Lessons from Multilingual Machine Translation

Rico Sennrich

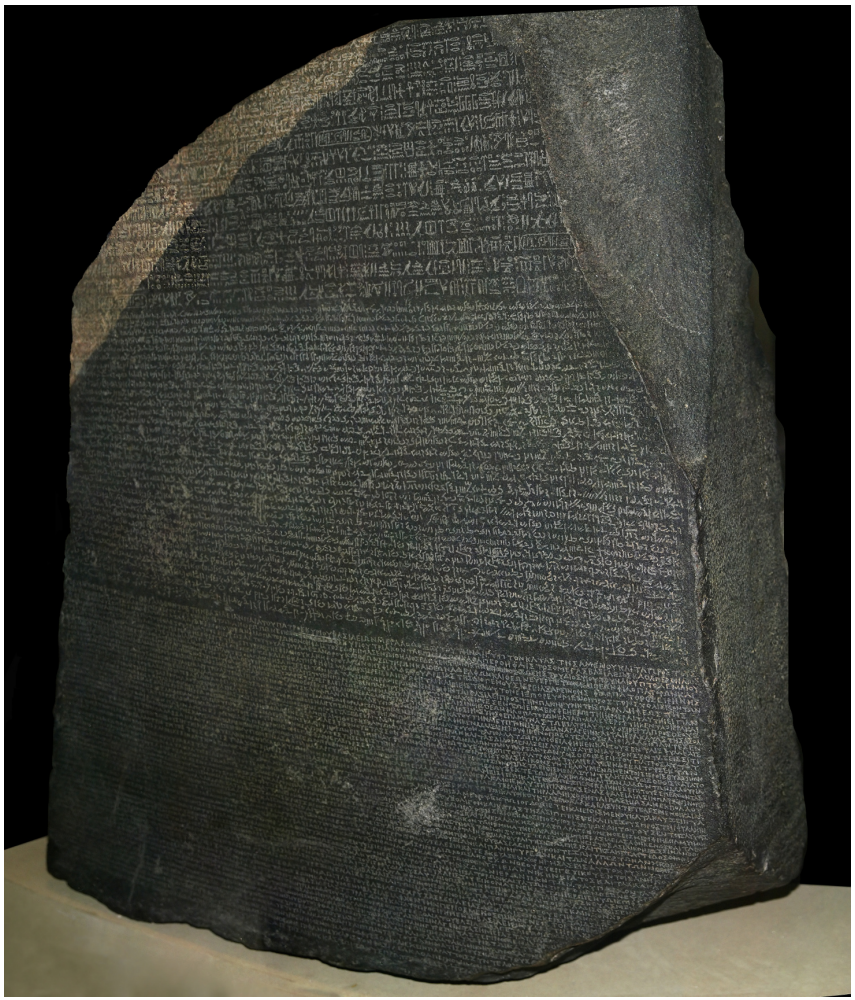

University of Zurich[UZH]

Lessons from Multilingual Machine Translation

1. Preliminaries: Neural Machine Translation
2. Cross-lingual Transfer for Machine Translation
3. Massively Multilingual Models and Zero-Shot Translation

Data-driven (Machine) Translation

the past



the future (?)

A Minimal Introduction to Machine Translation

- Suppose that we have:
  - a source sentence S of length m $(x_1, \ldots, x_m)$
  - a target sentence T of length n $(y_1, \ldots, y_n)$
- We can express translation as a probabilistic model

$$T^* = \arg\max_T p(T|S)$$

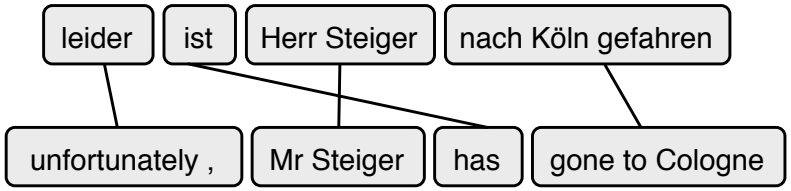
- Expanding using the chain rule gives

$$\begin{aligned} p(T|S) &= p(y_1, \ldots, y_n | x_1, \ldots, x_m) \\ &= \prod_{i=1}^{n} p(y_i | y_1, \ldots, y_{i-1}, x_1, \ldots, x_m) \end{aligned}$$

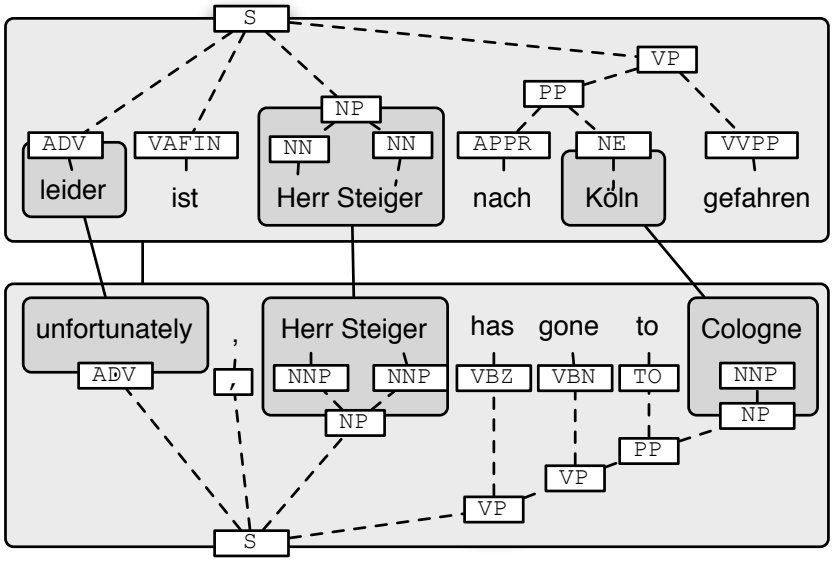
- how do we train model? Lots of human translations
- how do we measure quality? Similarity to human translation (BLEU)

(Non-neural) Statistical Machine Translation

phrase-based SMT



syntax-based SMT



data sparsity

→ independence assumptions

→ combination of weak models

→ poor generalisation

Neural Machine Translation

end-to-end training
of complex function

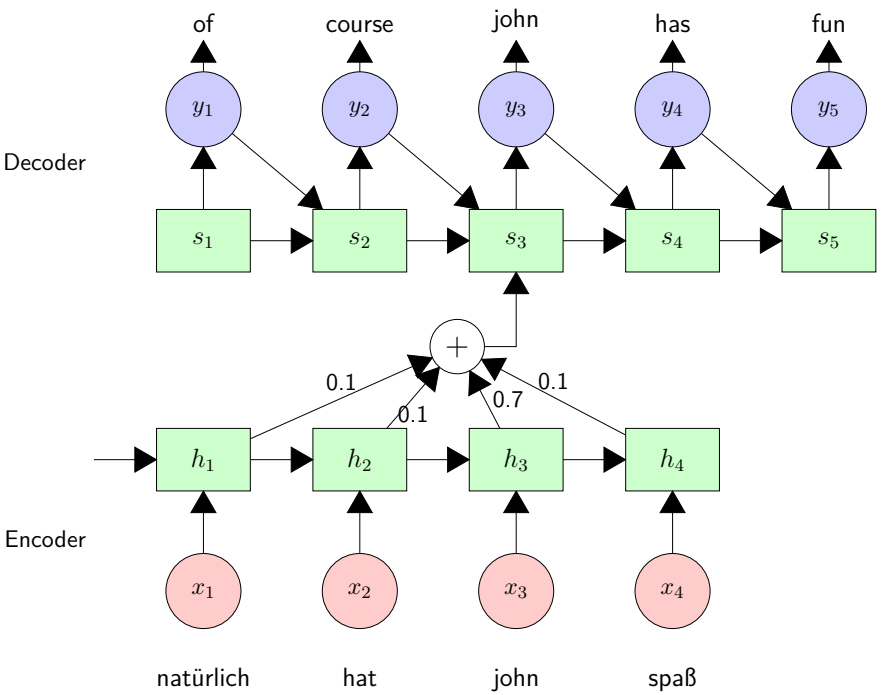
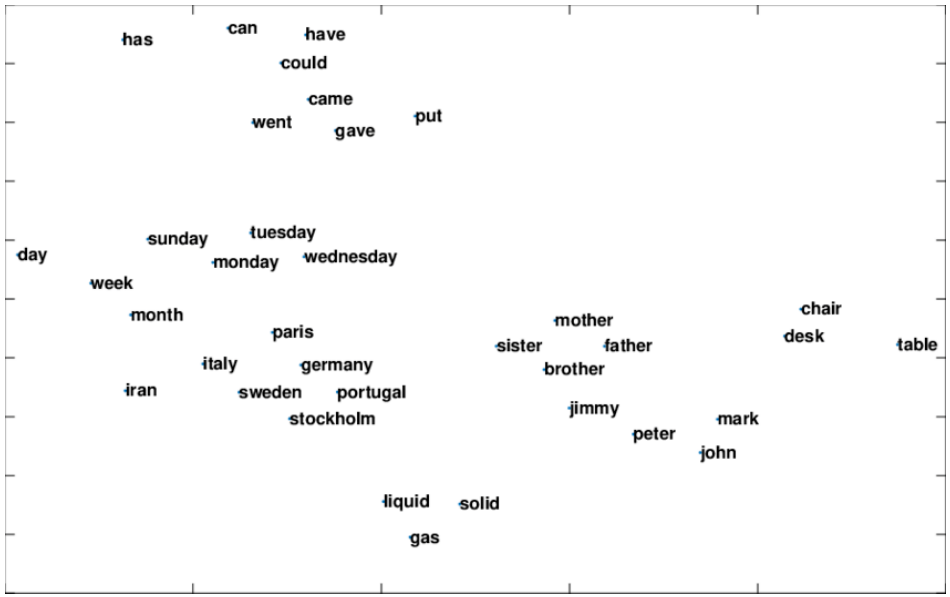


