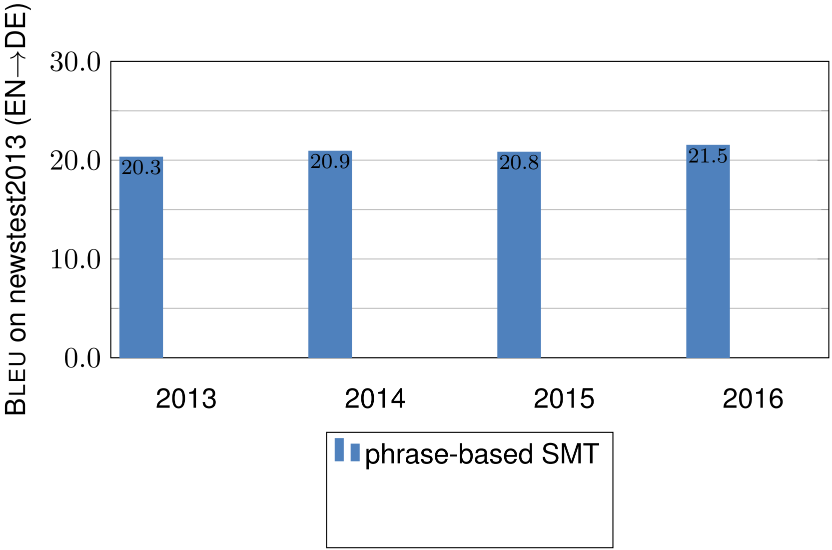
Neural Machine Translation: Breaking the Performance Plateau

Rico Sennrich

Institute for Language, Cognition and Computation
University of Edinburgh

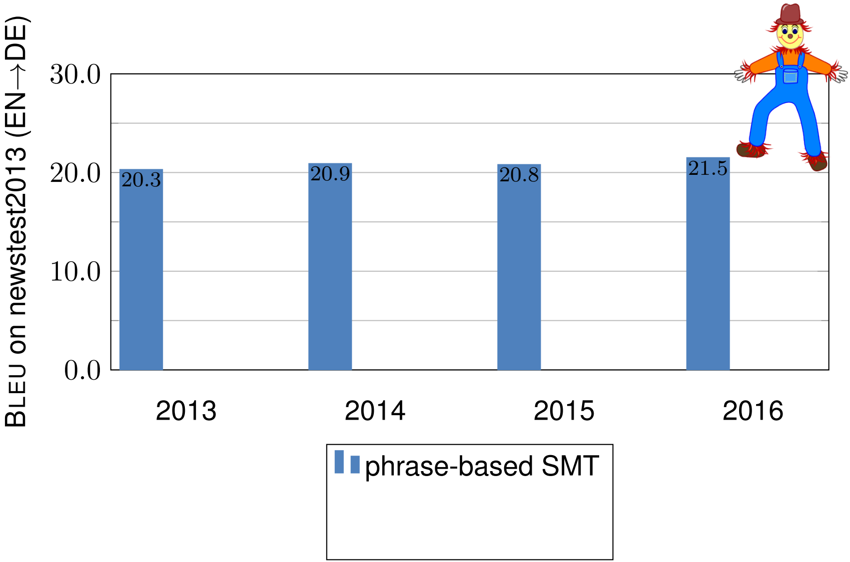
October 29 2016

Edinburgh's* WMT results over the years



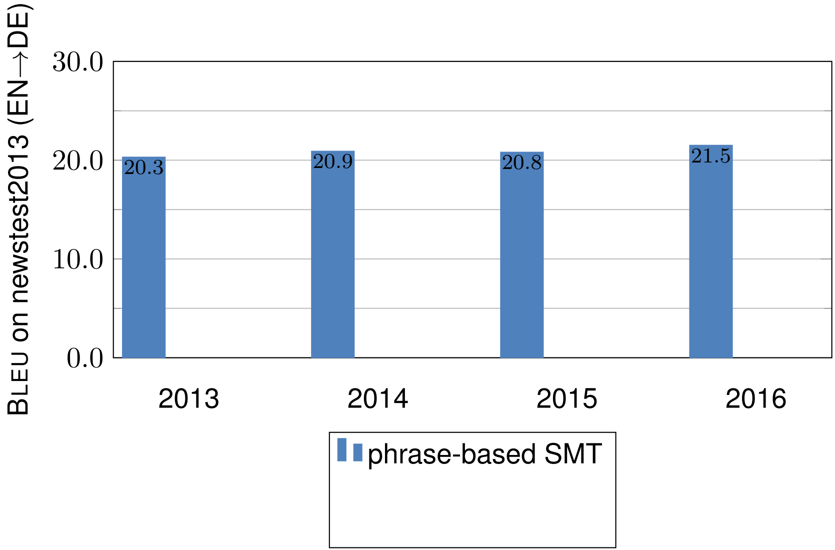
*NMT 2015 from U. Montréal: `https://sites.google.com/site/acl16nmt/`

Edinburgh's* WMT results over the years



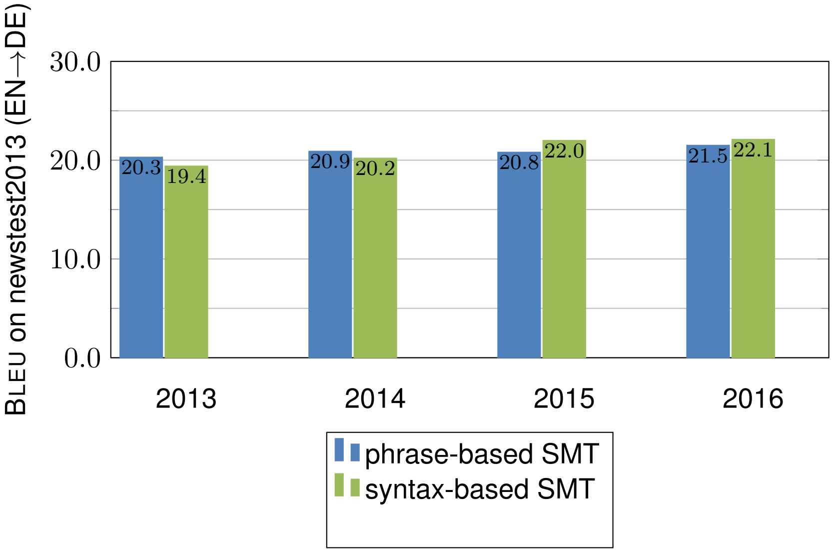
*NMT 2015 from U. Montréal: `https://sites.google.com/site/acl16nmt/`

