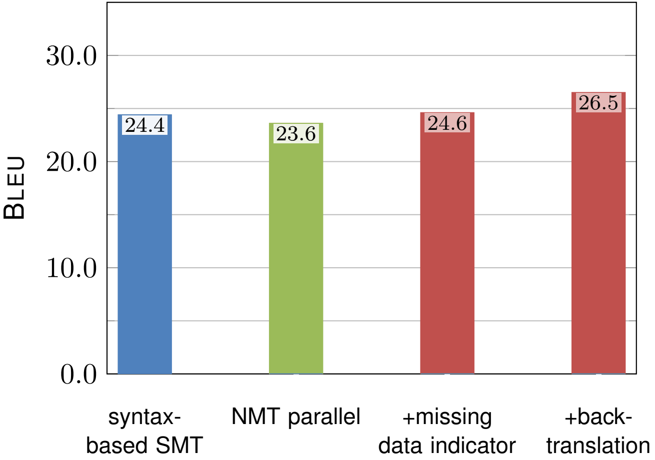
Monolingual Training Instances

solutions: missing data imputation for c_i

- missing data indicator: $\overrightarrow{0}$
  $\rightarrow$ works, but danger of catastrophic forgetting
- impute c_i with neural network
  $\rightarrow$ we do this indirectly by back-translating the target sentence



Evaluation: English→German

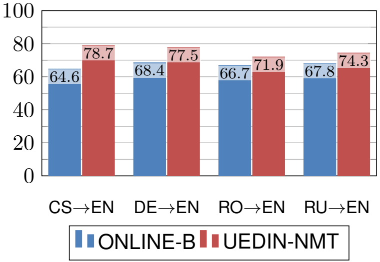


[Sennrich and Haddow, 2015]

(NMT systems are ensemble of 4)

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

Fluency
is translation good English?
+13%



Adequacy
is meaning preserved?
+1%

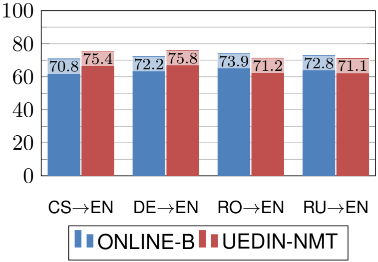


Figure: WMT16 direct assessment results

Adequacy in Neural Machine Translation

Word Sense Disambiguation

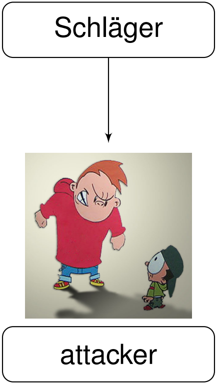
system	sentence
source	Dort wurde er von dem **Schläger** und einer weiteren männl. Person erneut angegriffen.
reference	There he was attacked again by his **original attacker** and another male.
our NMT	There he was attacked again by the **racket** and another male person.
Google	There he was again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

Schläger

Adequacy in Neural Machine Translation

Word Sense Disambiguation

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.

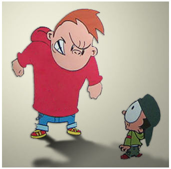


Adequacy in Neural Machine Translation

Word Sense Disambiguation

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



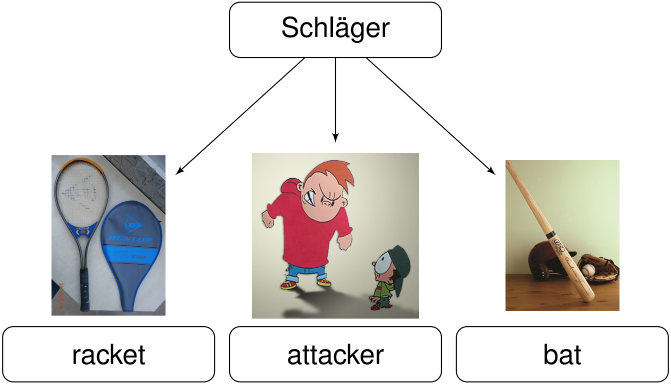
racket

attacker

Adequacy in Neural Machine Translation

Word Sense Disambiguation

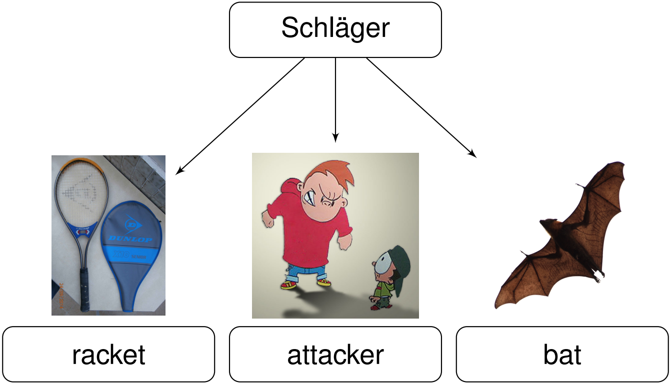
system	sentence
source	Dort wurde er von dem **Schläger** und einer weiteren männl. Person erneut angegriffen.
reference	There he was attacked again by his **original attacker** and another male.
our NMT	There he was attacked again by the **racket** and another male person.
Google	There he was again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



Adequacy in Neural Machine Translation

Word Sense Disambiguation

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.