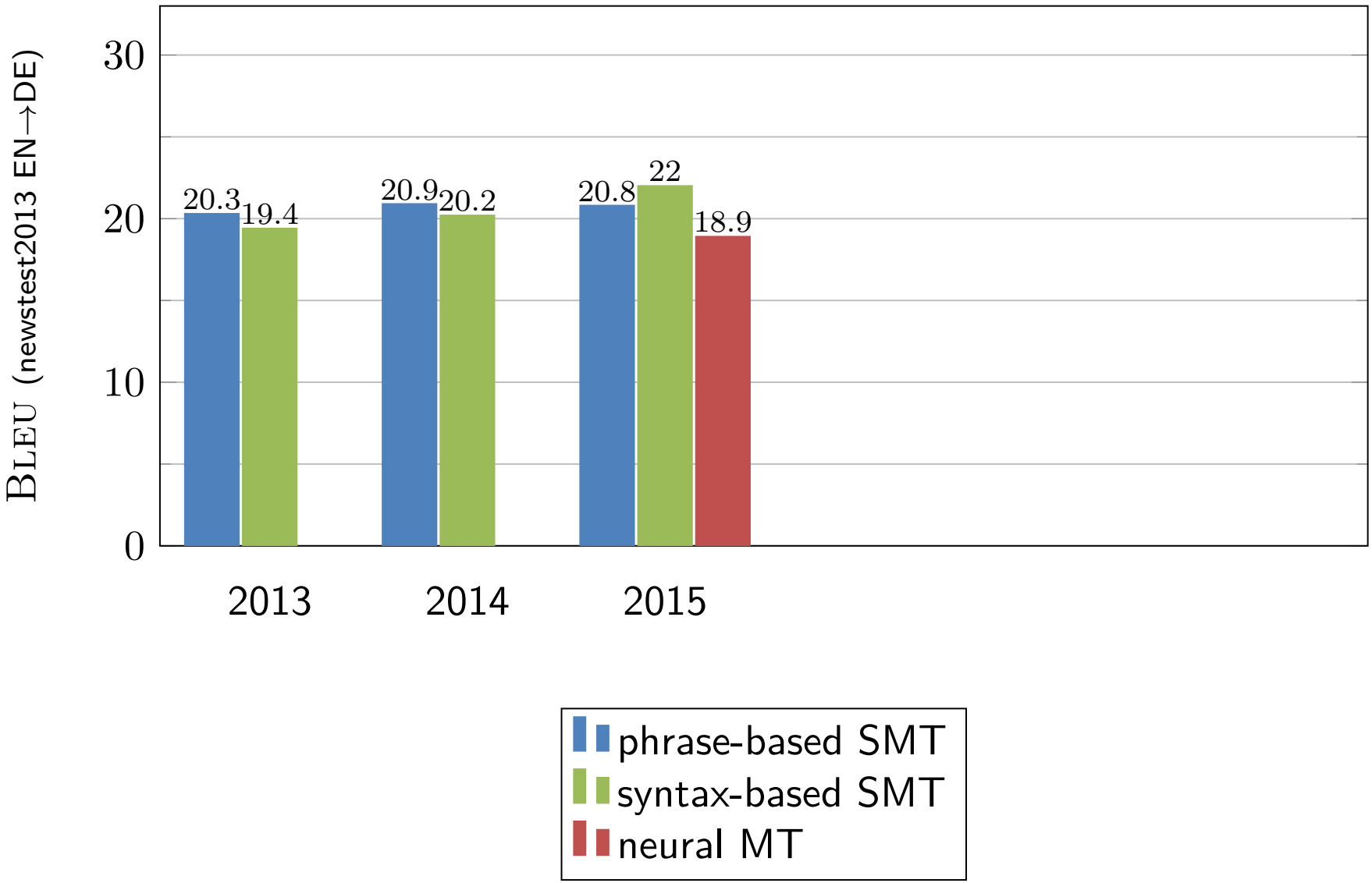
illustration by Barry Haddow

generalisation through
continuous representation



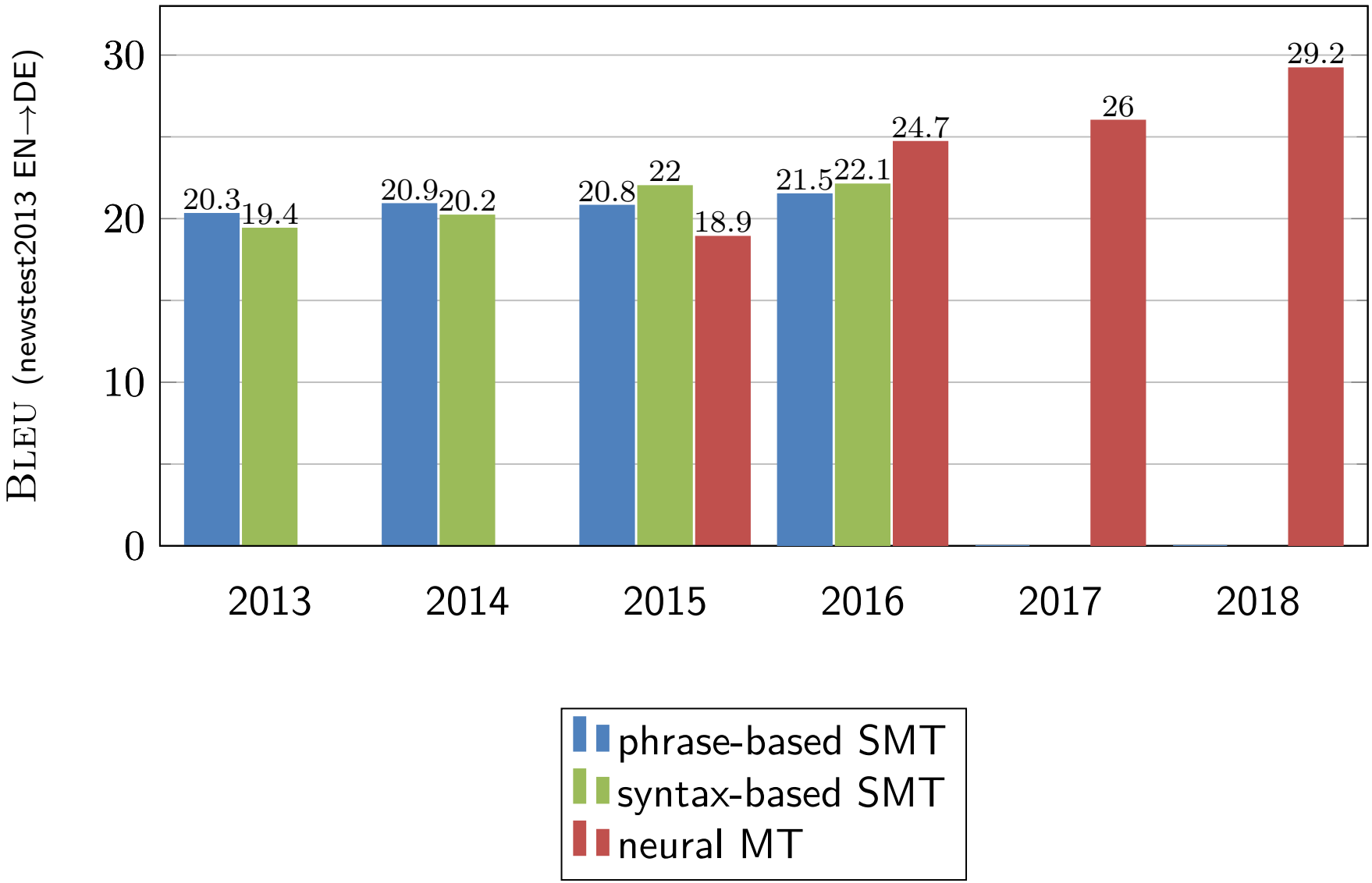
Ali Basirat, Principal Word Vectors

Edinburgh's* WMT Results over the Years



*NMT 2015 from U. Montréal: https://sites.google.com/site/acl16nmt/
*NMT 2018 from [Edunov et al., 2018]

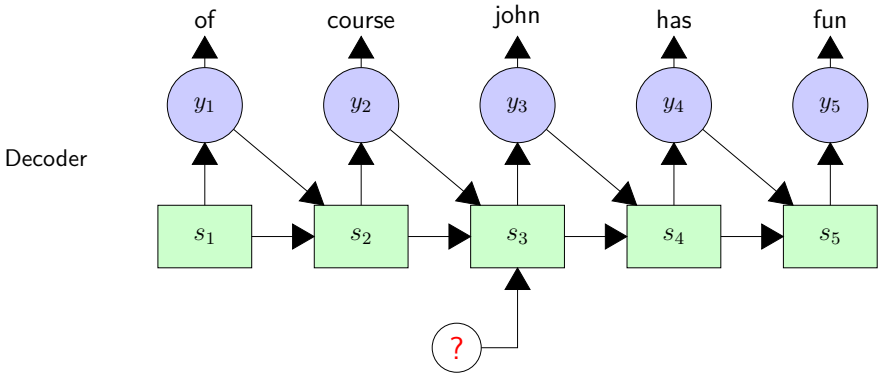
Edinburgh's* WMT Results over the Years



*NMT 2015 from U. Montréal: `https://sites.google.com/site/acl16nmt/`
*NMT 2018 from [Edunov et al., 2018]

