Revisiting Challenges in Neural Machine Translation

Rico Sennrich

University of Edinburgh

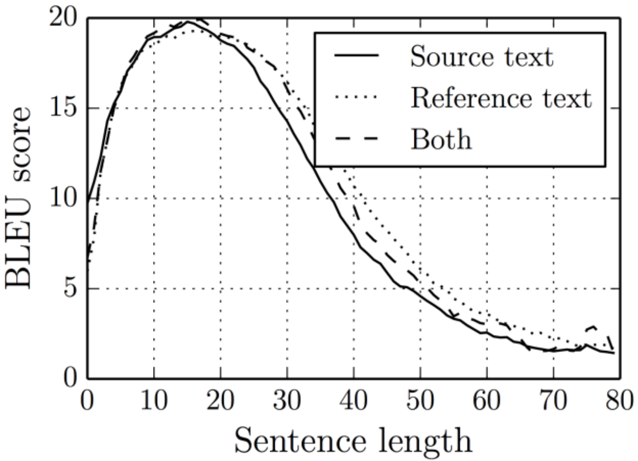
Why Revisit Challenges Regularly?

guide research directions

NMT facts
have expiration date

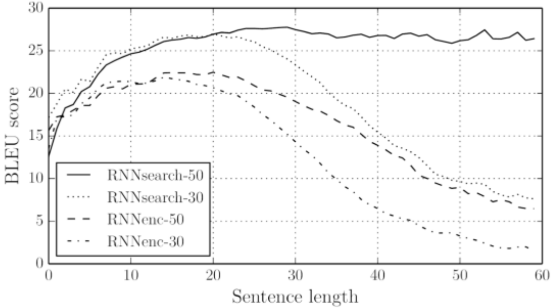
signpost: Ian Harding (CC BY-NC-SA 2.0)

Encoder–Decoder Has Information Bottleneck



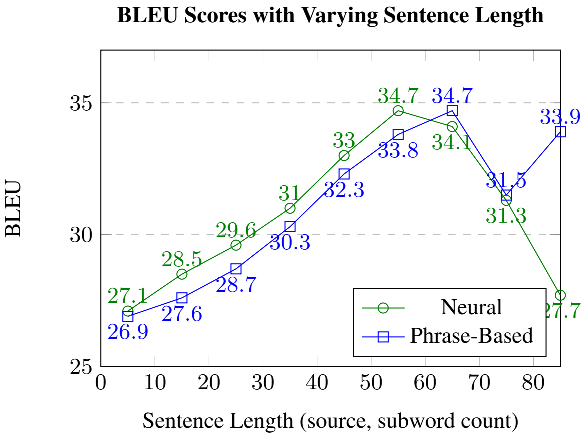
[Cho et al., 2014]

Attention Brings Improvement



[Bahdanau et al., 2015]

Still, Poor Performance Reported for Long Sentences



[Koehn and Knowles, 2017]

Adequacy vs. Fluency in WMT16 Evaluation

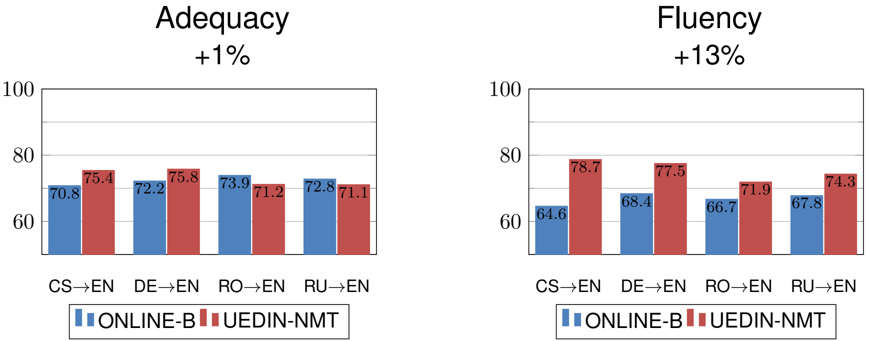


Figure: WMT16 direct assessment results

Human Evaluation in TraMOOC [Castilho et al., 2018]

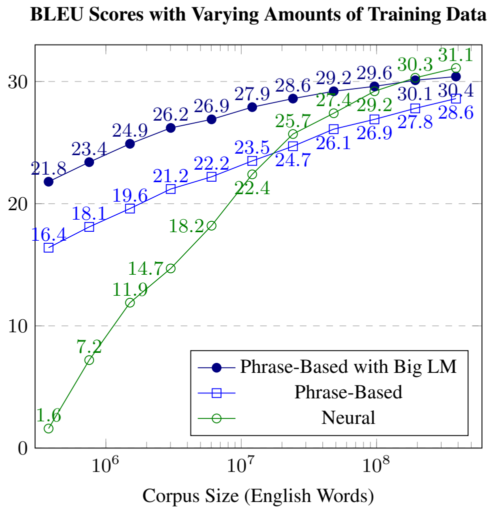
- comparison of NMT and PBSMT for EN$\rightarrow${DE,EL,PT,RU}
- direct assessment:
  - NMT obtains higher fluency judgment than PBSMT: +10%
  - NMT only obtains small improvement in adequacy judgment: +1%

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%**

Revisiting Challenges in NMT

Low-Resource Translation



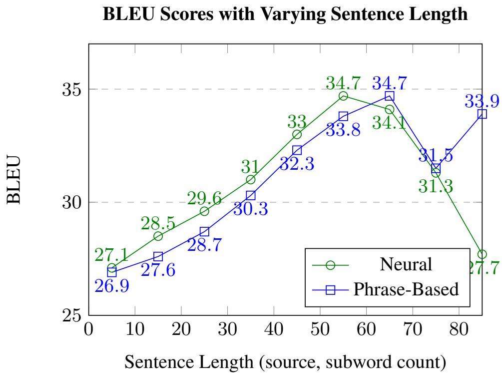
[Koehn and Knowles, 2017]

Why Are Long Sentences Hard?

different answers

- training on long sentences efficiently is challenging
  $\rightarrow$ training–test mismatch
- locally normalized models have bias towards low-entropy states
  $\rightarrow$ outputs too short (</s>)
- long-distance interactions may be challenging due to network path length (vanishing gradient)
- ...?