Edinburgh's* WMT results over the years



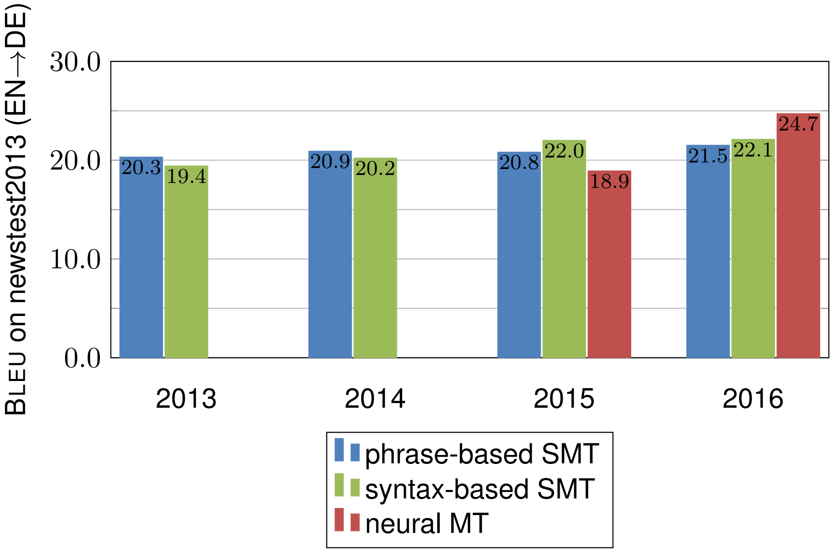
*NMT 2015 from U. Montréal: `https://sites.google.com/site/acl16nmt/`

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

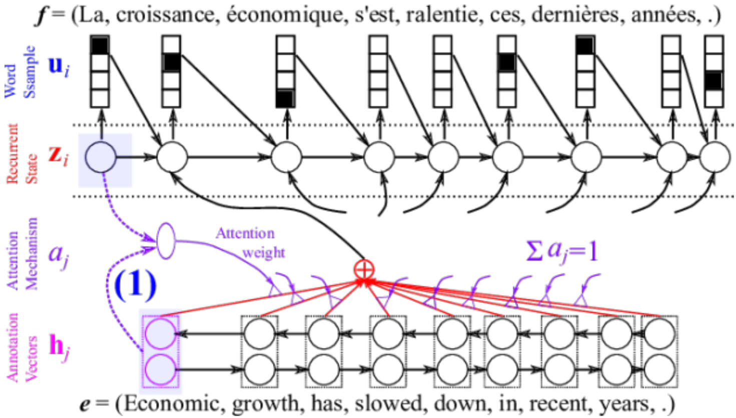


*NMT 2015 from U. Montréal: `https://sites.google.com/site/acl16nmt/`

1. Recent advances in neural MT

2. Towards using neural MT in production
   - things that are suddenly easy(er)
   - things that are suddenly hard(er)
   - things that are still hard

Neural Machine Translation [Bahdanau et al., 2015]



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

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?

Problem with word-level models

they charge a **carry-on** **bag** **fee**.
sie erheben eine **Handgepäckgebühr**.

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

Why subword units?

transparent translations

- some translations are semantically/phonologically transparent
- morphologically complex words (e.g. compounds):
  - solar system (English)
  - Sonnen|system (German)
  - Nap|rendszer (Hungarian)
- named entities:
  - Obama(English; German)
  - Обама (Russian)
  - オバマ (o-ba-ma) (Japanese)
- cognates and loanwords:
  - claustrophobia(English)
  - Klaustrophobie(German)
  - Клаустрофобия (Russian)

Choice of subword unit

- characters?
  $\rightarrow$ works, but inefficient
  (recent work on increasing efficiency [Lee et al., 2016])
- algorithms employed in SMT? (finite-state morphology; Morfessor)
  $\rightarrow$ no control over symbol vocabulary

byte pair encoding (BPE)

- compression algorithm adapted to word segmentation
- frequency-based
- single hyperparameter which controls symbol vocabulary size

Byte pair encoding for word segmentation

bottom-up character merging

- iteratively replace most frequent pair of symbols ('A','B') with 'AB'
- apply on dictionary, not on full text (for efficiency)
- output vocabulary: character vocabulary + one symbol per merge

word	freq
'l o w </w>'	5
'l o w e r </w>'	2
'n e w e s t </w>'	6
'w i d e s t </w>'	3

freq	symbol pair	new symbol

Byte pair encoding for word segmentation

bottom-up character merging

- iteratively replace most frequent pair of symbols ('A','B') with 'AB'
- apply on dictionary, not on full text (for efficiency)
- output vocabulary: character vocabulary + one symbol per merge

word	freq
'l o w </w>'	5
'l o w e r </w>'	2
'n e w es t </w>'	6
'w i d es t </w>'	3

freq	symbol pair		new symbol
9	('e', 's')	$\rightarrow$	'es'

Byte pair encoding for word segmentation

bottom-up character merging

- iteratively replace most frequent pair of symbols ('A','B') with 'AB'
- apply on dictionary, not on full text (for efficiency)
- output vocabulary: character vocabulary + one symbol per merge

word	freq
'l o w </w>'	5
'l o w e r </w>'	2
'n e w est </w>'	6
'w i d est </w>'	3

freq	symbol pair		new symbol
9	('e', 's')	$\rightarrow$	'es'
9	('es', 't')	$\rightarrow$	'est'

Byte pair encoding for word segmentation

bottom-up character merging

- iteratively replace most frequent pair of symbols ('A','B') with 'AB'
- apply on dictionary, not on full text (for efficiency)
- output vocabulary: character vocabulary + one symbol per merge

word	freq
'l o w </w>'	5
'l o w e r </w>'	2
'n e w est</w>'	6
'w i d est</w>'	3

freq	symbol pair		new symbol
9	('e', 's')	→	'es'
9	('es', 't')	→	'est'
9	('est', '</w>')	→	'est</w>'