Main Improvements*: 2016

- semi-supervised training (backtranslation) [Sennrich et al., 2016a]



- open-vocabulary modelling [Sennrich et al., 2016b]

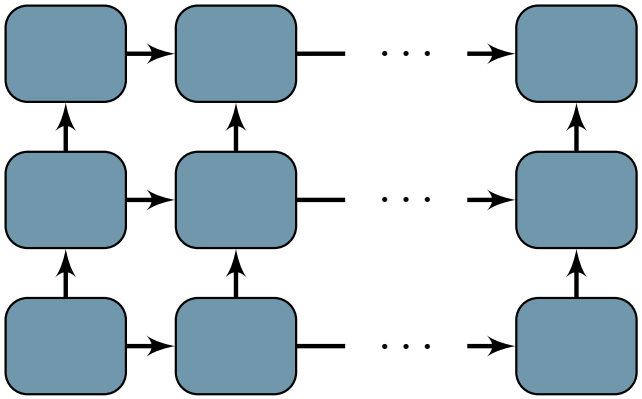
source	**indoor temperature**	
small-vocabulary	**UNK**	✗
+back-off	**Innenpool**	✗
byte-pair encoding	**Innen+ temperatur**	✓

- tweaks to neural RNN architectures and training

*for systems reported

Main Improvements*: 2017

- deeper models



- improvements to RNN architectures:
  layer normalisation; residual connections
- tweaks to training:
  Adam; diverse ensembles
- slightly more training data

*for systems reported

Main Improvements*: 2018

- Transformer architecture [Vaswani et al., 2017]
- larger and deeper models
- refined and larger-scale semi-supervised training [Edunov et al., 2018]
- tweaks to training:
  label smoothing; large batches

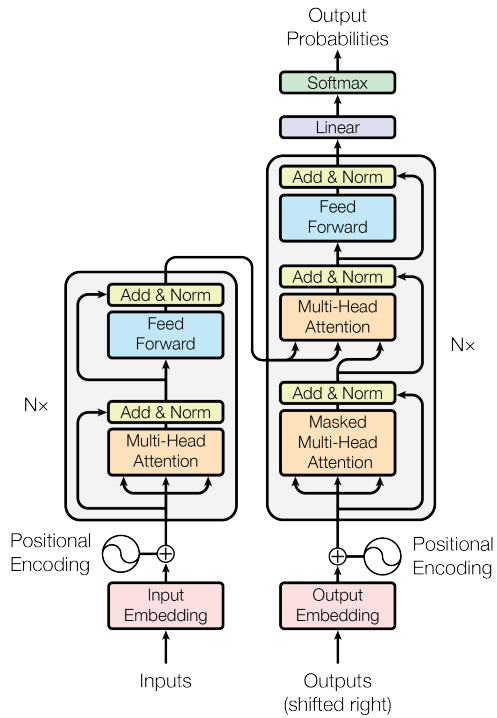


Figure 1: The Transformer - model architecture.

*for systems reported

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
→ 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
→ 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., 2015]	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., 2015]	Die **Innenpool** ist sehr angenehm.	✗
[Sennrich, Haddow, Birch, ACL 2016a]	Die **Innen+ temperatur** ist sehr angenehm.	✓

Subword Neural MT

[Sennrich, Haddow, Birch, ACL 2016]



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

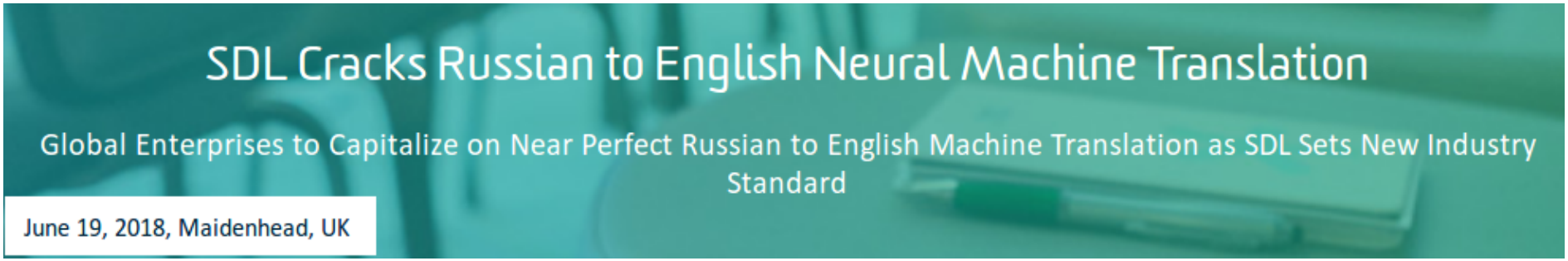
vocabulary size	text
300	t: h: e i: n: d: o: o: r t: e: m: p: e: r: a: t: u: r: e i: s v: e: r: y p: l: e: a: s: a: n: t
1300	the in: do: or t: em: per: at: ure is very p: le: as: ant
10300	the in: door temper: ature is very pleasant
50300	the indoor temperature is very pleasant

Where Is Machine Translation Now?

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

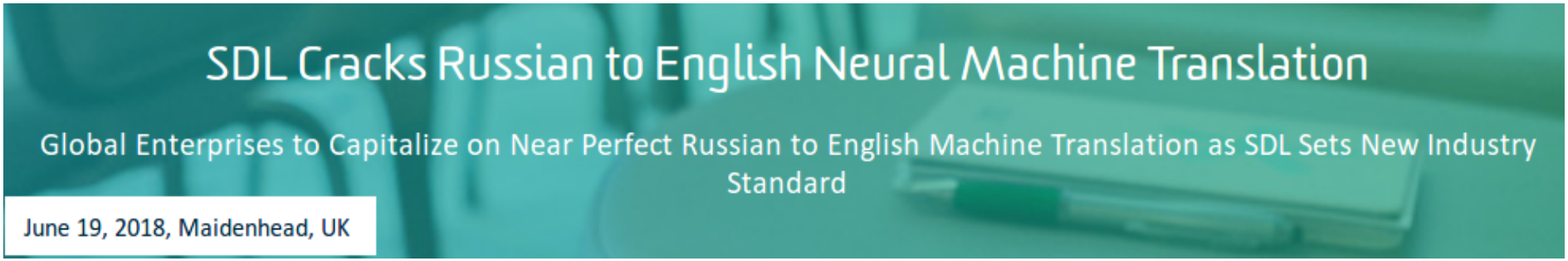


Where Is Machine Translation Now?

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

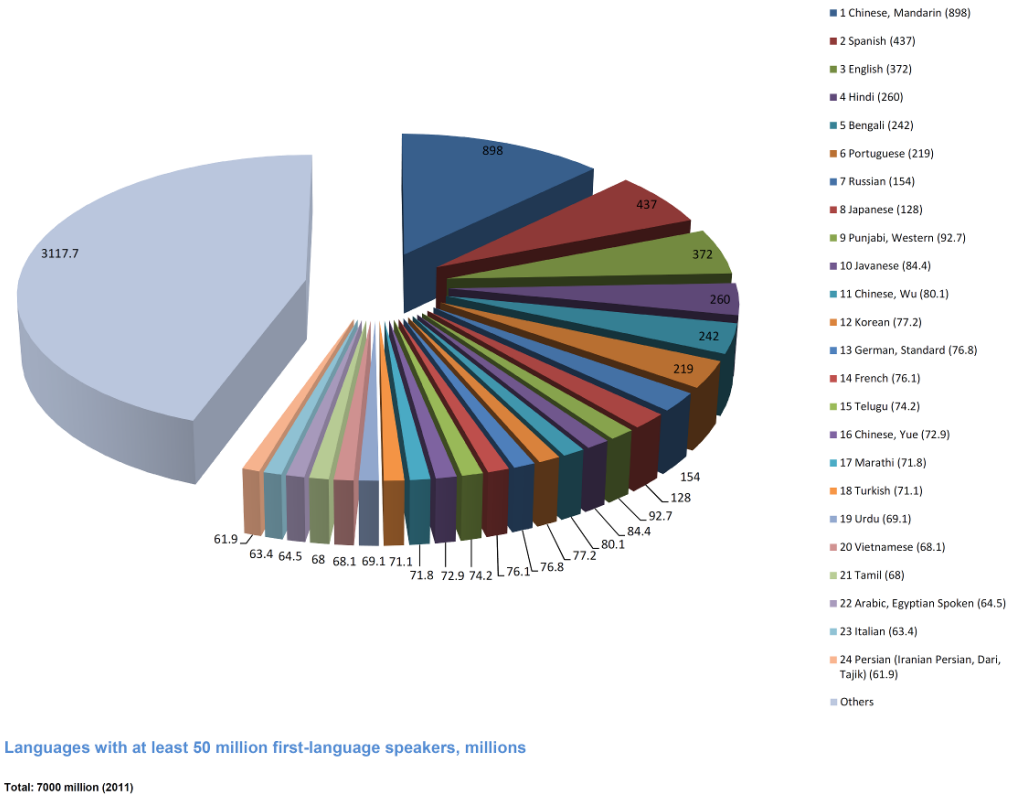


...but strong evaluation protocols still show gap [Läubli et al., 2018, Toral et al., 2018]

Talk Today

Neural machine translation is impressive for high-resource language pairs.

How do we build systems for low-resource pairs?