Long Sentences: Training–Test Mismatch



[Koehn and Knowles, 2017]

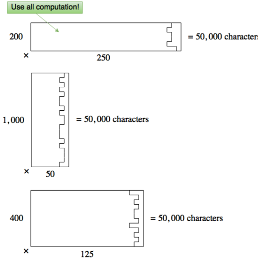
this is uedin-2016 system, trained with a maximum length of 50 subwords!

How to Train on Long Sentences

Problem: time and memory increases with longest sentence in batch

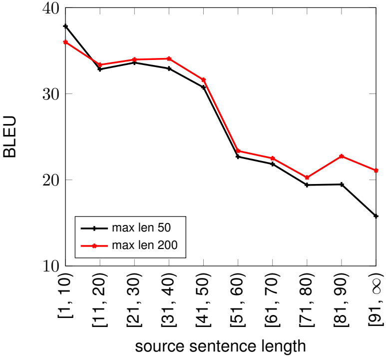
Solutions

- sort sentences of same length together [Sutskever et al., 2014]
- adjust batch size depending on length [Johansen et al., 2016]



[Johansen et al., 2016]

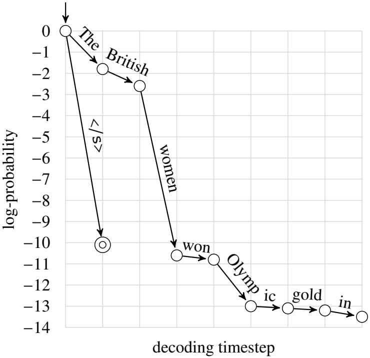
Training on Long Sentences Matters



BLEU score by source sentence length (number of subwords)

Long Sentences: Label Bias

locally normalized models have label bias [Murray and Chiang, 2018]



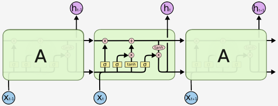
→ (tunable) length penalty and variants result in simple globally normalized model

→ other methods to escape local normalization include reconstruction [Tu et al., 2017]

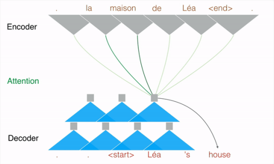
Long-Distance Interactions

does network architecture affect learning of long-distance dependencies?

architectures

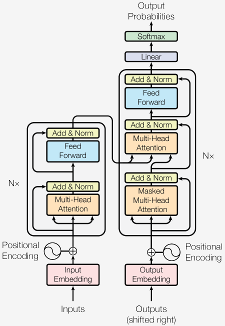


RNN/GRU/LSTM



convolution

[Gehring et al., 2017]



self-attention

[Vaswani et al., 2017]

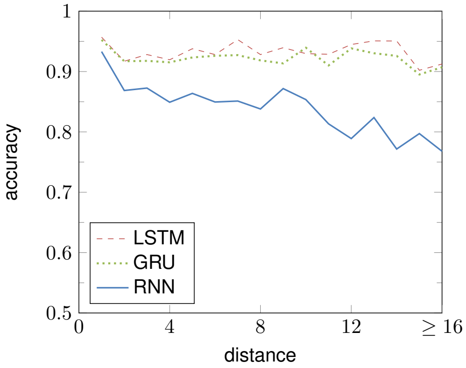
Long-Distance Interactions: Targeted Evaluation

evaluation with contrastive pairs: LingEval97 [Sennrich, EACL 2017]

	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

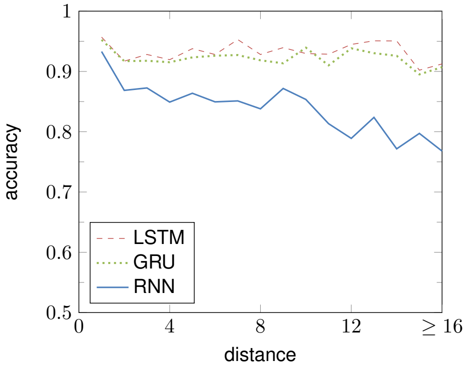
subject-verb agreement

- EN→DE WMT systems trained with Nematus
- targeted evaluation of subject-verb agreement with Lingeval97



Long-Distance Interactions: RNN vs. GRU vs. LSTM

- EN→DE WMT systems trained with Nematus
- targeted evaluation of subject-verb agreement with Lingeval97



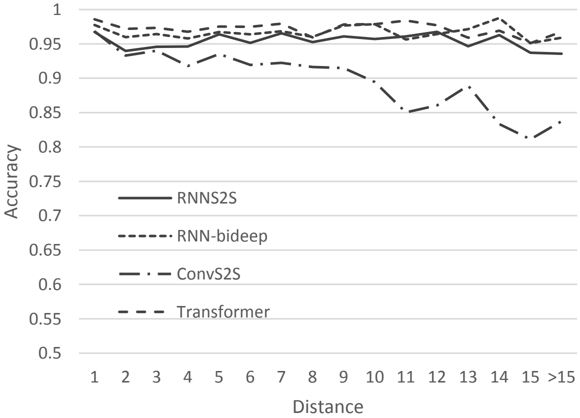
GRU/LSTM much more stable than RNN for long distances

