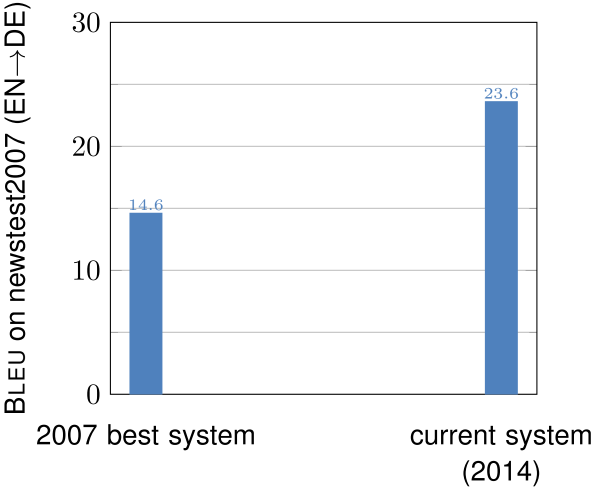
Neural Machine Translation: Breaking the Performance Plateau

Rico Sennrich

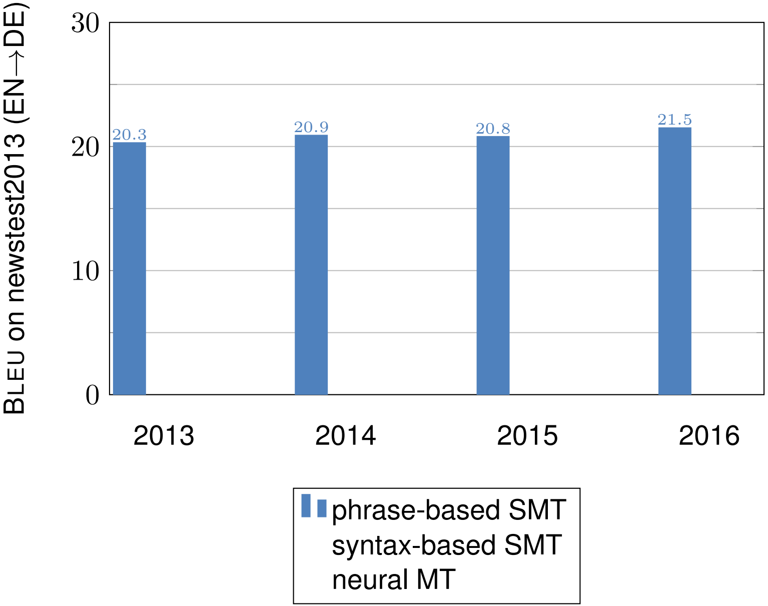
Institute for Language, Cognition and Computation
University of Edinburgh

July 4 2016

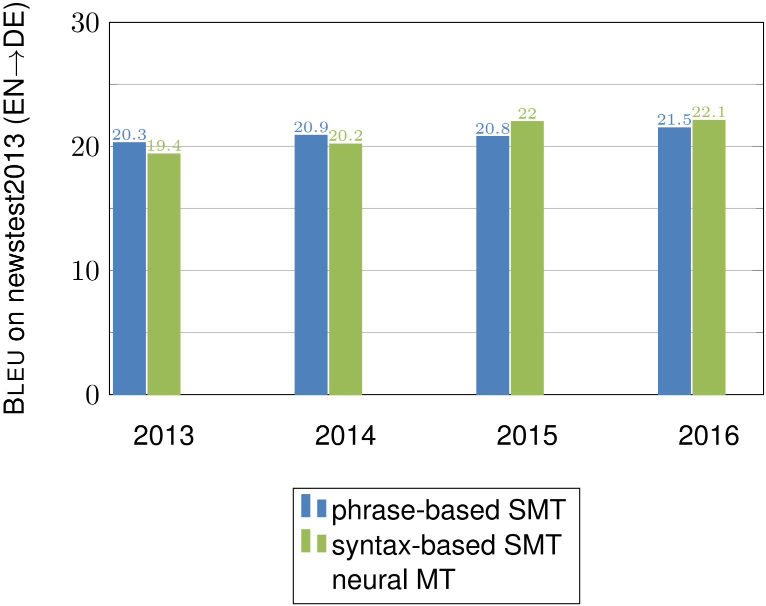
Is Machine Translation Getting Better Over Time? [Graham et al., 2014]