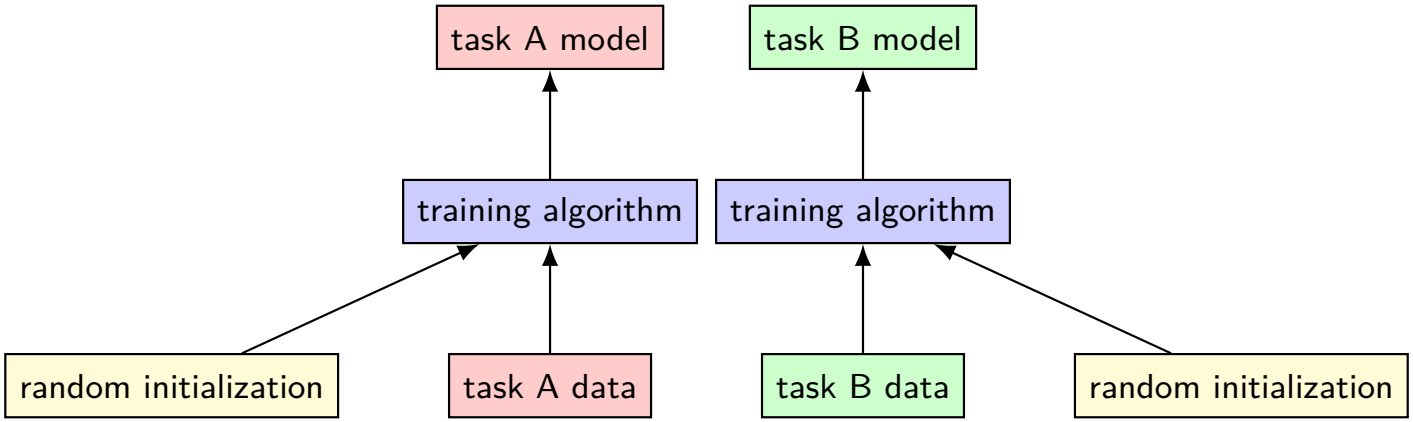


Lessons from Multilingual Machine Translation

1. Preliminaries: Neural Machine Translation
2. Cross-lingual Transfer for Machine Translation
3. Massively Multilingual Models and Zero-Shot Translation

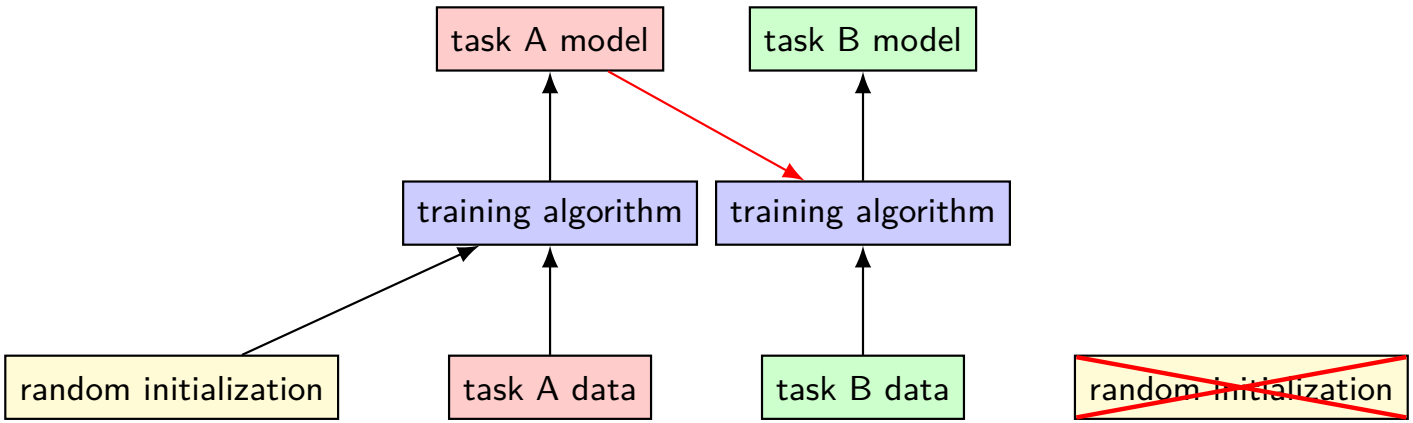
Transfer Learning

basic setup: models share architecture, but are trained separately



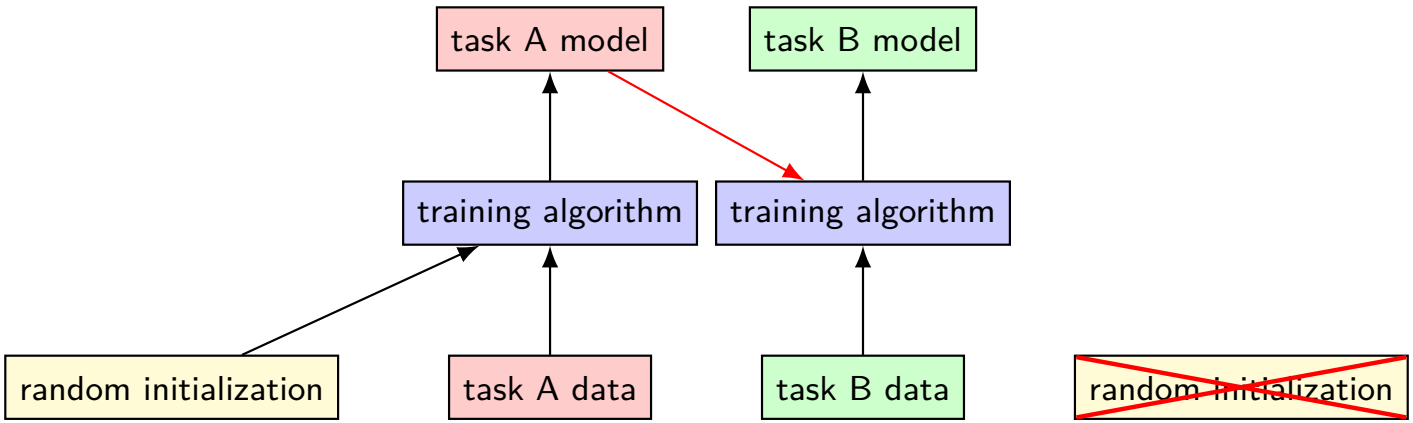
Transfer Learning

effective: use pre-trained model for initialization



Transfer Learning

effective: use pre-trained model for initialization



in MT, transfer from high-resource language pair to low-resource pair first proposed by [Zoph et al., 2016]

TL: What Does It Transfer?

[Aji, Bogoychev, Heafield, Sennrich, ACL 2020]

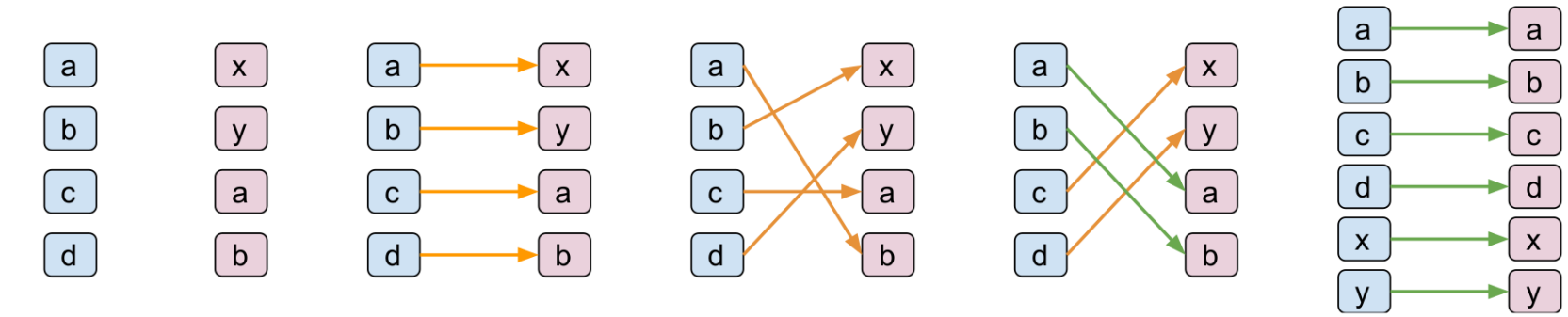



hypothesis: model with copying behavior is good initialization for NMT

experiments with three low-resource language pairs ({MY,ID,TR}→EN):

- real NMT model (DE→EN) as parent works best (+6 BLEU)
- but model that just copies EN→EN also does ok (+2 BLEU)
- copying an unrelated language (ZH→ZH) works similarly (+2 BLEU)
- some transfer from model that copies random sequences (+1 BLEU)

Transferring Embeddings



(a) Exclude embedding (b) Freq. assignment (c) Random assignment (d) Token Match (e) Joint vocab

Figure 1: Illustration of various strategies on how to transfer the embedding vector.

transfer strategy	average BLEU (EN↔{My,Id,Tr})
none	14.5
exclude embedding	18.3
frequency assign	19.2
random assign	19.0
token matching	21.7
joint vocabulary	22.0

joint vocabulary most effective, but least flexible
(we define joint subword vocabulary before training parent model)

Building Systems Fast

2010 Haiti Earthquake

Translator Fast-Tracks Haitian Creole

Published February 4, 2010

(from idea to deployed system in 5 days)

towards fast transfer of neural MT to new language pairs

- joint vocabulary has cumbersome workflow
- token matching does ok
- alternative: [Gheini and May, 2019]
  - re-use parent vocabulary
  - romanise/transliterate child languages to match parent vocabulary