Evaluating NMT Architectures

[Tang, Müller, Rios, Sennrich, EMNLP 2018]




- EN→DE WMT systems trained with Sockeye
- targeted evaluation of subject-verb agreement with Lingeval97

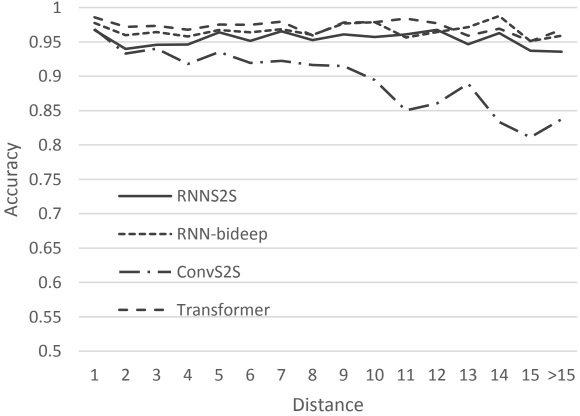


Evaluating NMT Architectures

[Tang, Müller, Rios, Sennrich, EMNLP 2018]




- EN→DE WMT systems trained with Sockeye
- targeted evaluation of subject-verb agreement with Lingeval97



no evidence that Transformer or ConvS2S outperform LSTM for long-distance interactions

Long Sentences: Conclusions

- strongest evidence for weakness of NMT on long sentences comes from old systems
- discarding long sentences no longer necessary in NMT training
- BLEU does not tell us **why** a system performs poorly on long sentences
  - are translations too short?
    $\rightarrow$ train on long sentences; use global scores
  - is grammaticality poor?
    $\rightarrow$ architectures matter, but long-distance interactions modelled well by GRU/LSTM and Transformer

Revisiting Challenges in NMT

Targeted Analysis: Word Sense Disambiguation

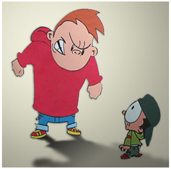
system	sentence
source	Dort wurde er von dem **Schläger** und einer weiteren männl. Person erneut angegriffen.
reference	There he was attacked again by his **original attacker** and another male.
our NMT	There he was attacked again by the **racket** and another male person.
Google	There he was again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

Schläger

Targeted Analysis: 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

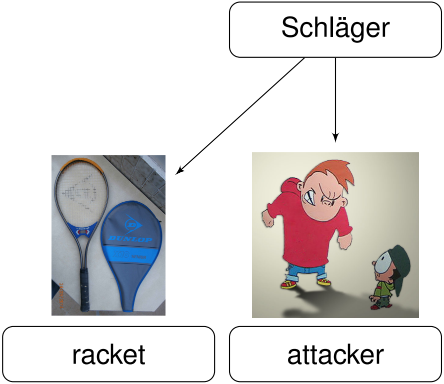


attacker

racket https://www.flickr.com/photos/128067141@N07/15157111178 / CC BY 2.0

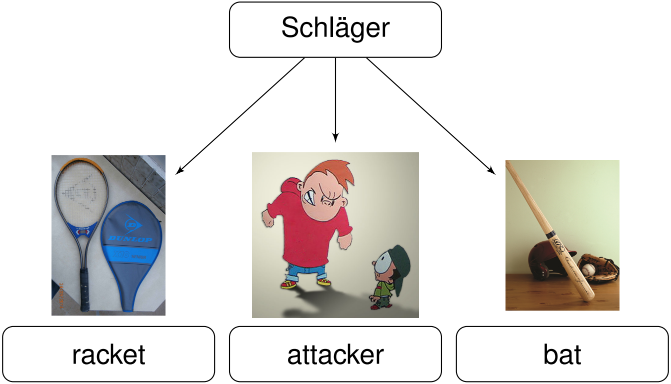
Targeted Analysis: Word Sense Disambiguation

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.



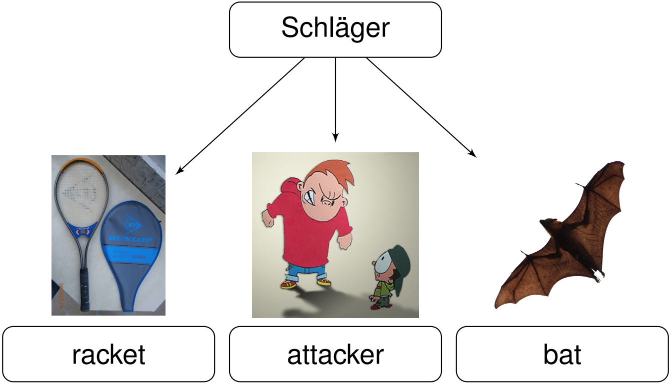
Targeted Analysis: Word Sense Disambiguation

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.



Targeted Analysis: 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.