Byte pair encoding for word segmentation

bottom-up character merging

- iteratively replace most frequent pair of symbols ('A','B') with 'AB'
- apply on dictionary, not on full text (for efficiency)
- output vocabulary: character vocabulary + one symbol per merge

word	freq
'l o w </w>'	5
'l o w e r </w>'	2
'n e w est</w>'	6
'w i d est</w>'	3

freq	symbol pair		new symbol
9	('e', 's')	→	'es'
9	('es', 't')	→	'est'
9	('est', '</w>')	→	'est</w>'
7	('l', 'o')	→	'lo'

Byte pair encoding for word segmentation

bottom-up character merging

- iteratively replace most frequent pair of symbols (’A’,’B’) with ’AB’
- apply on dictionary, not on full text (for efficiency)
- output vocabulary: character vocabulary + one symbol per merge

word	freq
’l o w </w>’	5
’l o w e r </w>’	2
’n e w est</w>’	6
’w i d est</w>’	3

freq	symbol pair		new symbol
9	(’e’, ’s’)	$\rightarrow$	’es’
9	(’es’, ’t’)	$\rightarrow$	’est’
9	(’est’, ’</w>’)	$\rightarrow$	’est</w>’
7	(’l’, ’o’)	$\rightarrow$	’lo’
7	(’lo’, ’w’)	$\rightarrow$	’low’
...			

Byte pair encoding for word segmentation

why BPE?

- open-vocabulary:
  learned operations can be applied to unknown words
- don't waste time on frequent character sequences
  $\rightarrow$ trade-off between text length and vocabulary size
- alternative view: character-level model on compressed text

'l o w e s t </w>'

('e', 's')	$\rightarrow$	'es'
('es', 't')	$\rightarrow$	'est'
('est', '</w>')	$\rightarrow$	'est</w>'
('l', 'o')	$\rightarrow$	'lo'
('lo', 'w')	$\rightarrow$	'low'

Byte pair encoding for word segmentation

why BPE?

- open-vocabulary:
  learned operations can be applied to unknown words
- don't waste time on frequent character sequences
  → trade-off between text length and vocabulary size
- alternative view: character-level model on compressed text

'l o w es t </w>'

('e', 's')	→	'es'
('es', 't')	→	'est'
('est', '</w>')	→	'est</w>'
('l', 'o')	→	'lo'
('lo', 'w')	→	'low'

Byte pair encoding for word segmentation

why BPE?

- open-vocabulary:
  learned operations can be applied to unknown words
- don't waste time on frequent character sequences
  → trade-off between text length and vocabulary size
- alternative view: character-level model on compressed text

'l o w est </w>'

('e', 's')	→	'es'
('es', 't')	→	'est'
('est', '</w>')	→	'est</w>'
('l', 'o')	→	'lo'
('lo', 'w')	→	'low'

Byte pair encoding for word segmentation

why BPE?

- open-vocabulary:
  learned operations can be applied to unknown words
- don't waste time on frequent character sequences
  $\rightarrow$ trade-off between text length and vocabulary size
- alternative view: character-level model on compressed text

'l o w est</w>'

('e', 's')	$\rightarrow$	'es'
('es', 't')	$\rightarrow$	'est'
('est', '</w>')	$\rightarrow$	'est</w>'
('l', 'o')	$\rightarrow$	'lo'
('lo', 'w')	$\rightarrow$	'low'

Byte pair encoding for word segmentation

why BPE?

- open-vocabulary:
  learned operations can be applied to unknown words
- don’t waste time on frequent character sequences
  → trade-off between text length and vocabulary size
- alternative view: character-level model on compressed text

’lo w est</w>’

(’e’, ’s’)	→	’es’
(’es’, ’t’)	→	’est’
(’est’, ’</w>’)	→	’est</w>’
(’l’, ’o’)	→	’lo’
(’lo’, ’w’)	→	’low’

Byte pair encoding for word segmentation

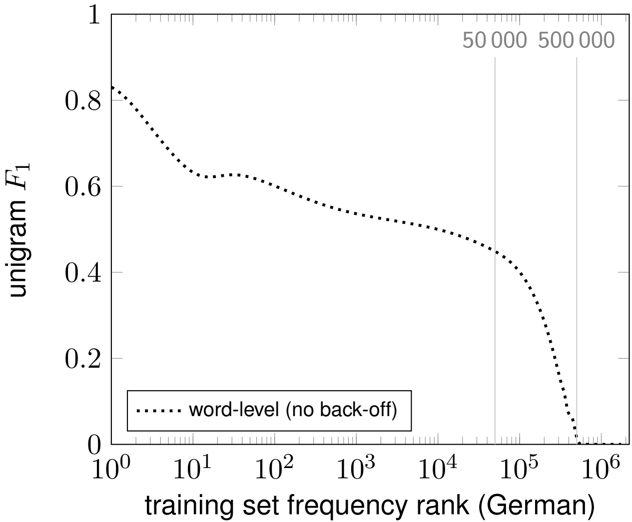
why BPE?

- open-vocabulary:
  learned operations can be applied to unknown words
- don't waste time on frequent character sequences
  $\rightarrow$ trade-off between text length and vocabulary size
- alternative view: character-level model on compressed text

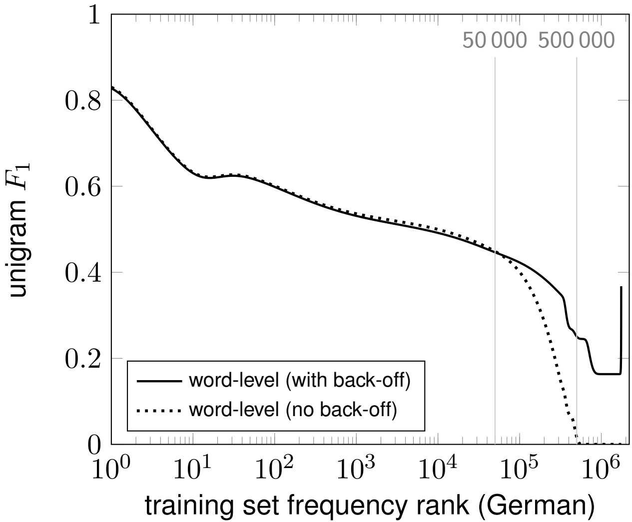
'low est</w>'