On Romanisation for NMT

[Amrhein and Sennrich, Findings of EMNLP 2020]

questions:

- can we recover original script if we romanise target side?
  $\rightarrow$ yes, deromanisation is easy with seq2seq models
- do we benefit from better transfer than with token matching strategy?
  $\rightarrow$ this is complicated...

experimental data:

- parent model for EN$\leftrightarrow${AR,DE,FR,RU,ZH}
  (1M sentence pairs each)
- child models for EN$\leftrightarrow${AM,HE,MR,MT,SH,TA,YI}
  (7000–200,000 sentence pairs)

On Romanisation for NMT

romanisation loses information

她到塔皓湖去了

uroman: ta dao ta hao hu qu le

uconv: tā dào tǎ hào hú qù le

"She went to Lake Tahoe."

information loss varies:

- 塔 (Pinyin *tǎ*): 'tower'
- 她 (Pinyin *tā*): 'she'
- 他 (Pinyin *tā*): 'he'

'universal' subword segmentation isn't optimal

Amharic	መልክተኞቹም ወቅት በተደረገላቸው ጊዜ ፡ ፡
	melikitenyochumi wek' iti betederegelachewi gīzē ፡ ፡
English	The Messengers will receive their appointments.
BPE	_መልክተኞ ቹም _ወ ቅ ት _በተ ደረገ ላቸው _ጊዜ _፡ _፡
	_melikitenyo chumi _we k' i ti _bete derege lachewi _gīzē _፡ _፡
romanised, re-using parent vocabulary	
uroman	_m ale ket any och ume _waq ete _bat ad ar ag ala a cha we _g iz ee
uconv	_mel ik iten y och u mi _we k ' iti _bet ed ere gel ache wi _g ī z ē _፡ _፡

On Romanisation for NMT: Results

	BLEU change to bilingual model		
		romanised	
	token match	uroman	uconv
new script, unrelated	+ 6.1	+ 4.5	+ 5.9
new script, related	+ 6.8	+ 9.8	+11.0
Latin, related to non-Latin parent	+13.9	+14.3	+14.8

On Romanisation for NMT: Results

	BLEU change to bilingual model		
		romanised	
	token match	uroman	uconv
new script, unrelated	+ 6.1	+ 4.5	+ 5.9
new script, related	+ 6.8	+ 9.8	+11.0
Latin, related to non-Latin parent	+13.9	+14.3	+14.8

- transfer learning with token matching strong baseline

On Romanisation for NMT: Results

	BLEU change to bilingual model		
		romanised	
	token match	uroman	uconv
new script, unrelated	+ 6.1	+ 4.5	+ 5.9
new script, related	+ 6.8	+ 9.8	+11.0
Latin, related to non-Latin parent	+13.9	+14.3	+14.8

- transfer learning with token matching strong baseline
- information loss matters: uconv outperforms uroman

On Romanisation for NMT: Results

	BLEU change to bilingual model		
		romanised	
	token match	uroman	uconv
new script, unrelated	+ 6.1	+ 4.5	+ 5.9
new script, related	+ 6.8	+ 9.8	+11.0
Latin, related to non-Latin parent	+13.9	+14.3	+14.8

- transfer learning with token matching strong baseline
- information loss matters: uconv outperforms uroman
- Romanisation is not generally superior...

On Romanisation for NMT: Results

	BLEU change to bilingual model		
		romanised	
	token match	uroman	uconv
new script, unrelated	+ 6.1	+ 4.5	+ 5.9
new script, related	+ 6.8	+ 9.8	+11.0
Latin, related to non-Latin parent	+13.9	+14.3	+14.8

- transfer learning with token matching strong baseline
- information loss matters: uconv outperforms uroman
- Romanisation is not generally superior...
- ...but better transfer between related languages with different scripts:
  - transfer to non-Latin script
    (Hebrew with Arabic parent; Yiddish with German parent)

On Romanisation for NMT: Results

	BLEU change to bilingual model		
		romanised	
	token match	uroman	uconv
new script, unrelated	+ 6.1	+ 4.5	+ 5.9
new script, related	+ 6.8	+ 9.8	+11.0
Latin, related to non-Latin parent	+13.9	+14.3	+14.8

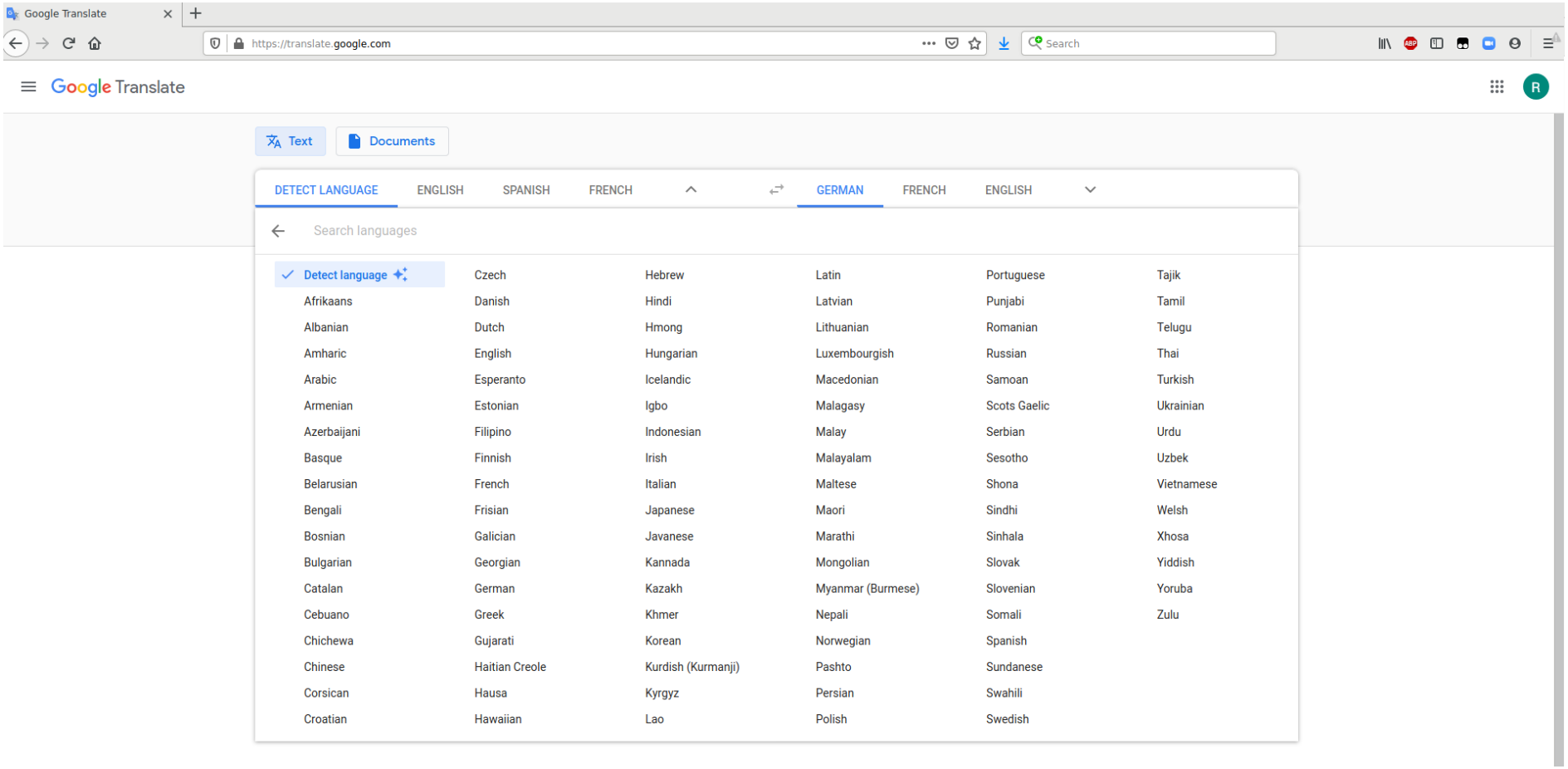
- transfer learning with token matching strong baseline
- information loss matters: uconv outperforms uroman
- Romanisation is not generally superior...
- ...but better transfer between related languages with different scripts:
  - transfer to non-Latin script
    (Hebrew with Arabic parent; Yiddish with German parent)
  - transfer from non-Latin script
    (Maltese with Arabic parent; Serbo-Croatian with Russian parent)

Lessons from Multilingual Machine Translation

1. Preliminaries: Neural Machine Translation
2. Cross-lingual Transfer for Machine Translation
3. Massively Multilingual Models and Zero-Shot Translation

Massively Multilingual Machine Translation

imagine you want to support 100 languages
that's $\approx$ 10 000 translation directions – how do you do MT for all of them?