Word Sense Disambiguation

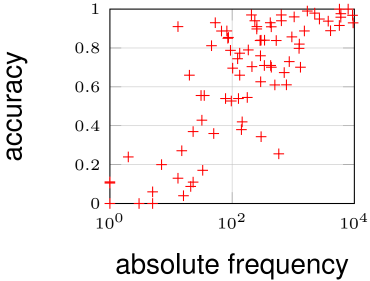
[Rios, Mascarell, Sennrich, WMT 2017]

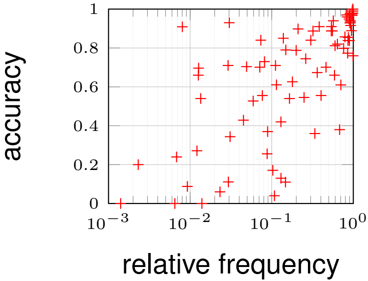
test set (ContraWSD)

- 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

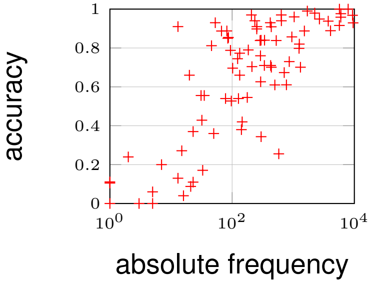
source:	*Also nahm ich meinen amerikanischen Reisepass und stellte mich in die **Schlange** für Extranjeros.*
reference:	*So I took my U.S. passport and got in the **line** for Extranjeros.*
contrastive:	*So I took my U.S. passport and got in the **snake** for Extranjeros.*
contrastive:	*So I took my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

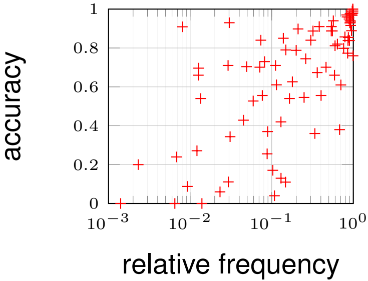
Word Sense Accuracy





Word Sense Accuracy





WSD is challenging, especially for rare word senses

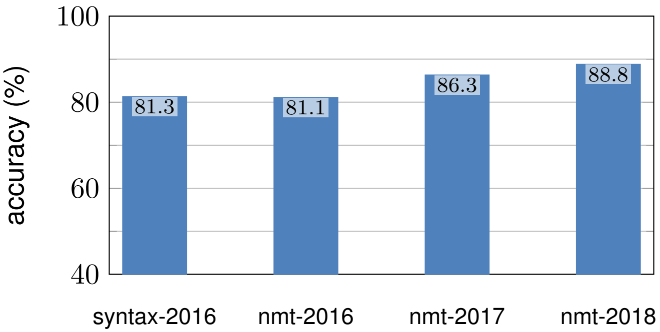
Word Sense Disambiguation: Measuring Progress

[Rios, Müller, Sennrich, WMT 2018]



- based on ContraWSD, but semi-automatic evaluation of 1-best output
- evaluating all WMT 2018 submissions, plus systems from previous years

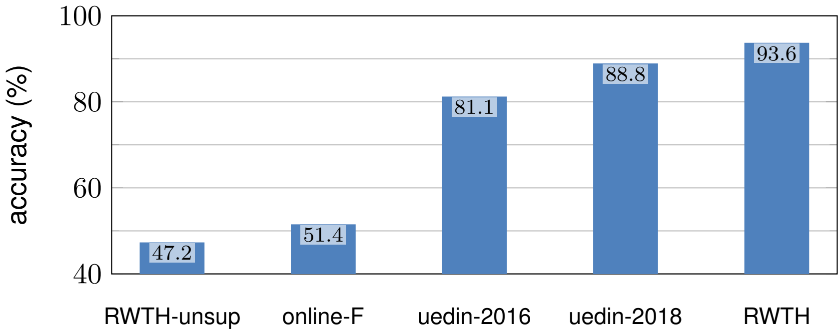
Results: Word Sense Disambiguation (uedin systems)



improvements to NMT system

- 2016: shallow RNN
- 2017: deep RNN; layer normalization; better ensembles; slightly more training data
- 2018: Transformer; more (noisy) training data

Results: Word Sense Disambiguation (selected systems)



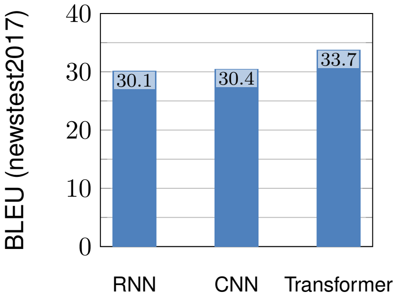
- WSD is big challenge for unsupervised NMT and rule-based system
- all neural systems at WMT18 > 81%
- big reduction in WSD errors in last 2 years

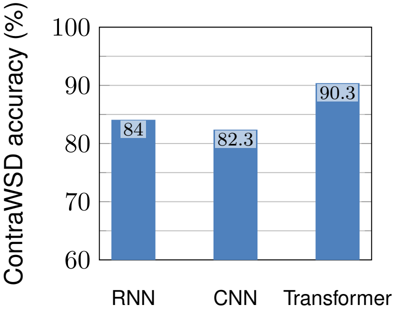
Evaluating NMT Architectures

[Tang, Müller, Rios, Sennrich, EMNLP 2018]




- comparing different architectures on same dataset
- Transformer no better than RNN at long-distance agreement
- interesting differences for word sense disambiguation:





Evaluating NMT Architectures

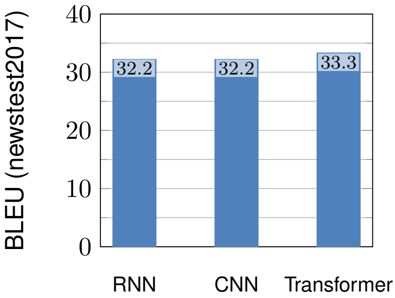
[Tang, Müller, Rios, Sennrich, EMNLP 2018]

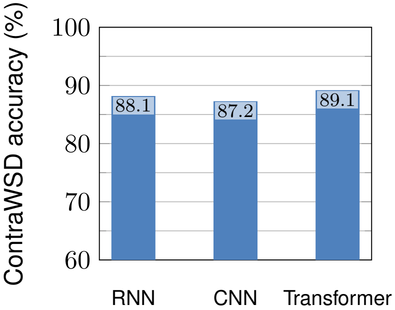



post-publication experiments:
models can be made more similar to Transformer with:

- multihead attention
- feedforward block
- layer normalization
- ...