('e', 's')	$\rightarrow$	'es'
('es', 't')	$\rightarrow$	'est'
('est', '</w>')	$\rightarrow$	'est</w>'
('l', 'o')	$\rightarrow$	'lo'
('lo', 'w')	$\rightarrow$	'low'

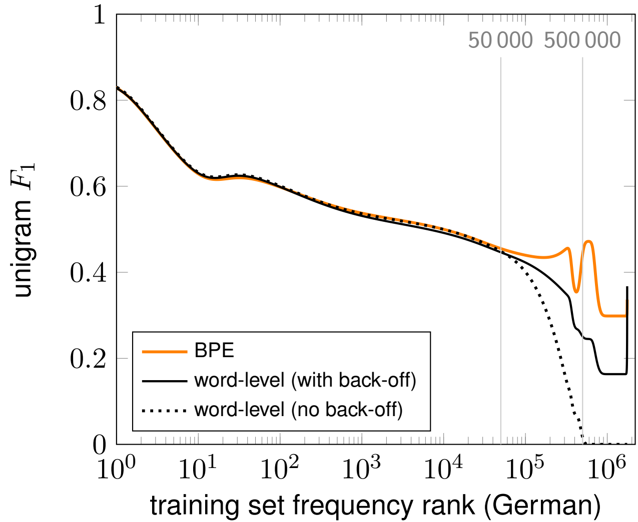
Unigram F_1 EN→DE



Unigram F_1 EN→DE



Unigram F_1 EN→DE



Examples

system	sentence
source	health research institutes
reference	Gesundheitsforschungsinstitute
word-level (with back-off)	Forschungsinstitute
BPE	Gesundheits|forsch|ungsin|stitute
source	rakfisk
reference	ракфиска (rakfiska)
word-level (with back-off)	rakfisk $\to$ UNK $\to$ rakfisk
BPE	rak|f|isk $\to$ рак|ф|иска (rak|f|iska)

Monolingual training data

Why Monolingual Data for Phrase-based SMT?

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

Why Monolingual Data for NMT?

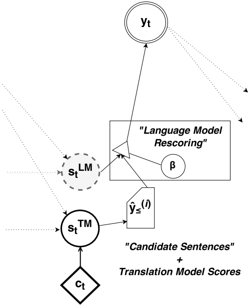
- more training data ✓
- more appropriate training data (domain adaptation) ✓
- relax independence assumptions ✗

Monolingual training data

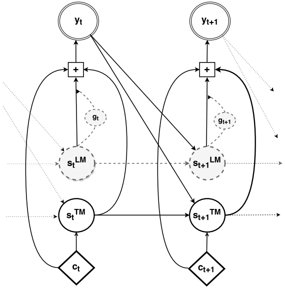
Related work [Gülçehre et al., 2015]

shallow fusion: rescore beam with language model

deep fusion: extra, LM-specific hidden layer



(a) Shallow Fusion (Sec. 4.1)



(b) Deep Fusion (Sec. 4.2)

[Gülçehre et al., 2015]

Training data: monolingual

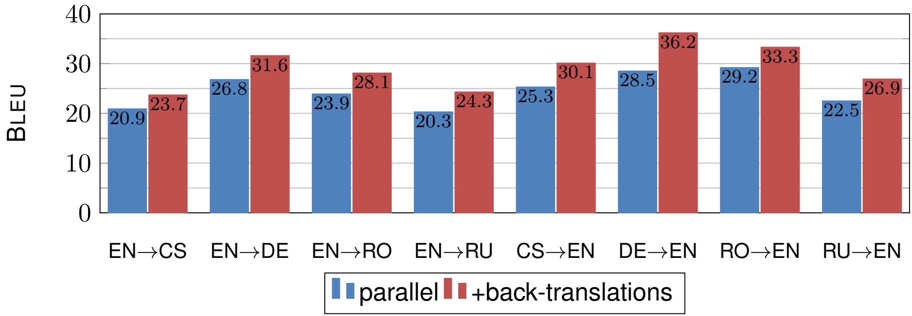
train NMT with monolingual data [Sennrich et al., 2016b]

- decoder is already a language model. Train encoder-decoder with added monolingual data
- how do we get approximation of context vector c_i?
  - dummy source context (moderately effective)
  - automatically back-translate monolingual data into source language $\rightarrow$ synthetic training instances with approximate c_i

Training data: monolingual

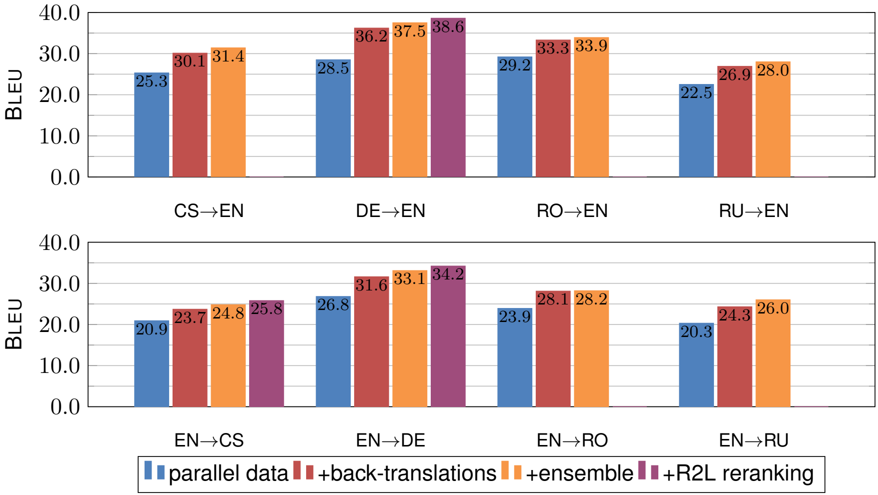
train NMT with monolingual data [Sennrich et al., 2016b]

- decoder is already a language model. Train encoder-decoder with added monolingual data
- how do we get approximation of context vector c_i?
  - dummy source context (moderately effective)
  - automatically back-translate monolingual data into source language $\rightarrow$ synthetic training instances with approximate c_i