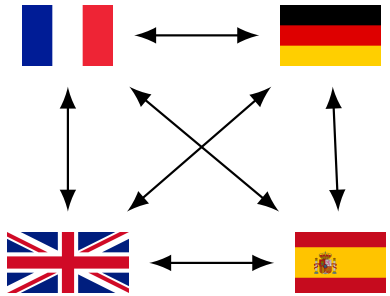


Massively Multilingual Machine Translation

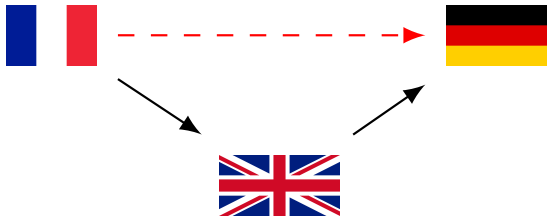
imagine you want to support 100 languages
that's $\approx 10\,000$ translation directions – how do you do MT for all of them?

traditional answer:

- build lots of systems ☹



- pivot via English if you have no dedicated system for that direction ☹

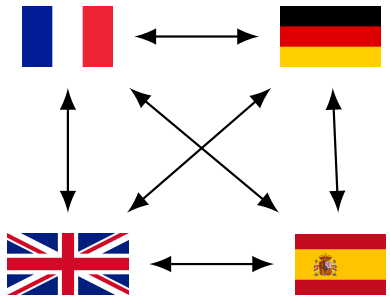


Massively Multilingual Machine Translation

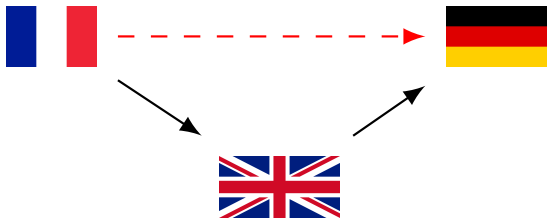
imagine you want to support 100 languages
that's $\approx 10\,000$ translation directions – how do you do MT for all of them?

traditional answer:

- build lots of systems ☹



- pivot via English if you have no dedicated system for that direction ☹



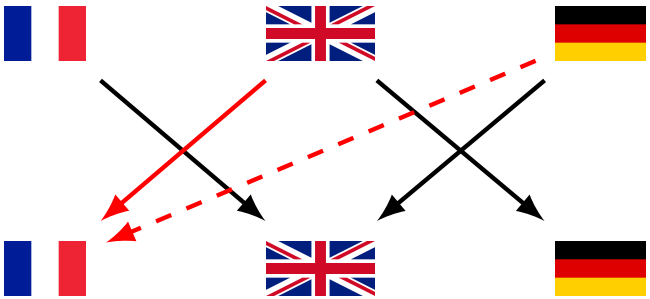
crazy idea: train one massively multilingual model that supports all

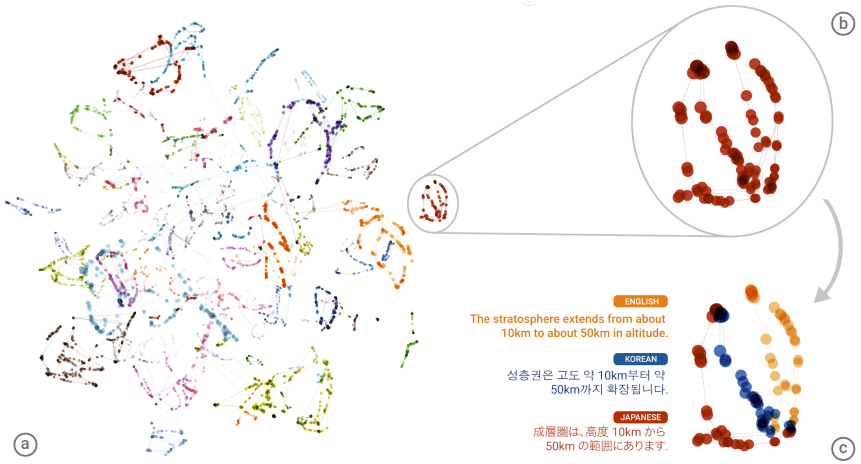
Multilingual Machine Translation Made Easy

- use standard model, just share parameters between languages
- only modification: language tag for each training example: indicate desired output language, for example as extra input token

  <2FR> Wir hoffen, dass dies nicht der Fall sein wird.

- at test time, can translate in unseen directions: **zero-shot translation**
  $\rightarrow$ how? by exploiting representational similarity across languages





[Johnson et al., 2017]

A Focus on Zero-Shot Generalisation

my focus today:

- zero-shot translation for massively multilingual models:
  - is model capacity a problem?
  - does semi-supervised learning scale?
- what biases affect zero-shot translation?

Massively Multilingual MT: Prior Work

multilingual systems and zero-shot translation work in principle

what happens if we actually scale this up to 100 languages?

experiments by [Aharoni et al., 2019]: EN-X parallel data for 102 languages

Massively Multilingual MT: Prior Work

average performance decreases with number of languages ☹

	Ar-En	En-Ar	Fr-En	En-Fr	Ru-En	En-Ru	Uk-En	En-Uk	Avg.
5-to-5	**23.87**	**12.42**	**38.99**	**37.3**	29.07	**24.86**	**26.17**	16.48	**26.14**
25-to-25	23.43	11.77	38.87	36.79	**29.36**	23.24	25.81	**17.17**	25.8
50-to-50	23.7	11.65	37.81	35.83	29.22	21.95	26.02	15.32	25.18
75-to-75	22.23	10.69	37.97	34.35	28.55	20.7	25.89	14.59	24.37
103-to-103	21.16	10.25	35.91	34.42	27.25	19.9	24.53	13.89	23.41

Table 7: Supervised performance while varying the number of languages involved

[Aharoni et al., 2019]

Massively Multilingual MT: Prior Work

zero-shot performance generally poor and unpredictable ☹

	Ar-Fr	Fr-Ar	Ru-Uk	Uk-Ru	Avg.
5-to-5	1.66	4.49	3.7	3.02	3.21
25-to-25	1.83	**5.52**	**16.67**	4.31	7.08
50-to-50	**4.34**	4.72	15.14	**20.23**	**11.1**
75-to-75	1.85	4.26	11.2	15.88	8.3
103-to-103	2.87	3.05	12.3	18.49	9.17

Table 8: Zero-Shot performance while varying the number of languages involved

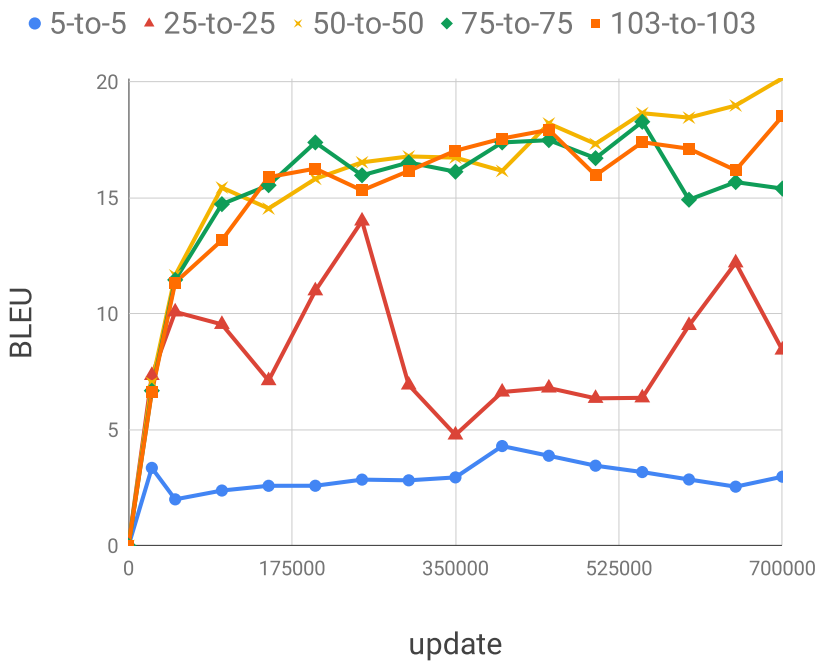


Figure 2: Zero-shot BLEU during training for Ukranian to Russian

[Aharoni et al., 2019]

Large-Capacity Massively Multilingual NMT

[Zhang, Williams, Titov, Sennrich, ACL 2020]




- introducing OPUS-100: an open massively multilingual dataset
- exploring effects of deep models and language-specific components on massively multilingual models
- scaling of semi-supervised training to 10 000 zero-resource translation directions

OPUS-100: A Massively Multilingual Dataset

- based on OPUS collection of parallel corpora: `http://opus.nlpl.eu/opus-100.php`
- English-centric: EN$\leftrightarrow$X parallel corpora for 99 languages
- up to 1M sentence pairs per corpus sampled; 55M in total

Large-Capacity Modelling

- deep models with depth-scaled initialization [Zhang, Titov and Sennrich, EMNLP 2019]
- models with parameters specific to target language:
  - biases g and b for layer normalization:

$$\bar{a}_i = \frac{a_i - \mu}{\sigma} g_i + b_i$$

  - linear transformation on top of encoder

Experiments

- BLEU_4: average of DE, ZH, BR, TE
  (for comparison to bilingual baselines)
- BLEU_{94}: average of 94 translation directions with test data
- BLEU_{zero}: average of 15 zero-shot directions:
  AR, ZH, NL, FR, DE, RU (all pairs)

Results: EN→X

system	$BLEU_4$	$BLEU_{94}$
Transformer base (bilingual)	20.3	-
Transformer base (multilingual)	15.4	19.7
+ deep (12 layers)	16.1	20.9
+ language-specific parameters	19.3	22.9
+ deeper (24 layers)	19.8	24.0

Results: EN→X

system	BLEU_4	BLEU_{94}
Transformer base (bilingual)	20.3	-
Transformer base (multilingual)	15.4	19.7
+ deep (12 layers)	16.1	20.9
+ language-specific parameters	19.3	22.9
+ deeper (24 layers)	19.8	24.0

- deep models help ☺

Results: EN→X

system	$BLEU_4$	$BLEU_{94}$
Transformer base (bilingual)	20.3	-
Transformer base (multilingual)	15.4	19.7
+ deep (12 layers)	16.1	20.9
+ language-specific parameters	19.3	22.9
+ deeper (24 layers)	19.8	24.0