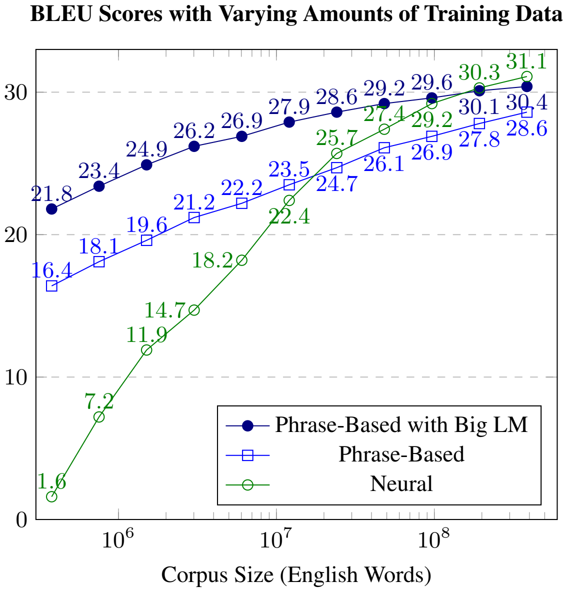
Transformer still ahead in WSD accuracy





Revisiting Challenges in NMT

Low-Resource Translation



Low-Resource Translation

aspects worth revisiting:

- phrase-based benefit from large LM
  but NMT can also improve with monolingual data
- there were general improvements in NMT
  do they move the point where NMT outperforms SMT?
- the NMT system was not optimized for low-resource NMT
  does tuning model to low-resource NMT help?

Low-Resource Translation: Monolingual Data

There is a large pool of methods to exploit monolingual data for NMT:

- ensembling with LM [Gülçehre et al., 2015]
- training objective: language modelling [Sennrich et al., 2016, Ramachandran et al., 2016]
- training objective: autoencoders [Luong et al., 2016, Currey et al., 2017]
- training objective: round-trip translation

  [Sennrich et al., 2016, He et al., 2016, Cheng et al., 2016]
- unsupervised NMT [Artetxe et al., 2017, Lample et al., 2017]

similarly, parallel data from other language pairs can help

[Zoph et al., 2016, Chen et al., 2017, Nguyen and Chiang, 2017]

Low-Resource Translation: Monolingual Data

There is a large pool of methods to exploit monolingual data for NMT:

- ensembling with LM [Gülçehre et al., 2015]
- training objective: language modelling [Sennrich et al., 2016, Ramachandran et al., 2016]
- training objective: autoencoders [Luong et al., 2016, Currey et al., 2017]
- training objective: round-trip translation

  [Sennrich et al., 2016, He et al., 2016, Cheng et al., 2016]
- unsupervised NMT [Artetxe et al., 2017, Lample et al., 2017]

similarly, parallel data from other language pairs can help

[Zoph et al., 2016, Chen et al., 2017, Nguyen and Chiang, 2017]

..but how far can we get with just parallel data?

Low-Resource Translation: Experiments

setup

- IWSLT14 German→English:
  - full set: 160 000 sentences (3.2M words)
  - smallest subset: 5000 sentences (100 000 words)
- phrase-based SMT with Moses
- neural MT with Nematus and BPE
- baseline: hyperparameters similar to uedin@WMT16
  shallow RNN, no dropout

comparison to [Koehn and Knowles, 2017]

[Koehn and Knowles, 2017]:
0.4 million to 385 million words of data (EN→ES WMT)

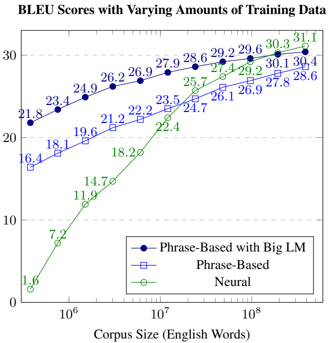
our experiments:
0.1 million to 3.2 million words of data (DE→EN IWSLT)

Low-Resource Translation: Experiments

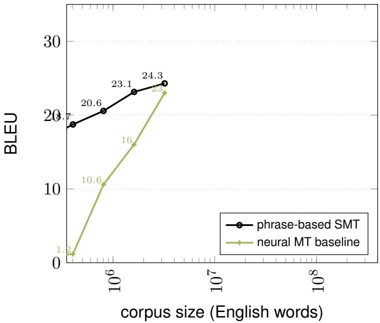
- general architecture improvements:
  - bideep RNN [Miceli Barone et al., 2017]
  - layer normalization [Ba et al., 2016]
  - label smoothing [Szegedy et al., 2016]
- choices optimized for low-resource scenario:
  - dropout [Srivastava et al., 2014]
  - tied embeddings [Press and Wolf, 2017]
  - smaller BPE vocabulary size for smaller data sets
  - smaller batch size for smaller data sets
  - lexical model [Nguyen and Chiang, 2018]

Low-Resource Translation: Results

[Koehn and Knowles, 2017]