Other techniques @WMT16

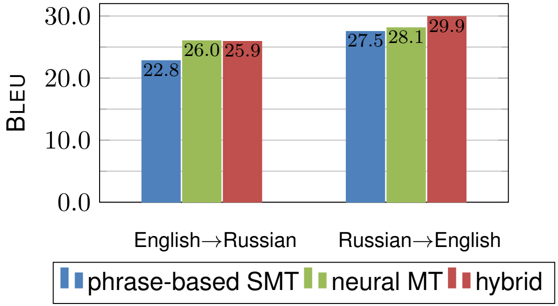
- ensembling of checkpoints
- bidirectional decoding (R2L reranking)



Phrase-based/neural MT hybridization

[Junczys-Dowmunt et al., 2016]

- use NMT as a feature function in phrase-based SMT
  $\rightarrow$ approximations and batching for efficiency
- effectiveness depends on quality of phrase-based and NMT system



WMT16 results

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

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

DE→EN

WMT16 results

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

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

DE→EN

- pure NMT

WMT16 results

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

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

DE→EN

- pure NMT
- NMT component

WMT16 results

uedin-nmt	25.8	1
NYU-UMontreal	23.6	2
jhu-pbmt	23.6	3
cu-chimera	21.0	4-5
cu-tamchyna	20.8	4-5
uedin-cu-syntax	20.9	6-7
online-B	22.7	6-7
online-A	19.5	15
cu-TectoMT	14.7	16
cu-mergedtrees	8.2	18

EN→CS

uedin-nmt	31.4	1
jhu-pbmt	30.4	2
online-B	28.6	3
PJATK	28.3	8-10
online-A	25.7	11
cu-mergedtrees	13.3	12

CS→EN

online-B	39.2	1-2
uedin-nmt	33.9	1-2
uedin-pbmt	35.2	3
uedin-syntax	33.6	4-5
online-A	30.8	4-6
jhu-pbmt	32.2	5-7
LIMSI	31.0	6-7

RO→EN

uedin-nmt	28.1	1-2
QT21-HimL-SysComb	28.9	1-2
KIT	25.8	3-7
uedin-pbmt	26.8	3-7
online-B	25.4	3-7
uedin-lmu-hiero	25.9	3-7
RWTH-SYSCOMB	27.1	3-7
LIMSI	23.9	8-10
lmu-cuni	24.3	8-10
jhu-pbmt	23.5	8-11
usfd-rescoring	23.1	10-12
online-A	19.2	11-12

EN→RO

WMT16 results

PROMT-rule	22.3	1
amu-uedin	25.3	2-4
online-B	23.8	2-5
uedin-nmt	26.0	2-5
online-G	26.2	3-5
NYU-UMontreal	23.1	6
jhu-pbmt	24.0	7-8
LIMSI	23.6	7-10
online-A	20.2	8-10
AFRL-MITLL-phr	23.5	9-10
AFRL-MITLL-verb	20.9	11
online-F	8.6	12

EN→RU

uedin-pbmt	23.4	1-4
online-G	20.6	1-4
online-B	23.6	1-4
UH-opus	23.1	1-4
PROMT-SMT	20.3	5
UH-factored	19.3	6-7
uedin-syntax	20.4	6-7
online-A	19.0	8
jhu-pbmt	19.1	9

FI→EN

amu-uedin	29.1	1-2
online-G	28.7	1-3
NRC	29.1	2-4
online-B	28.1	3-5
uedin-nmt	28.0	4-5
online-A	25.7	6-7
AFRL-MITLL-phr	27.6	6-7
AFRL-MITLL-contrast	27.0	8-9
PROMT-rule	20.4	8-9
online-F	13.5	10

RU→EN

online-G	15.4	1-3
abumatra-nmt	17.2	1-4
online-B	14.4	1-4
abumatran-combo	17.4	3-5
UH-opus	16.3	4-5
NYU-UMontreal	15.1	6-8
abumatran-pbsmt	14.6	6-8
online-A	13.0	6-8
jhu-pbmt	13.8	9-10
UH-factored	12.8	9-12
aalto	11.6	10-13
jhu-hltcoe	11.9	10-13
UUT	11.6	11-13

EN→FI

Neural Machine Translation

1. Recent advances in neural MT

2. Towards using neural MT in production
   - things that are suddenly easy(er)
   - things that are suddenly hard(er)
   - things that are still hard

Production use of neural MT

use of neural MT in production is only a matter of time

Production use of neural MT

use of neural MT in production ~~is only a matter of time~~ has begun

Production use of neural MT

use of neural MT in production ~~is only a matter of time~~ has begun

SYSTRAN announces the launch of its "Purely Neural MT" engine, a revolution for the machine translation market

Production use of neural MT

use of neural MT in production ~~is only a matter of time~~ has begun

SYSTRAN announces the launch of its "Purely Neural MT" engine, a revolution for the machine translation market

Google announces Neural Machine Translation to improve Google Translate

Production use of neural MT

use of neural MT in production ~~is only a matter of time~~ has begun

SYSTRAN announces the launch of its "Purely Neural MT" engine, a revolution for the machine translation market

Google announces Neural Machine Translation to improve Google Translate

WIPO goes Neural

Oct 4, 2016 | 590 views 41 Likes 3 Comments

Production use of neural MT

use of neural MT in production ~~is only a matter of time~~ has begun

SYSTRAN announces the launch of its "Purely Neural MT" engine, a revolution for the machine translation market

Google announces Neural Machine Translation to improve Google Translate

WIPO goes Neural

Oct 4, 2016 | 590 views | 41 Likes | 3 Comments

TraMOOC
Translation for Massive Open Online Courses

Production use of neural MT

use of neural MT in production ~~is only a matter of time~~ has begun

SYSTRAN announces the launch of its "Purely Neural MT" engine, a revolution for the machine translation market

Google announces Neural Machine Translation to improve Google Translate

WIPO goes Neural

Oct 4, 2016 | 590 views 41 Likes 3 Comments

TraMOOC Translation for Massive Open Online Courses

Health in my Language HimL

Production use of neural MT

use of neural MT in production ~~is only a matter of time~~ has begun

SYSTRAN announces the launch of its "Purely Neural MT" engine, a revolution for the machine translation market