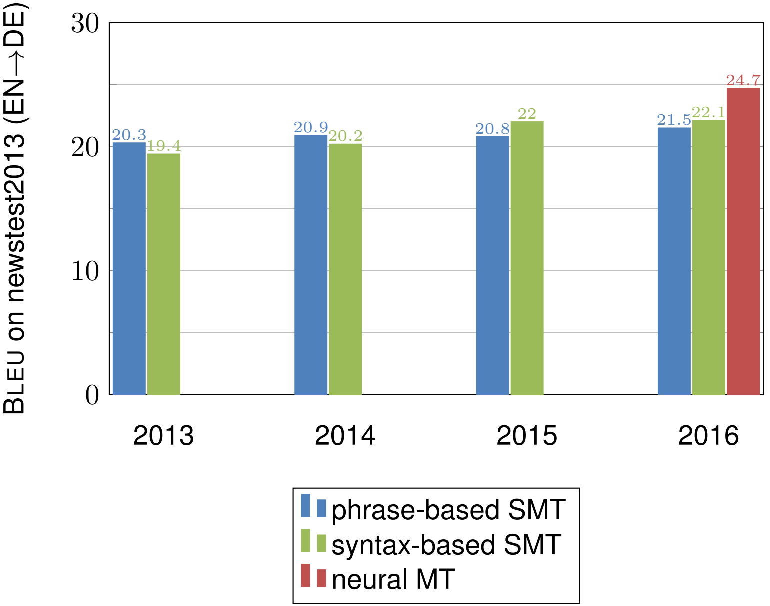
Edinburgh's WMT Results Over the Years



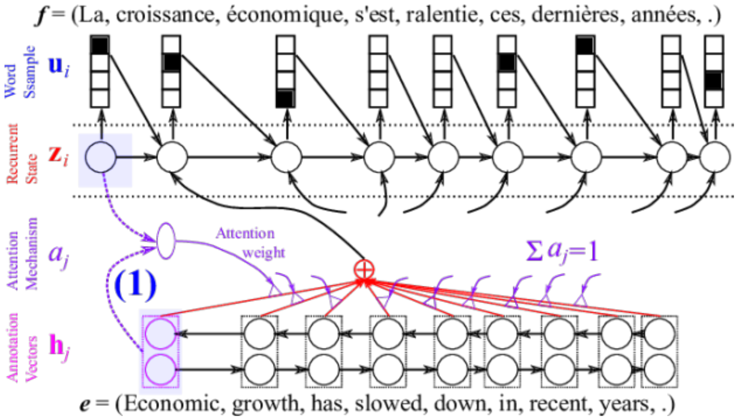
Edinburgh's WMT Results Over the Years



Edinburgh's WMT Results Over the Years



Neural Machine Translation [Bahdanau et al., 2015]



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

Why Neural Machine Translation?

qualitative differences

- main strength of neural MT: improved grammaticality [Neubig et al., 2015]

phrase-based SMT

- strong independence assumptions
- log-linear combination of many “weak” features

neural MT

- output conditioned on full source text and target history
- end-to-end trained model

Example (WMT16 EN→DE)

source	But he wants an international reporter **to be there** to write about it.
reference	Aber er will , **dass** ein internationaler Reporter **anwesend ist** , um dort zu schreiben .
PBSMT	Aber er will einen internationalen Reporter **zu sein** , darüber zu schreiben .
SBSMT	Aber er will einen internationalen Reporter , **um dort zu sein** , über sie zu schreiben .
neural MT	Aber er will , **dass** ein internationaler Reporter **da ist** , um darüber zu schreiben .

Recent Advances in Neural MT

- some problems:
  - networks have fixed vocabulary
    $\rightarrow$ poor translation of rare/unknown words
  - models are trained on parallel data; how do we use monolingual data?
- recent solutions:
  - subword models allow translation of rare/unknown words [Sennrich et al., 2016b]
  - train on back-translated monolingual data [Sennrich et al., 2016a]

Problem with Word-level Models

they charge a **carry-on bag fee**.
sie erheben eine **Hand|gepäck|gebühr**.