- deep models help ☺
- language-specific parameters help a lot ☺☺

Results: EN→X

system	$BLEU_4$	$BLEU_{94}$
Transformer base (bilingual)	20.3	-
Transformer base (multilingual)	15.4	19.7
+ deep (12 layers)	16.1	20.9
+ language-specific parameters	19.3	22.9
+ deeper (24 layers)	19.8	24.0

- deep models help ☺
- language-specific parameters help a lot ☺☺
- hard to beat bilingual models ☹

Results: X→EN

system	$BLEU_4$	$BLEU_{94}$
Transformer base (bilingual)	21.2	-
Transformer base (multilingual)	23.4	27.6
+ deep (12 layers)	24.2	29.2
+ language-specific parameters	24.5	29.5
+ deeper (24 layers)	26.0	31.4

Results: X→EN

system	$BLEU_4$	$BLEU_{94}$
Transformer base (bilingual)	21.2	-
Transformer base (multilingual)	23.4	27.6
+ deep (12 layers)	24.2	29.2
+ language-specific parameters	24.5	29.5
+ deeper (24 layers)	26.0	31.4

- deep models help ☺

Results: X→EN

system	$BLEU_4$	$BLEU_{94}$
Transformer base (bilingual)	21.2	-
Transformer base (multilingual)	23.4	27.6
+ deep (12 layers)	24.2	29.2
+ language-specific parameters	24.5	29.5
+ deeper (24 layers)	26.0	31.4

- deep models help ☺
- common output: English
  - gains over bilingual models ☺

Results: X→EN

system	$BLEU_4$	$BLEU_{94}$
Transformer base (bilingual)	21.2	-
Transformer base (multilingual)	23.4	27.6
+ deep (12 layers)	24.2	29.2
+ language-specific parameters	24.5	29.5
+ deeper (24 layers)	26.0	31.4

- deep models help ☺
- common output: English
  - gains over bilingual models ☺
  - language-specific parameters help little ☹

Results: Zero-shot

system	BLEU_{zero}
Transformer base (bilingual; pivoting)	13.0
Transformer base (multilingual)	4.0
+ deep (12 layers)	4.7
+ language-specific parameters	5.4
+ deeper (24 layers)	5.2

Results: Zero-shot

system	BLEU_{zero}
Transformer base (bilingual; pivoting)	13.0
Transformer base (multilingual)	4.0
+ deep (12 layers)	4.7
+ language-specific parameters	5.4
+ deeper (24 layers)	5.2

- little gains on zero-shot languages ☹
- performance lags far behind pivoting ☹

Semi-Supervised Massive Multitask Training

- successful strategy: **backtranslate** monolingual target text to create synthetic training data
- can we use back-translation to train $\approx 10\,000$ zero-shot directions?
  $\rightarrow$ **Random Online Back-Translation**
  - sample training batch
  - for each training example (x_k, y_k), pick random source language L_k
  - backtranslate y_k into L_k to obtain x'_k
  - use (x'_k, y_k) as new training example

Results: Zero-shot, Semisupervised

system	BLEU_{zero}	BLEU_{zero} with ROBT
Transformer base (bilingual; pivoting)	13.0	-
Transformer base (multilingual)	4.0	10.1
+ deep (12 layers)	4.7	11.9
+ language-specific parameters	5.4	12.6
+ deeper (24 layers)	5.2	14.1

Results: Zero-shot, Semisupervised

system	BLEU_{zero}	BLEU_{zero} with ROBT
Transformer base (bilingual; pivoting)	13.0	-
Transformer base (multilingual)	4.0	10.1
+ deep (12 layers)	4.7	11.9
+ language-specific parameters	5.4	12.6
+ deeper (24 layers)	5.2	14.1

- ROBT scales to $\approx 10\,000$ translation directions ☺
- large-capacity models help a lot ☺

- ROBT slightly hurts quality for EN$\leftrightarrow$X ☹
- pivoting not clearly beaten ☹
  (our best results: pivot with deep multilingual models)

Off-target translations

why does model tend to copy or produce English in zero-shot condition?

Source	Jusqu'à ce qu'on trouve le moment clé, celui où tu as su que tu l'aimais.
Reference	Bis wir den unverkennbaren Moment gefunden haben, den Moment, wo du wusstest, du liebst ihn.
Zero-Shot	Jusqu'à ce qu'on trouve le moment clé, celui où tu as su que tu l'aimais.
Source	Les États membres ont été consultés et ont approuvé cette proposition.
Reference	Die Mitgliedstaaten wurden konsultiert und sprachen sich für diesen Vorschlag aus.
Zero-Shot	Les Member States have been consultedés and have approved this proposal.

Off-target translations

bias towards copying:
untranslated segments, but also matching strings (named entities)

bias towards English:
very strong spurious correlation in training: French input only seen with English output

How Subwords and Bridge Languages Affect Zero-Shot NMT

[Rios, Müller, Sennrich, WMT 2020]



subword segmentation affects copy bias

shared subword segmentation (most widespread):

...Herrn Don: ald Tus: k , den Premi: er: minister von P: olen ...

...Don: ald Tus: k , the Prime Minister of Poland ...

language-specific segmentation with shared vocabulary:

...Herrn Don: ald T: us: k , den Premi: er: minister von Polen ...

...D: on: al: d T: us: k , the Pri: me Minister of Pol: and

How Subwords and Bridge Languages Affect Zero-Shot NMT

[Rios, Müller, Sennrich, WMT 2020]



reduce English bias

ensure each source language is paired with (small amounts of) non-English output, so that language token is respected

<2DE> espérons que ce ne sera pas le cas .

Results

reducing copy bias

	shared BPE	language-specific BPE shared vocabulary
EN↔X	31.6 ±0.12	31.2 ±0.60
zero-shot	15.4 ±3.14	20.5 ±0.43

reducing English bias

	Anglocentric	multi-directional
EN↔{CS,DE,FI,FR}	31.2 ±0.60	31.2 ±0.56
zero-shot	20.9 ±0.43	24.0 ±0.62

Results

reducing copy bias

	shared BPE	language-specific BPE shared vocabulary
EN↔X	31.6 ±0.12	31.2 ±0.60
zero-shot	15.4 ±3.14	20.5 ±0.43

reducing English bias

	Anglocentric	multi-directional
EN↔{CS,DE,FI,FR}	31.2 ±0.60	31.2 ±0.56
zero-shot	20.9 ±0.43	24.0 ±0.62

- no gains for EN↔X

Results

reducing copy bias

	shared BPE	language-specific BPE shared vocabulary
EN↔X	31.6 ±0.12	31.2 ±0.60
zero-shot	15.4 ±3.14	20.5 ±0.43

reducing English bias

	Anglocentric	multi-directional
EN↔{CS,DE,FI,FR}	31.2 ±0.60	31.2 ±0.56
zero-shot	20.9 ±0.43	24.0 ±0.62

- no gains for EN↔X
- large gains in zero-shot performance ☺

Conclusions: Lessons Learned

- cross-lingual transfer exciting for low-resource machine translation
- massively multilingual models have potential...
- ...but capacity is crucial
- if you want zero-shot generalisation, be careful about design:
  - model learns spurious correlations from Anglocentric model
  - synthetic (or real) training data better than being fully zero-shot
- language representations matter:
  - sharing subword segmentation improves transfer...
  - ...but can also cause undesirable copy behaviour

Thank you for your attention

Resources

- code and data for massively multilingual models:
  `https://github.com/EdinburghNLP/opus-100-corpus`
  `http://opus.nlpl.eu/opus-100.php`

Acknowledgments

This work has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreements 825460 (ELITR) and 825299 (GoURMET). Further funding from the European Research Council was received from the ERC Starting Grant BroadSem (678254).

This work has received funding from the Swiss National Science Foundation (SNF) in the projects CoNTra (grant number 105212_169888) and MUTAMUR (grant number 176727).

This work has been performed using resources provided by the Cambridge Tier-2 system operated by the University of Cambridge Research Computing Service (http://www.hpc.cam.ac.uk) funded by EPSRC Tier-2 capital grant EP/P020259/1.

This project has received support from Samsung Electronics Polska sp. z o.o. - Samsung R&D Institute Poland.

This work has received research credits from the Google Cloud Platform.

Bibliography I

Aharoni, R., Johnson, M., and Firat, O. (2019).

Massively multilingual neural machine translation.

In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 3874–3884, Minneapolis, Minnesota. Association for Computational Linguistics.

Aji, A. F., Bogoychev, N., Heafield, K., and Sennrich, R. (2020).

In neural machine translation, what does transfer learning transfer?

In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 7701–7710, Online. Association for Computational Linguistics.

Amrhein, C. and Sennrich, R. (2020).

On Romanization for model transfer between scripts in neural machine translation.

In Findings of the Association for Computational Linguistics: EMNLP 2020, pages 2461–2469, Online. Association for Computational Linguistics.

Arivazhagan, N., Bapna, A., Firat, O., Aharoni, R., Johnson, M., and Macherey, W. (2019a).

The Missing Ingredient in Zero-Shot Neural Machine Translation.

CoRR, abs/1903.07091.

Arivazhagan, N., Bapna, A., Firat, O., Lepikhin, D., Johnson, M., Krikun, M., Chen, M. X., Cao, Y., Foster, G., Cherry, C., Macherey, W., Chen, Z., and Wu, Y. (2019b).

Massively multilingual neural machine translation in the wild: Findings and challenges.

CoRR, abs/1907.05019.

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).

