Multilingual Machine Translation

Ondřej Bojar

April 30, 2020




EUROPEAN UNION
European Structural and Investment Fund
Operational Programme Research,
Development and Education

Charles University
Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics

Outline

- Motivation for using more than 2 languages.
- Transfer Learning.
  - Catastrophic Forgetting.
  - Trivial Transfer Learning.
- Multi-Lingual NMT.
- Massively Multi-Lingual NMT.

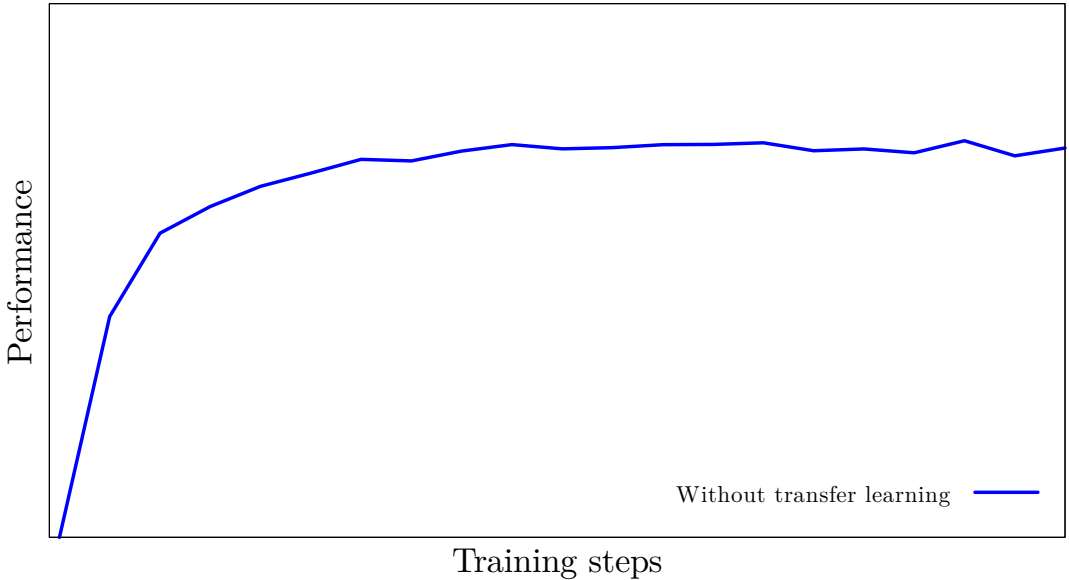
Many slides on transfer learning by Tom Kocmi.
Many slides on multilingual models by Rico Sennrich and Adam Lopez.

Why Multilingual MT

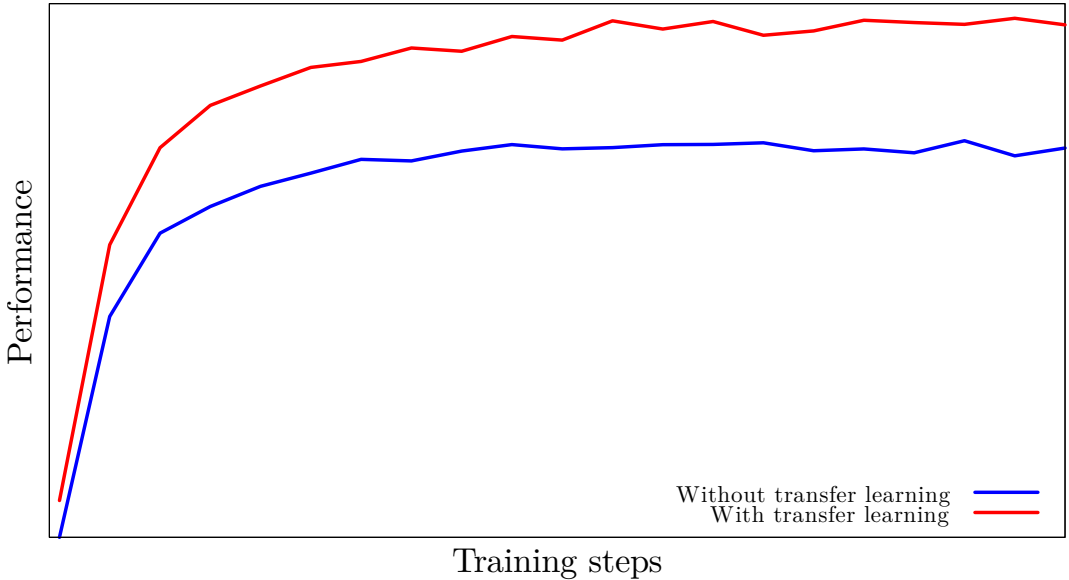
- Help in low-resource settings.
  - Words, morphemes or syntactic patterns common to more languages.
  - Learning can reuse patterns seen in another dataset.
- Improve translation quality.
  - Words are ambiguous, the third language can disambiguate.
- Truly multi-lingual environments.
  - United Nations: 6 languages.
  - EU official languages: 24.
  - EUROSAI official languages: 43.
  - INTOSAI official languages...