Google announces Neural Machine Translation to improve Google Translate

WIPO goes Neural

Oct 4, 2016 | 590 views 41 Likes 3 Comments

TraMOOC
Translation for Massive Open Online Courses

Health in my Language HimL

SUMMA

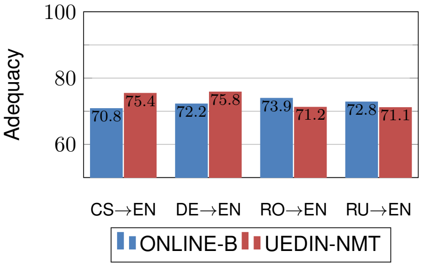
Neural Machine Translation

1. Recent advances in neural MT

2. Towards using neural MT in production
   - things that are suddenly easy(er)
   - things that are suddenly hard(er)
   - things that are still hard

Fluency

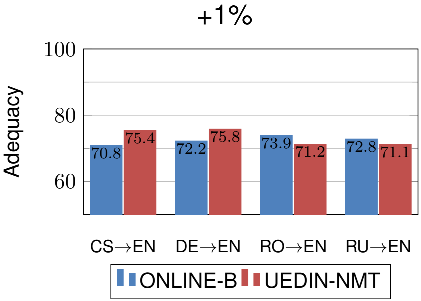
main strength of neural MT [Neubig et al., 2015, Bojar et al., 2016, Bentivogli et al., 2016]



WMT16 direct assessment results

Fluency

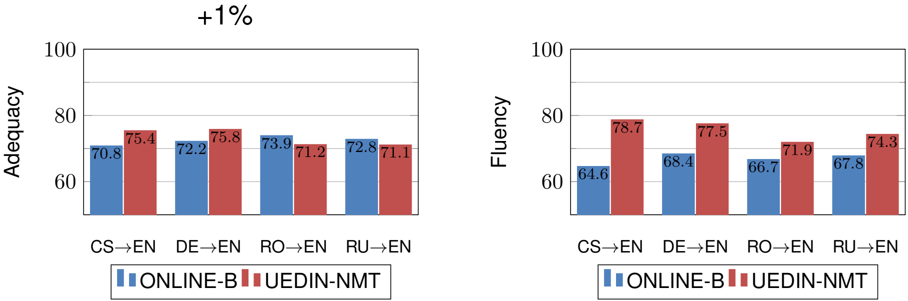
main strength of neural MT [Neubig et al., 2015, Bojar et al., 2016, Bentivogli et al., 2016]



WMT16 direct assessment results

Fluency

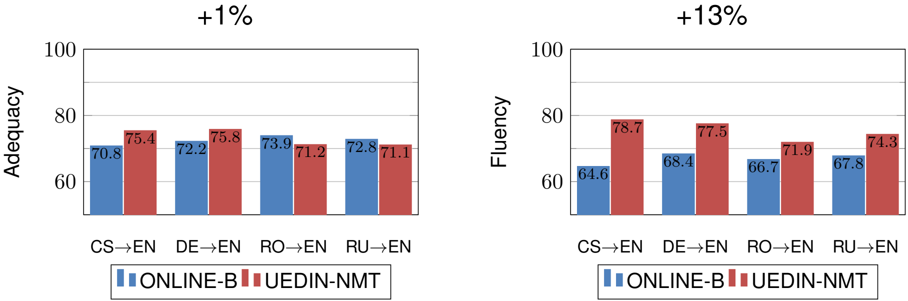
main strength of neural MT [Neubig et al., 2015, Bojar et al., 2016, Bentivogli et al., 2016]



WMT16 direct assessment results

Fluency

main strength of neural MT [Neubig et al., 2015, Bojar et al., 2016, Bentivogli et al., 2016]



WMT16 direct assessment results

Why is neural MT so much more fluent?

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

Fluency: example (WMT16; UEDIN submissions)

system	sentence
SRC	Unsere digitalen Leben haben die Notwendigkeit, stark, lebenslustig und erfolgreich zu erscheinen, verdoppelt [...]
REF	Our digital lives have doubled the need to appear strong, fun-loving and successful [...]
PBSMT	Our digital lives are lively, strong, and to be successful, doubled [...]
NMT	Our digital lives have doubled the need to appear strong, lifelike and successful [...]

Controlling neural MT

T-V distinction

language	informal (T)	formal (V)
Latin	tu	vos
Chinese	你(nǐ)	您 (nín)
French	tu	vous
German	du	Sie
Italian	tu	Lei
Polish	ty	pan
Spanish	tú	usted

Controlling neural MT

T-V distinction

language	informal (T)	formal (V)
Latin	tu	vos
Chinese	你(nǐ)	您 (nín)
French	tu	vous
German	du	Sie
Italian	tu	Lei
Polish	ty	pan
Spanish	tú	usted
Early Modern English	thou	ye
Modern English	you	

- inconsistency in T-V choice is a “limitation of MT technology” that is “often frustrat[ing]” to post-editors [Etchegoyhen et al., 2014]

Controlling neural MT

T-V distinction

language	informal (T)	formal (V)
Latin	tu	vos
Chinese	你(nǐ)	您 (nín)
French	tu	vous
German	du	Sie
Italian	tu	Lei
Polish	ty	pan
Spanish	tú	usted
Early Modern English	thou	ye
Modern English	you	

What users want

- inconsistency in T-V choice is a “limitation of MT technology” that is “often frustrat[ing]” to post-editors [Etchegoyhen et al., 2014]

Side constraints [Sennrich et al., 2016a]

Core idea

- additional input feature that is based on target-side information
  $\rightarrow$ extra word at end of source sentence
- mark in English text if German translation is polite or not (+noise)

- Are you ok?
- Sind Sie in Ordnung?

- are you ok?
- Bist du in Ordnung?

At test time

- we can control level of politeness by adding side constraints to input

Side constraints [Sennrich et al., 2016a]

Core idea

- additional input feature that is based on target-side information
  $\rightarrow$ extra word at end of source sentence
- mark in English text if German translation is polite or not (+noise)

- Are you ok?
- Sind Sie in Ordnung?

- are you ok?
- Bist du in Ordnung?