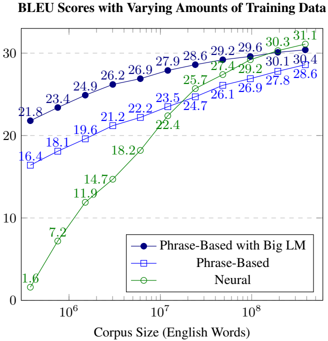


our experiments

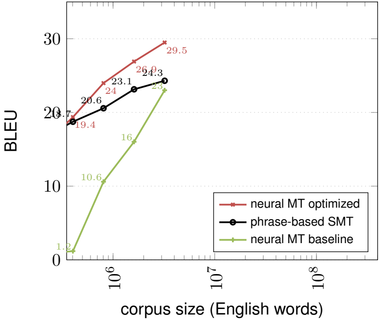


Low-Resource Translation: Results

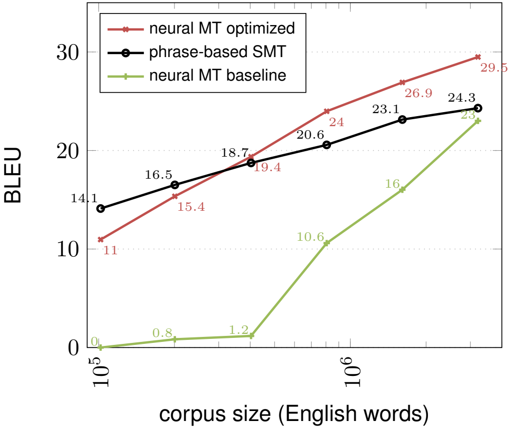
[Koehn and Knowles, 2017]



our experiments



Low-Resource Translation: Results



Low-Resource Translation: Results

system	BLEU by corpus size (words) 100k	3.2M
phrase-based SMT	14.1	24.3
NMT baseline	0.0	23.0
+dropout, tied embeddings, layer normalization, bideep RNN, label smoothing	6.0	30.0
+reduce BPE vocabulary (14k $\rightarrow$ 2k symbols)	9.3	-
+reduce batch size (4k $\rightarrow$ 1k tokens)	9.7	29.9
+lexical model	11.0	29.5

Low-Resource Translation: Conclusions

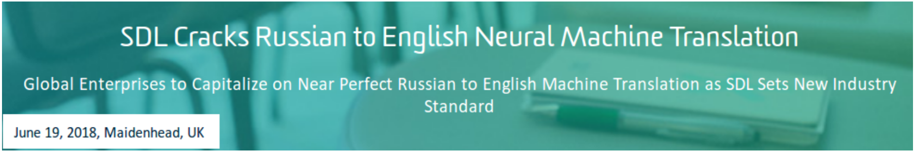
- the balance between PBSMT and NMT for low-resource settings is shifting with
  - general improvements in NMT
  - careful choice of hyperparameters and architectures for low-resource setting
- it is no longer true that we cannot train NMT on less than 1M words...
- ...but low-resource machine translation remains a challenge

Are There Any Challenges Left?

Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation

Microsoft reaches a historic milestone, using AI to match human performance in translating news from Chinese to English

March 14, 2018 | Allison Linn

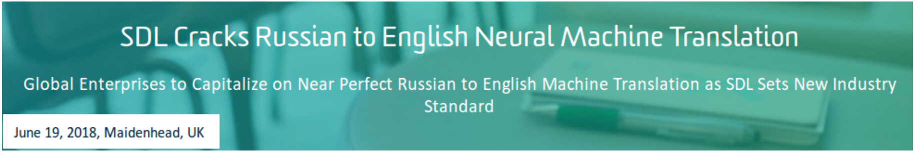


Are There Any Challenges Left?

Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation

Microsoft reaches a historic milestone, using AI to match human performance in translating news from Chinese to English

March 14, 2018 | Allison Linn



...extraordinary claims require extraordinary evidence

Achieving Human Parity

Microsoft reaches a historic milestone, using AI to match human performance in translating news from Chinese to English

March 14, 2018 | Allison Linn

laudable...

- follows best practices with WMT-style evaluation
- data released for scientific scrutiny (outputs, references, rankings)

Achieving Human Parity

Microsoft reaches a historic milestone, using AI to match human performance in translating news from Chinese to English

March 14, 2018 | Allison Linn

...but warrants further scrutiny

- failure to reject null hypothesis is not evidence of parity
- alternative hypothesis:
  human raters prefer human translations on a **document-level**
- rationale:
  - context helps raters understand text and spot semantic errors
  - discourse errors are invisible in sentence-level evaluation

A Case for Document-level Evaluation

[Läubli, Sennrich, Volk, EMNLP 2018]



can we reproduce Microsoft's finding with different evaluation protocol?

	original evaluation	our evaluation
test set	WMT17	WMT17 (native Chinese part)
system	Microsoft COMBO-6	Microsoft COMBO-6
raters	crowd-workers	**professional translators**
experimental unit	sentence	sentence / **document**
measurement	direct assessment	**pairwise ranking**
raters see reference	no	no
raters see source	yes	yes / no
ratings	$\geq$ 2,520 per system	$\approx$ 200 per setting

Which Text is Better?

Members of the public who find their cars obstructed by unfamiliar vehicles during their daily journeys can use the "Twitter Move Car" feature to address this distress when the driver of the unfamiliar vehicle cannot be reached.

A citizen whose car is obstructed by vehicle and is unable to contact the owner of the obstructing vehicle can use the "WeChat Move the Car" function to address the issue.

Which Text is Better?

Members of the public who find their cars obstructed by unfamiliar vehicles during their daily journeys can use the "Twitter Move Car" feature to address this distress when the driver of the unfamiliar vehicle cannot be reached.

On August 11, Xi'an traffic police WeChat service number "Xi'an traffic police" launched "WeChat mobile" service.

With the launch of the service, members of the public can tackle such problems in their daily lives by using the "WeChat Move" feature when an unfamiliar vehicle obstructs the movement of their vehicle while the driver is not at the scene. [...]

A citizen whose car is obstructed by vehicle and is unable to contact the owner of the obstructing vehicle can use the "WeChat Move the Car" function to address the issue.