- Neural MT architectures have small and fixed vocabulary
- translation is an **open-vocabulary** problem
  - productive word formation (example: compounding)
  - names (may require transliteration)

Why Subword Models?

transparent translations

- many translations are semantically/phonologically transparent
  $\rightarrow$ translation via subword units possible
- morphologically complex words (e.g. compounds):
  - solar system (English)
  - Sonnen|system (German)
  - Nap|rendszer (Hungarian)
- named entities:
  - Barack Obama (English; German)
  - Барак Обама (Russian)
  - バラク・オバマ (ba-ra-ku o-ba-ma) (Japanese)
- cognates and loanwords:
  - claustrophobia (English)
  - Klaustrophobie (German)
  - Клаустрофобия (Russian)

Examples

system	sentence
source	health research institutes
reference	Gesundheitsforschungsinstitute
word-level	Forschungsinstitute
character bigrams	Fo|rs|ch|un|gs|in|st|it|ut|io|ne|n
joint BPE	Gesundheits|forsch|ungsin|stitute
source	rakfisk
reference	ракфиска (rakfiska)
word-level	rakfisk $\rightarrow$ UNK $\rightarrow$ rakfisk
character bigrams	ra|kf|is|k $\rightarrow$ ра|кф|ис|к (ra|kf|is|k)
joint BPE	rak|f|isk $\rightarrow$ рак|ф|иска (rak|f|iska)

Monolingual Training Data

why monolingual data for phrase-based SMT?

- relax independence assumptions ✓
- more training data ✓
- more appropriate training data (domain adaptation) ✓

why monolingual data for neural MT?

- relax independence assumptions ✗
- more training data ✓
- more appropriate training data (domain adaptation) ✓

Monolingual Data in NMT

solutions

- previous work: combine NMT with separately trained LM [Gülçehre et al., 2015]
- our idea: decoder is already a language model
  $\rightarrow$ train encoder-decoder with added monolingual data

monolingual training instances

- how do we get approximation of source context?
  - dummy source context (moderately effective)
  - automatically back-translate monolingual data into source language

Results: WMT 15 English→German

system	BLEU
syntax-based	24.4
Neural MT baseline	22.0
+subwords	22.8
+back-translated data	25.7
+ensemble of 4	26.5

WMT16 Results (BLEU)

uedin-nmt	34.2
metamind	32.3
NYU-UMontreal	30.8
cambridge	30.6
uedin-syntax	30.6
KIT/LIMSI	29.1
KIT	29.0
uedin-pbmt	28.4
jhu-syntax	26.6

EN→DE

uedin-nmt	38.6
uedin-pbmt	35.1
jhu-pbmt	34.5
uedin-syntax	34.4
KIT	33.9
jhu-syntax	31.0

DE→EN

uedin-nmt	25.8
NYU-UMontreal	23.6
jhu-pbmt	23.6
cu-chimera	21.0
uedin-cu-syntax	20.9
cu-tamchyna	20.8
cu-TectoMT	14.7
cu-mergedtrees	8.2

EN→CS

uedin-nmt	31.4
jhu-pbmt	30.4
PJATK	28.3
cu-mergedtrees	13.3

CS→EN

uedin-pbmt	35.2
uedin-nmt	33.9
uedin-syntax	33.6
jhu-pbmt	32.2
LIMSI	31.0

RO→EN

QT21-HimL-SysComb	28.9
uedin-nmt	28.1
RWTH-SYSCOMB	27.1
uedin-pbmt	26.8
uedin-lmu-hiero	25.9
KIT	25.8
lmu-cuni	24.3
LIMSI	23.9
jhu-pbmt	23.5
usfd-rescoring	23.1

EN→RO

uedin-nmt	26.0
amu-uedin	25.3
jhu-pbmt	24.0
LIMSI	23.6
AFRL-MITLL	23.5
NYU-UMontreal	23.1
AFRL-MITLL-verb-annot	20.9

EN→RU

amu-uedin	29.1
NRC	29.1
uedin-nmt	28.0
AFRL-MITLL	27.6
AFRL-MITLL-contrast	27.0

RU→EN

WMT16 Results (BLEU)

uedin-nmt	34.2
metamind	32.3
NYU-UMontreal	30.8
cambridge	30.6
uedin-syntax	30.6
KIT/LIMSI	29.1
KIT	29.0
uedin-pbmt	28.4
jhu-syntax	26.6

EN→DE

uedin-nmt	38.6
uedin-pbmt	35.1
jhu-pbmt	34.5
uedin-syntax	34.4
KIT	33.9
jhu-syntax	31.0

DE→EN

uedin-nmt	25.8
NYU-UMontreal	23.6
jhu-pbmt	23.6
cu-chimera	21.0
uedin-cu-syntax	20.9
cu-tamchyna	20.8
cu-TectoMT	14.7
cu-mergedtrees	8.2