At test time

- we can control level of politeness by adding side constraints to input

Side constraints [Sennrich et al., 2016a]

Core idea

- additional input feature that is based on target-side information
  $\rightarrow$ extra word at end of source sentence
- mark in English text if German translation is polite or not (+noise)

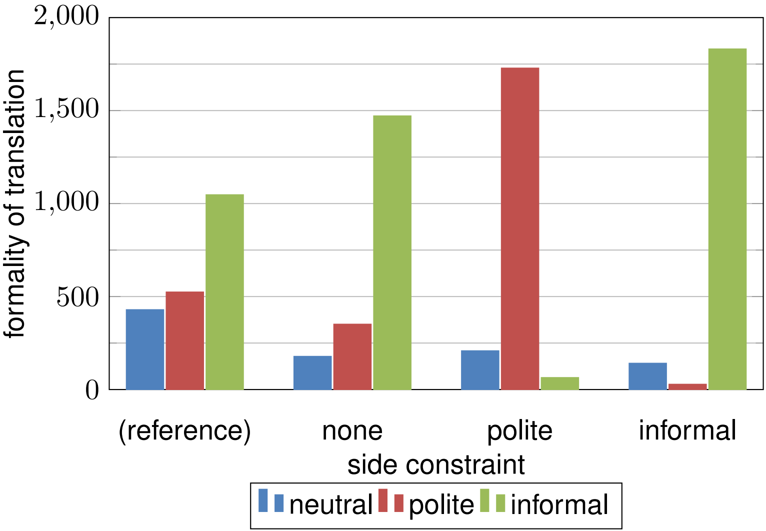
- Are you ok? <polite>
- Sind Sie in Ordnung?

- are you ok? <informal>
- Bist du in Ordnung?

At test time

- we can control level of politeness by adding side constraints to input

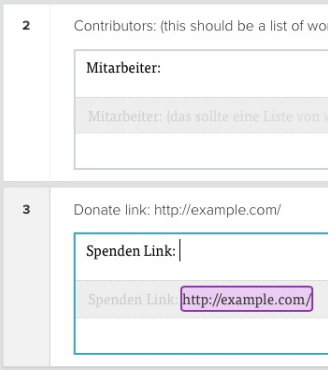
Results: politeness as a function of side constraint



Interactive MT: constrained decoding

[Wuebker et al., 2016]

- prefix-constrained decoding of high interest for interactive MT
- phrase-based MT has problems with reachability; requires new algorithms
- prefix-constrained decoding with neural MT is very natural



Incremental/online training

- Neural MT uses iterative training (SGD or Reinforcement Learning)
  $\rightarrow$ stopping/continuing training trivial
- problematic: expanding vocabulary
  $\rightarrow$ unnecessary with subword models
- multi-BLEU improvements reported with minutes of training time
  [Sennrich et al., 2016b, Luong and Manning, 2015, Crego et al., 2016]

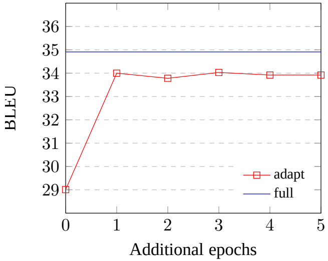


Figure 3: Adaptation with In-Domain data.

[Crego et al., 2016]

1. Recent advances in neural MT

2. Towards using neural MT in production
   - things that are suddenly easy(er)
   - things that are suddenly hard(er)
   - things that are still hard

Model interpretation/manipulation

- limited interpretability of neural network
- limited ability to manipulate neural network

- more research on terminology integration needed

Model interpretation/manipulation

- limited interpretability of neural network
- limited ability to manipulate neural network

Lifestyle › Tech

Thousands sign petition asking [redacted] to remove homophobic slurs from translation service

Company later obliged and slurs were taken down

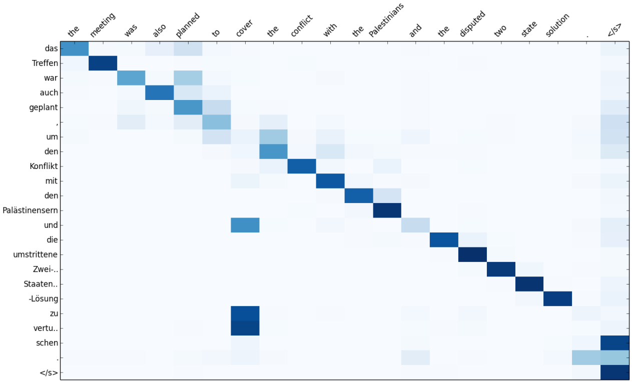
- more research on terminology integration needed

Alignment

attention model

- attends to states that are relevant for next translation decision
- ...bearing in mind that information can travel along RNN

→ no ’traditional’ word alignment



Neural Machine Translation

1. Recent advances in neural MT

2. Towards using neural MT in production
   - things that are suddenly easy(er)
   - things that are suddenly hard(er)
   - things that are still hard

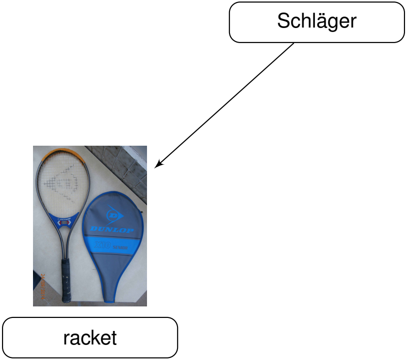
Ambiguity

system	sentence
SRC	Dort wurde er von dem Schläger und einer weiteren männl. Person erneut angegriffen.
REF	There he was attacked again by his original attacker and another male.
PBSMT	There, he was at the club and another male person attacked again.
NMT	There he was attacked again by the racket and another male person.

Schläger

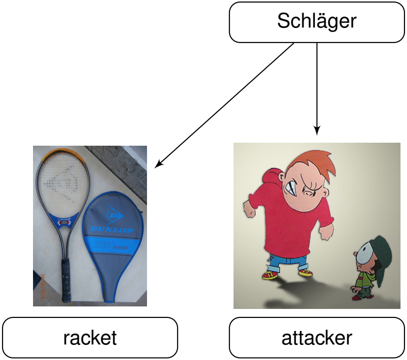
Ambiguity

system	sentence
SRC	Dort wurde er von dem Schläger und einer weiteren männl. Person erneut angegriffen.
REF	There he was attacked again by his original attacker and another male.
PBSMT	There, he was at the club and another male person attacked again.
NMT	There he was attacked again by the racket and another male person.