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

core idea

- NMT assigns score to every translation hypothesis
- we provide NMT system with several translations:
  - correct human reference translation
  - contrastive variants which introduces error
- we count how often model prefers correct translation

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

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

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.*

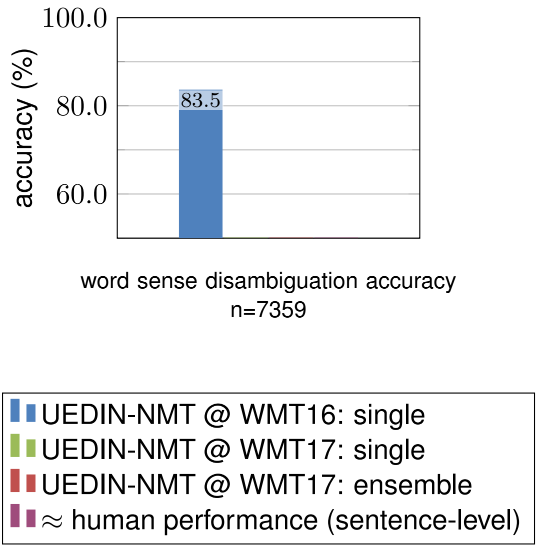
Word Sense Disambiguation: Progress over Time

UEDIN-NMT at WMT (German→English)

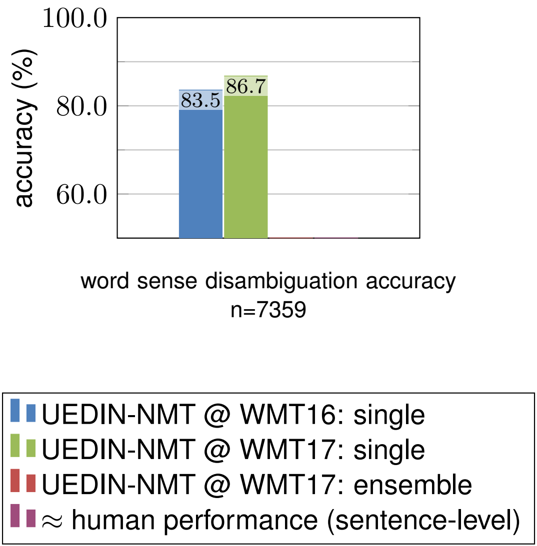
[Sennrich, Birch, Currey, Germann, Haddow, Heafield, Miceli Barone, Williams, WMT 2017]

- 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?

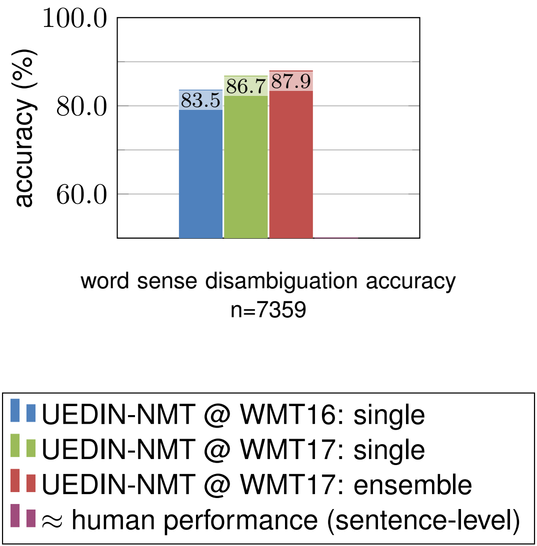
Results: Word Sense Disambiguation



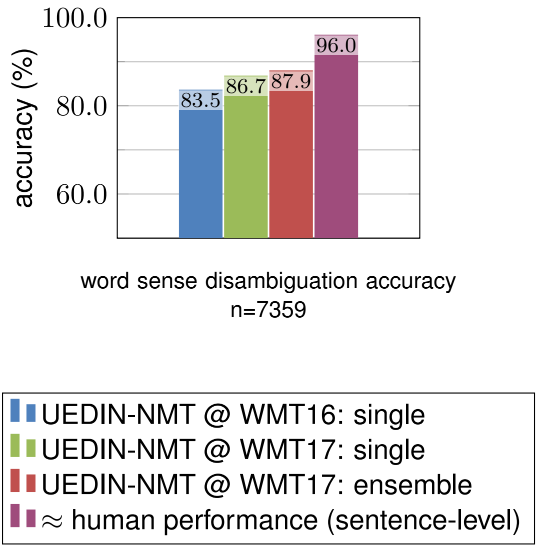
Results: Word Sense Disambiguation



Results: Word Sense Disambiguation



Results: Word Sense Disambiguation



What Did We Learn?