Bibliography II

Edunov, S., Ott, M., Auli, M., and Grangier, D. (2018).
Understanding back-translation at scale.
In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 489–500, Brussels, Belgium. Association for Computational Linguistics.

Fan, A., Bhosale, S., Schwenk, H., Ma, Z., El-Kishky, A., Goyal, S., Baines, M., Celebi, O., Wenzek, G., Chaudhary, V., Goyal, N., Birch, T., Liptchinsky, V., Edunov, S., Grave, E., Auli, M., and Joulin, A. (2020).
Beyond english-centric multilingual machine translation.

Freitag, M. and Firat, O. (2020).
Complete multilingual neural machine translation.

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

Gheini, M. and May, J. (2019).
A universal parent model for low-resource neural machine translation transfer.
CoRR, abs/1909.06516.

Gu, J., Wang, Y., Cho, K., and Li, V. O. (2019).
Improved zero-shot neural machine translation via ignoring spurious correlations.
In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 1258–1268, Florence, Italy. Association for Computational Linguistics.

Bibliography III

Jean, S., Cho, K., Memisevic, R., and Bengio, Y. (2015).
On Using Very Large Target Vocabulary for Neural Machine Translation.
In
Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Co
pages 1–10, Beijing, China.

Johnson, M., Schuster, M., Le, Q. V., Krikun, M., Wu, Y., Chen, Z., Thorat, N., Viégas, F., Wattenberg, M., Corrado, G., Hughes, M., and Dean, J. (2017).
Google's Multilingual Neural Machine Translation System: Enabling Zero-Shot Translation.
Transactions of the Association for Computational Linguistics, 5:339–351.

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.

Rios, A., Müller, M., and Sennrich, R. (2020).
Subword segmentation and a single bridge language affect zero-shot neural machine translation.
In Proceedings of the Fifth Conference on Machine Translation, pages 451–460, Online. Association for Computational Linguistics.

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 (Volume 1: Long Papers), pages 86–96, Berlin, Germany.

Sennrich, R., Haddow, B., and Birch, A. (2016b).
Neural Machine Translation of Rare Words with Subword Units.
In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1715–1725, Berlin, Germany.

Bibliography IV

Tiedemann, J. (2020).

The tatoeba translation challenge – realistic data sets for low resource and multilingual mt.
In Proceedings of the Fifth Conference on Machine Translation, pages 1172–1180, Online. Association for Computational Linguistics.

Toral, A., Castilho, S., Hu, K., and Way, A. (2018).

Attaining the unattainable? reassessing claims of human parity in neural machine translation.
In Proceedings of the Third Conference on Machine Translation: Research Papers, pages 113–123, Brussels, Belgium. Association for Computational Linguistics.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017).

Attention is All you Need.
In Advances in Neural Information Processing Systems 30, pages 5998–6008.

Zhang, B., Titov, I., and Sennrich, R. (2019).

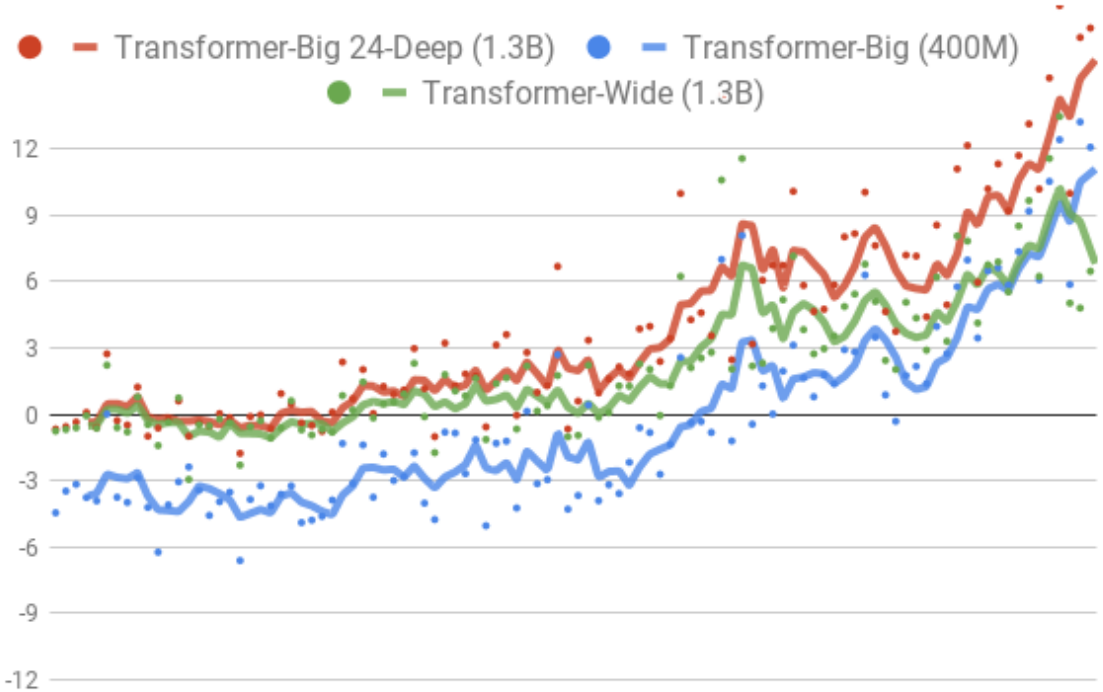
Improving Deep Transformer with Depth-Scaled Initialization and Merged Attention.
In
Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Confe
Hong Kong, China. Association for Computational Linguistics.

Zoph, B., Yuret, D., May, J., and Knight, K. (2016).

Transfer Learning for Low-Resource Neural Machine Translation.
In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pages 1568–1575, Austin, Texas.

Massively Multilingual MT: Prior Work

deeper/wider models help:



but what about zero-shot generalization?

[Arivazhagan et al., 2019b]

Zero-Shot MT: Prior Work

system	BLEU
supervised (multi-way parallel)	23.9
zero-shot (multi-way parallel)	21.3
zero-shot (English-centric)	14.8

DE→IT performance (IWSLT corpus {EN,DE,IT,NL,RO})

good zero-shot generalisation in multi-way parallel training ☺
performance much worse in (more realistic) English-centric setup ☹

[Gu et al., 2019]

Zero-Shot MT: Prior Work

in imbalanced (or English-centric) settings, source and target language are correlated...
...but we don't want model to learn this correlation

proposed solutions:

- LM pretraining
- back-translation

how to scale to massively multilingual setting?

[Gu et al., 2019]

Reduce copy bias: results

	EN↔X	zero-shot
shared BPE	31.6 ±0.12	15.4 ±3.14
language-specific BPE, shared vocabulary	31.2 ±0.60	20.5 ±0.43

Reduce copy bias: results

	EN↔X	zero-shot
shared BPE	31.6 ±0.12	15.4 ±3.14
language-specific BPE, shared vocabulary	31.2 ±0.60	20.5 ±0.43

- small effects on trained directions

Reduce copy bias: results

	EN↔X	zero-shot
shared BPE	31.6 ±0.12	15.4 ±3.14
language-specific BPE, shared vocabulary	31.2 ±0.60	20.5 ±0.43

- small effects on trained directions
- on zero-shot directions, representation sharing has large effects

Reduce copy bias: results

string overlap between source and target text:

	BPE	subwords	words
training set	shared	9.7%	5.7%
	language-specific	8.0%	5.7%
translations	shared	24.8%	20.6%
	language-specific	7.0%	4.7%

Reduce English bias: the problem with Anglocentricity

<2DE> espérons que ce ne sera pas le cas .

can model learn to respect indicator token if French input is always paired with English output in training?

Reduce English bias: making indicator tokens matter

language pair(s)	sentence pairs
EN↔{CS,DE,FI,FR}	5 000 000
CS↔DE	350 000
FI↔FR	350 000
DE↔FI	0
DE↔FR	0
CS↔FR	0

- each source language is paired with 2nd target
- small increase in total data: 20M → 20.7M
  → much more lightweight than fully multi-way corpus
- generalization tested on remaining zero-shot languages

Result on off-target translation

language of MT output (according to langdetect) for zero-shot directions

	target ✓	EN ✗	source ✗
Anglocentric	93.55%	1.69%	0.89%
multi-directional	95.30%	0.78%	0.44%

Comparison to related work

- minimize cosine distance of sentence pair representations

  [Arivazhagan et al., 2019a]

- zero-shot backtranslation [Gu et al., 2019]: 250k sentence pairs per direction

	EN↔X	zero-shot
Anglocentric	31.2	20.9
+cosine	31.1	21.3
+bt	30.4	21.6
multi-directional	31.2	24.0
+cosine	31.0	23.9
+bt	30.5	25.2

Comparison to related work

- minimize cosine distance of sentence pair representations

  [Arivazhagan et al., 2019a]

- zero-shot backtranslation [Gu et al., 2019]: 250k sentence pairs per direction

	EN↔X	zero-shot
Anglocentric	31.2	20.9
+cosine	31.1	21.3
+bt	30.4	21.6
multi-directional	31.2	24.0
+cosine	31.0	23.9
+bt	30.5	25.2

- other methods also work... ☺

Comparison to related work

- minimize cosine distance of sentence pair representations

  [Arivazhagan et al., 2019a]

- zero-shot backtranslation [Gu et al., 2019]: 250k sentence pairs per direction

	EN↔X	zero-shot
Anglocentric	31.2	20.9
+cosine	31.1	21.3
+bt	30.4	21.6
multi-directional	31.2	24.0
+cosine	31.0	23.9
+bt	30.5	25.2

- other methods also work... ☺
- ...but reducing Anglocentricity is worth it ☺☺