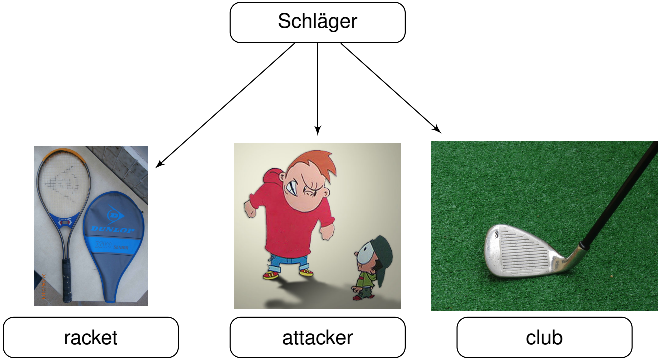
Ambiguity

system	sentence
SRC	Dort wurde er von dem Schläger und einer weiteren männl. Person erneut angegriffen.
REF	There he was attacked again by his original attacker and another male.
PBSMT	There, he was at the club and another male person attacked again.
NMT	There he was attacked again by the racket and another male person.



Ambiguity

system	sentence
SRC	Dort wurde er von dem Schläger und einer weiteren männl. Person erneut angegriffen.
REF	There he was attacked again by his original attacker and another male.
PBSMT	There, he was at the club and another male person attacked again.
NMT	There he was attacked again by the racket and another male person.



Rare words

system	sentence
SRC	Titelverteidiger ist Drittligaabsteiger SpVgg Unterhaching.
REF	The defending champions are SpVgg Unterhaching, who have been relegated to the third league.
PBSMT	Title defender Drittligaabsteiger Week 2.
NMT	Defending champion is third-round pick SpVgg Underhaching.

fully character-level models [Lee et al., 2016]

(a) Spelling mistakes

DE ori	Warum sollten wir nicht Freunde sei ?
DE src	Warum solltne wir nich Freunde sei ?
EN ref	Why should not we be friends ?
bpe2char	Why are we to be friends ?
char2char	Why should we not be friends ?

(b) Rare words

DE src	Siebentausendzweihundertvierundfünfzig .
EN ref	Seven thousand two hundred fifty four .
bpe2char	Fifty-five Decline of the Seventy .
char2char	Seven thousand hundred thousand fifties .

Discourse

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.

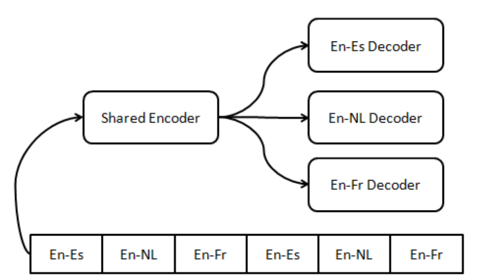
- most MT systems do not take discourse context into account...
- ... but neural MT is a promising architecture to solve this problem

Low-resourced language pairs

- most language pairs have few parallel resources
- is NMT more data efficient than phrase-based SMT?

- new potential: sharing of model parameters between language pairs

[Zoph et al., 2016, Dong et al., 2015, Firat et al., 2016, Lee et al., 2016]



Conclusions

- neural MT has achieved state of the art on many tasks...
  ... and is still improving quickly
- industry adoption is happening, but beware:
  - some things are suddenly easy(er)
  - some things are suddenly hard(er)
- machine translation still has hard problems to tackle...
- ...and neural MT offers exciting new ways to address them

Thanks

Collaborators

Alexandra Birch

Barry Haddow

Marcin Junczys-Dowmunt

Kenneth Heafield

Antonio Valerio Miceli Barone

Tomasz Dwojak

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).

Thanks

Collaborators


Alexandra Birch


Barry Haddow


Marcin Junczys-Dowmunt


Kenneth Heafield


Antonio Valerio Miceli Barone


Tomasz Dwojak

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).

Thank you for your attention

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).

Bentivogli, L., Bisazza, A., Cettolo, M., and Federico, M. (2016).
Neural versus Phrase-Based Machine Translation Quality: a Case Study.
In EMNLP 2016.

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. Association for Computational Linguistics.

Crego, J., Kim, J., Klein, G., Rebollo, A., Yang, K., Senellart, J., Akhanov, E., Brunelle, P., Coquard, A., Deng, Y., Enoue, S., Geiss, C., Johanson, J., Khalsa, A., Khiari, R., Ko, B., Kobus, C., Lorieux, J., Martins, L., Nguyen, D.-C., Priori, A., Riccardi, T., Segal, N., Servan, C., Tiquet, C., Wang, B., Yang, J., Zhang, D., Zhou, J., and Zoldan, P. (2016).
SYSTRAN's Pure Neural Machine Translation Systems.
ArXiv e-prints.

Dong, D., Wu, H., He, W., Yu, D., and Wang, H. (2015).
Multi-Task Learning for Multiple Language Translation.
In
Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference
pages 1723–1732, Beijing, China. Association for Computational Linguistics.

Bibliography II

Etchegoyhen, T., Bywood, L., Fishel, M., Georgakopoulou, P., Jiang, J., Loenhout, G. V., Pozo, A. D., Maucec, M. S., Turner, A., and Volk, M. (2014).
Machine Translation for Subtitling: A Large-Scale Evaluation.
In Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), Reykjavik, Iceland. European Language Resources Association (ELRA).

Firat, O., Cho, K., and Bengio, Y. (2016).
Multi-Way, Multilingual Neural Machine Translation with a Shared Attention Mechanism.
In
Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Langua
pages 866–875. 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.

Junczys-Dowmunt, M., Dwojak, T., and Sennrich, R. (2016).
The AMU-UEDIN Submission to the WMT16 News Translation Task: Attention-based NMT Models as Feature Functions in Phrase-based SMT.
In Proceedings of the First Conference on Machine Translation, Volume 2: Shared Task Papers, pages 316–322, Berlin, Germany. Association for Computational Linguistics.

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

Luong, M.-T. and Manning, C. D. (2015).
Stanford Neural Machine Translation Systems for Spoken Language Domains.
In Proceedings of the International Workshop on Spoken Language Translation 2015, Da Nang, Vietnam.

Bibliography III

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).
Controlling Politeness in Neural Machine Translation via Side Constraints.
In
Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Langua
pages 35–40, San Diego, California. Association for Computational Linguistics.

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. Association for Computational Linguistics.

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. Association for Computational Linguistics.

Wuebker, J., Green, S., DeNero, J., Hasan, S., and Luong, M.-T. (2016).
Models and Inference for Prefix-Constrained Machine Translation.
In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 66–75. Association for Computational Linguistics.

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