- word sense disambiguation remains challenging problem in MT, but measurable progress 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**?*

→ targeted evaluation of document-level modelling [Bawden et al., 2018]

Conclusion

neural sequence–to–sequence models

- neural models have revolutionized MT
- we have overcome early limitations
- many methods shared with general deep learning

open challenges

- increasing semantic faithfulness
  - scaling up sequence length (documents)
  - novel objective functions (NCE, GANs etc.)
  - word sense disambiguation
- data efficiency
  - interactive MT
  - one-shot learning
  - low-resourced translation

Collaborators

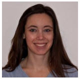
Alexandra Birch

Barry Haddow

Annette Rios

Laura Mascarell

Martin Volk



Anna Currey

Ulrich Germann

Kenneth Heafield

Antonio Valerio Miceli Barone

Phil Williams

Thanks

Thank you for your attention

Resources

- BPE scripts: `https://github.com/rsennrich/subword-nmt`
- 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

Allen, R. (1987).
Several Studies on Natural Language and Back-Propagation.
In IEEE First International Conference on Neural Networks, pages 335–341, San Diego, California, USA.

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).

Bawden, R., Sennrich, R., Birch, A., and Haddow, B. (2018).
Evaluating Discourse Phenomena in Neural Machine Translation.
In NAACL 2018, New Orleans, USA.

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.

Brown, P. F., Della Pietra, V. J., Della Pietra, S. A., and Mercer, R. L. (1993).
The Mathematics of Statistical Machine Translation: Parameter Estimation.
Computational Linguistics, 19(2):263–311.

Cho, K., Courville, A., and Bengio, Y. (2015).
Describing Multimedia Content using Attention-based Encoder-Decoder Networks.

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

Bibliography II

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

Gülçehre, c., Firat, O., Xu, K., Cho, K., Barrault, L., Lin, H., Bougares, F., Schwenk, H., and Bengio, Y. (2015).
On Using Monolingual Corpora in Neural Machine Translation.
CoRR, abs/1503.03535.

Haddow, B., Huck, M., Birch, A., Bogoychev, N., and Koehn, P. (2015).
The Edinburgh/JHU Phrase-based Machine Translation Systems for WMT 2015.
In Proceedings of the Tenth Workshop on Statistical Machine Translation, pages 126–133, Lisbon, Portugal. Association for Computational Linguistics.

Jean, S., Cho, K., Memisevic, R., and Bengio, Y. (2015a).
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
pages 1–10, Beijing, China. Association for Computational Linguistics.

Jean, S., Firat, O., Cho, K., Memisevic, R., and Bengio, Y. (2015b).
Montreal Neural Machine Translation Systems for WMT'15 .
In Proceedings of the Tenth Workshop on Statistical Machine Translation, pages 134–140, Lisbon, Portugal. Association for Computational Linguistics.

Kalchbrenner, N. and Blunsom, P. (2013).
Recurrent Continuous Translation Models.
In Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, Seattle. Association for Computational Linguistics.

Bibliography III

Och, F. J. (2003).
Minimum Error Rate Training in Statistical Machine Translation.
In ACL '03: Proceedings of the 41st Annual Meeting on Association for Computational Linguistics, pages 160–167, Sapporo, Japan. Association for Computational Linguistics.

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.

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., Firat, O., Cho, K., Birch, A., Haddow, B., Hitschler, J., Junczys-Dowmunt, M., Läubli, S., Miceli Barone, A. V., Mokry, J., and Nadejde, M. (2017).
Nematus: a Toolkit for Neural Machine Translation.
In
Proceedings of the Software Demonstrations of the 15th Conference of the European Chapter of the Association for Computational L
pages 65–68, Valencia, Spain.

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.

Bibliography IV

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).
Improving Neural Machine Translation Models with Monolingual Data.
In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 86–96, Berlin, Germany.

Sennrich, R., Haddow, B., and Birch, A. (2016c).
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.

Sennrich, R., Williams, P., and Huck, M. (2015).
A tree does not make a well-formed sentence: Improving syntactic string-to-tree statistical machine translation with more linguistic knowledge.
Computer Speech & Language, 32(1):27–45.
Hybrid Machine Translation: integration of linguistics and statistics.

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.