The Xi'an Traffic Police WeChat official account "Xi'an Jiaojing" released the "WeChat Move the Car" service since August 11.

Once the service was released, a fellow citizen whose car was obstructed by another vehicle and where the driver of the vehicle was not present, the citizen could use the "WeChat Move the Car" function to address the issue. [...]

Which Text is Better?

市民在日常出行中,发现爱车被陌生车辆阻碍了,在联系不上陌生车辆司机的情况下,可以使用“微信挪车”功能解决这一困扰。

8月11日起,西安交警微信服务号“西安交警”推出“微信挪车”服务。

这项服务推出后,日常生活中,市民如遇陌生车辆在驾驶人不在现场的情况下阻碍自己车辆行驶时,就可通过使用“微信挪车”功能解决此类问题。[...]

Members of the public who find their cars obstructed by unfamiliar vehicles during their daily journeys can use the "Twitter Move Car" feature to address this distress when the driver of the unfamiliar vehicle cannot be reached.

On August 11, Xi'an traffic police WeChat service number "Xi'an traffic police" launched "WeChat mobile" service.

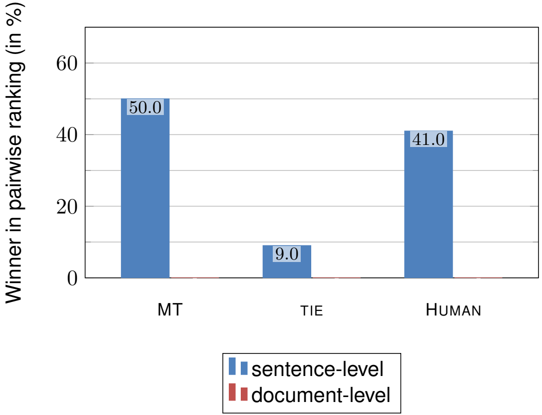
With the launch of the service, members of the public can tackle such problems in their daily lives by using the "WeChat Move" feature when an unfamiliar vehicle obstructs the movement of their vehicle while the driver is not at the scene. [...]

A citizen whose car is obstructed by vehicle and is unable to contact the owner of the obstructing vehicle can use the "WeChat Move the Car" function to address the issue.

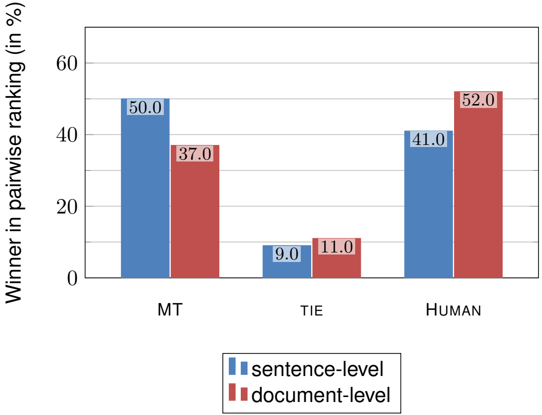
The Xi'an Traffic Police WeChat official account "Xi'an Jiaojing" released the "WeChat Move the Car" service since August 11.

Once the service was released, a fellow citizen whose car was obstructed by another vehicle and where the driver of the vehicle was not present, the citizen could use the "WeChat Move the Car" function to address the issue. [...]

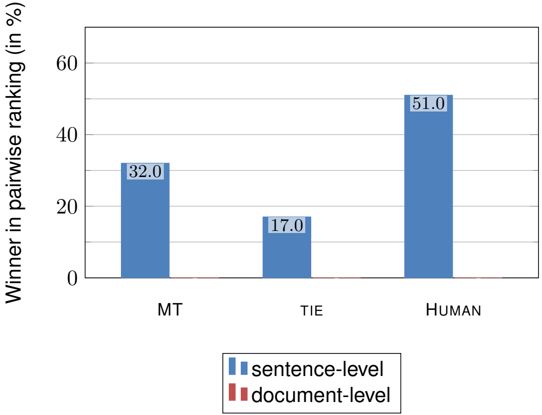
Evaluation Results: Bilingual Assessment



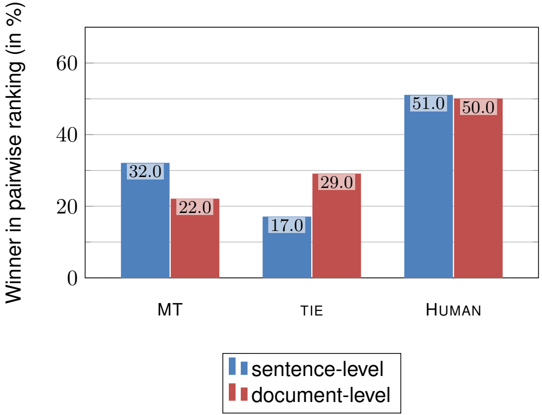
Evaluation Results: Bilingual Assessment



Evaluation Results: Monolingual Assessment



Evaluation Results: Monolingual Assessment



A Case for Document-level Evaluation

- document-level ratings show significant preference for HUMAN
- preference for HUMAN is even stronger in monolingual evaluation

Conclusions

- distinguishing MT from human translations becomes harder with increasing quality
- document-level evaluation shows some limitations of current NMT systems

Conclusions

- NMT has made tremendous progress in past years
  $\rightarrow$ to make progress, we need to regularly re-evaluate its weaknesses
- word sense disambiguation, long sentences, low-resource settings are still challenging, but no longer embarassingly bad
- plenty of challenges remain, such as document-level translation

Thank you for your attention

Resources

- WSD Test Suite:
  `https://github.com/a-rios/ContraWSD`
- Evaluation data on human parity:
  `https://github.com/laeubli/parity`

Bibliography I

Artetxe, M., Labaka, G., Agirre, E., and Cho, K. (2017).
Unsupervised Neural Machine Translation.
CoRR, abs/1710.11041.