EN→CS

uedin-nmt	31.4
jhu-pbmt	30.4
PJATK	28.3
cu-mergedtrees	13.3

CS→EN

uedin-pbmt	35.2
uedin-nmt	33.9
uedin-syntax	33.6
jhu-pbmt	32.2
LIMSI	31.0

RO→EN

QT21-HimL-SysComb	28.9
uedin-nmt	28.1
RWTH-SYSCOMB	27.1
uedin-pbmt	26.8
uedin-lmu-hiero	25.9
KIT	25.8
lmu-cuni	24.3
LIMSI	23.9
jhu-pbmt	23.5
usfd-rescoring	23.1

EN→RO

uedin-nmt	26.0
amu-uedin	25.3
jhu-pbmt	24.0
LIMSI	23.6
AFRL-MITLL	23.5
NYU-UMontreal	23.1
AFRL-MITLL-verb-annot	20.9

EN→RU

amu-uedin	29.1
NRC	29.1
uedin-nmt	28.0
AFRL-MITLL	27.6
AFRL-MITLL-contrast	27.0

RU→EN

- Edinburgh NMT

WMT16 Results (BLEU)

uedin-nmt	34.2
metamind	32.3
NYU-UMontreal	30.8
cambridge	30.6
uedin-syntax	30.6
KIT/LIMSI	29.1
KIT	29.0
uedin-pbmt	28.4
jhu-syntax	26.6

EN→DE

uedin-nmt	38.6
uedin-pbmt	35.1
jhu-pbmt	34.5
uedin-syntax	34.4
KIT	33.9
jhu-syntax	31.0

DE→EN

uedin-nmt	25.8
NYU-UMontreal	23.6
jhu-pbmt	23.6
cu-chimera	21.0
uedin-cu-syntax	20.9
cu-tamchyna	20.8
cu-TectoMT	14.7
cu-mergedtrees	8.2

EN→CS

uedin-nmt	31.4
jhu-pbmt	30.4
PJATK	28.3
cu-mergedtrees	13.3

CS→EN

uedin-pbmt	35.2
uedin-nmt	33.9
uedin-syntax	33.6
jhu-pbmt	32.2
LIMSI	31.0

RO→EN

QT21-HimL-SysComb	28.9
uedin-nmt	28.1
RWTH-SYSCOMB	27.1
uedin-pbmt	26.8
uedin-lmu-hiero	25.9
KIT	25.8
lmu-cuni	24.3
LIMSI	23.9
jhu-pbmt	23.5
usfd-rescoring	23.1

EN→RO

uedin-nmt	26.0
amu-uedin	25.3
jhu-pbmt	24.0
LIMSI	23.6
AFRL-MITLL	23.5
NYU-UMontreal	23.1
AFRL-MITLL-verb-annot	20.9

EN→RU

amu-uedin	29.1
NRC	29.1
uedin-nmt	28.0
AFRL-MITLL	27.6
AFRL-MITLL-contrast	27.0

RU→EN

- Edinburgh NMT
- System Combination with Edinburgh NMT

Neural MT and Phrase-based SMT

	Neural MT	Phrase-based SMT
translation quality	✓	
model size	✓	
training time		✓
model interpretability		✓
decoding efficiency	✓	✓
toolkits	✓ (for simplicity)	✓ (for maturity)
special hardware requirement	GPU	lots of RAM

Conclusions and Outlook

conclusions

- neural MT is SOTA on many tasks
- subword models and back-translated data contributed to success

future predictions

- performance lead over phrase-based SMT will increase
- industry adoption will happen, but beware:
  - some hard things are suddenly easy (incremental training)
  - some easy things are suddenly hard (manual changes to model)
- exciting research opportunities
  - relax independence assumptions:
    document-level translation, multimodal input, ...
  - share parts of network between tasks:
    universal translation models, multi-task models, ...

Bibliography I

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).

Graham, Y., Baldwin, T., Moffat, A., and Zobel, J. (2014).
Is Machine Translation Getting Better over Time?
In Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, pages 443–451, Gothenburg, Sweden. Association for Computational Linguistics.

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.

Neubig, G., Morishita, M., and Nakamura, S. (2015).
Neural Reranking Improves Subjective Quality of Machine Translation: NAIST at WAT2015.
In Proceedings of the 2nd Workshop on Asian Translation (WAT2015), pages 35–41, Kyoto, Japan.

Sennrich, R., Haddow, B., and Birch, A. (2016a).
Improving Neural Machine Translation Models with Monolingual Data.
In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL 2016), 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 (ACL 2016), Berlin, Germany.