Ba, L. J., Kiros, R., and Hinton, G. E. (2016).
Layer Normalization.
CoRR, abs/1607.06450.

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).

Castilho, S., Moorkens, J., Gaspari, F., Sennrich, R., Way, A., and Georgakopoulou, P. (2018).
Evaluating MT for massive open online courses.
Machine Translation.

Chen, Y., Liu, Y., Cheng, Y., and Li, V. O. (2017).
A Teacher-Student Framework for Zero-Resource Neural Machine Translation.
In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1925–1935, Vancouver, Canada. Association for Computational Linguistics.

Cheng, Y., Xu, W., He, Z., He, W., Wu, H., Sun, M., and Liu, Y. (2016).
Semi-Supervised Learning for Neural Machine Translation.
In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1965–1974, Berlin, Germany.

Bibliography II

Cho, K., van Merrienboer, B., Bahdanau, D., and Bengio, Y. (2014).
On the Properties of Neural Machine Translation: Encoder–Decoder Approaches.
In Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation, pages 103–111, Doha, Qatar. Association for Computational Linguistics.

Currey, A., Miceli Barone, A. V., and Heafield, K. (2017).
Copied Monolingual Data Improves Low-Resource Neural Machine Translation.
In Proceedings of the Second Conference on Machine Translation, pages 148–156, Copenhagen, Denmark. 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.

Gülçehre, Ç., 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.

He, D., Xia, Y., Qin, T., Wang, L., Yu, N., Liu, T., and Ma, W.-Y. (2016).
Dual Learning for Machine Translation.
In Lee, D. D., Sugiyama, M., Luxburg, U. V., Guyon, I., and Garnett, R., editors, Advances in Neural Information Processing Systems 29, pages 820–828. Curran Associates, Inc.

Johansen, A. R., Hansen, J. M., Obeid, E. K., Sønderby, C. K., and Winther, O. (2016).
Neural Machine Translation with Characters and Hierarchical Encoding.
CoRR, abs/1610.06550.

Bibliography III

Koehn, P. and Knowles, R. (2017).
Six Challenges for Neural Machine Translation.
In Proceedings of the First Workshop on Neural Machine Translation, pages 28–39, Vancouver. Association for Computational Linguistics.

Lample, G., Denoyer, L., and Ranzato, M. (2017).
Unsupervised Machine Translation Using Monolingual Corpora Only.
CoRR, abs/1711.00043.

Läubli, S., Sennrich, R., and Volk, M. (2018).
Has Neural Machine Translation Achieved Human Parity? A Case for Document-level Evaluation.
In EMNLP 2018, Brussels, Belgium. Association for Computational Linguistics.

Luong, M., Le, Q. V., Sutskever, I., Vinyals, O., and Kaiser, L. (2016).
Multi-task Sequence to Sequence Learning.
In ICLR 2016.

Miceli Barone, A. V., Helcl, J., Sennrich, R., Haddow, B., and Birch, A. (2017).
Deep Architectures for Neural Machine Translation.
In Proceedings of the Second Conference on Machine Translation, Volume 1: Research Papers, Copenhagen, Denmark. Association for Computational Linguistics.

Murray, K. and Chiang, D. (2018).
Correcting Length Bias in Neural Machine Translation.
In Proceedings of the Third Conference on Machine Translation, pages 212–223, Belgium, Brussels. Association for Computational Linguistics.

Bibliography IV

Nguyen, T. and Chiang, D. (2018).
Improving Lexical Choice in Neural Machine Translation.
In
Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Langua
pages 334–343, New Orleans, Louisiana. Association for Computational Linguistics.

Nguyen, T. Q. and Chiang, D. (2017).
Transfer Learning across Low-Resource, Related Languages for Neural Machine Translation.
In Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers), pages 296–301, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Press, O. and Wolf, L. (2017).
Using the Output Embedding to Improve Language Models.
In Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL), Valencia, Spain.

Ramachandran, P., Liu, P. J., and Le, Q. V. (2016).
Unsupervised pretraining for sequence to sequence learning.
arXiv preprint arXiv:1611.02683.

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.

Rios, A., Müller, M., and Sennrich, R. (2018).
The Word Sense Disambiguation Test Suite at WMT18.
In Proceedings of the Third Conference on Machine Translation, pages 594–602, Belgium, Brussels. Association for Computational Linguistics.

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., Haddow, B., and Birch, A. (2016).
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.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014).
Dropout: A Simple Way to Prevent Neural Networks from Overfitting.
Journal of Machine Learning Research, 15:1929–1958.

Sutskever, I., Vinyals, O., and Le, Q. V. (2014).
Sequence to Sequence Learning with Neural Networks.
In
Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, pages 3104–3112, Montreal, Quebec, Canada.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z. (2016).
Rethinking the Inception Architecture for Computer Vision.
In 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 2818–2826.

Tang, G., Müller, M., Rios, A., and Sennrich, R. (2018).
Why Self-Attention? A Targeted Evaluation of Neural Machine Translation Architectures.
In EMNLP 2018, Brussels, Belgium. Association for Computational Linguistics.

Bibliography VI

Tu, Z., Liu, Y., Shang, L., Liu, X., and Li, H. (2017).
Neural Machine Translation with Reconstruction.
In
Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA., pages 3097–3103.

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.

Zoph, B., Yuret, D., May, J., and Knight, K. (2016).
Transfer Learning for Low-Resource Neural Machine Translation.
CoRR, abs/1604.02201.