Transfer Learning

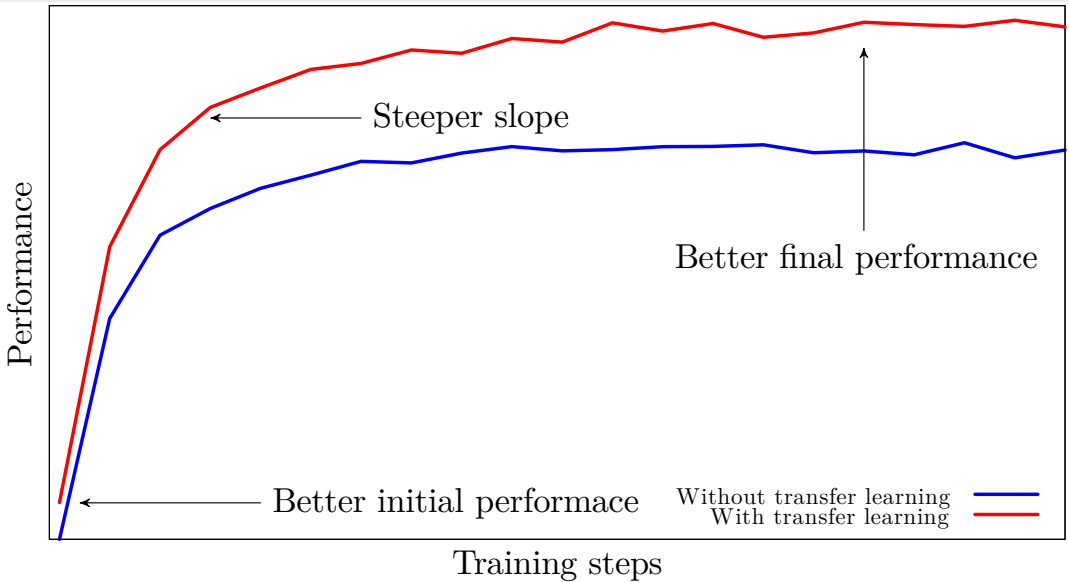
Motivation for NN Transfer Learning



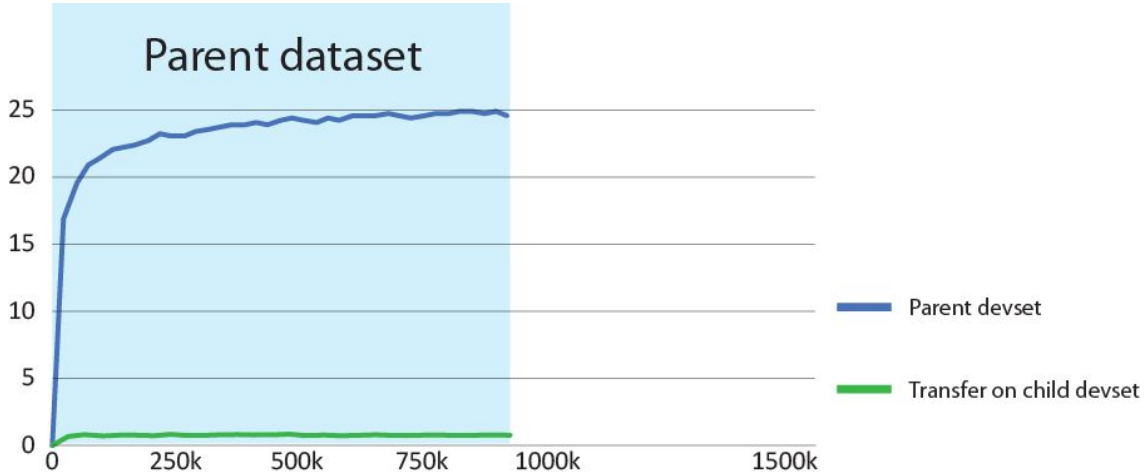
Motivation for NN Transfer Learning



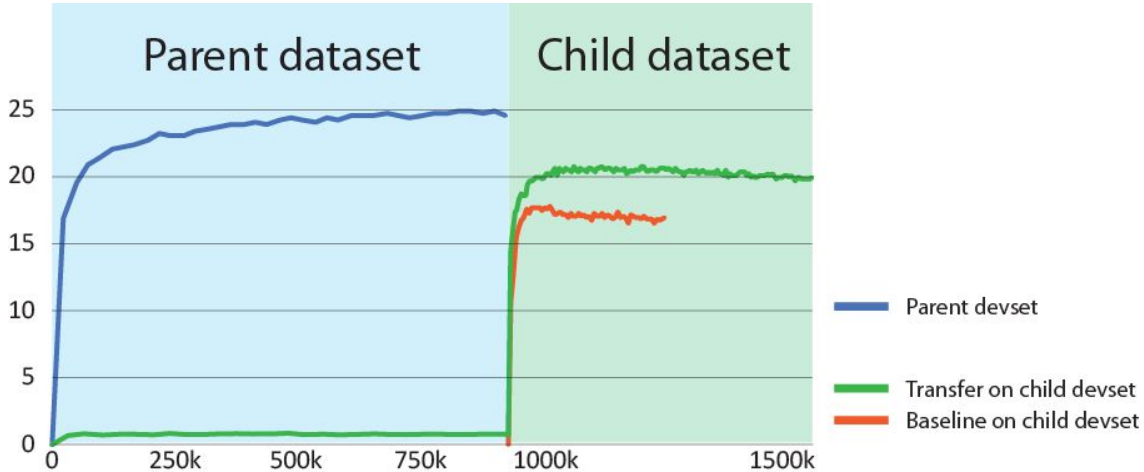
Motivation for NN Transfer Learning



Steps of Transfer Learning

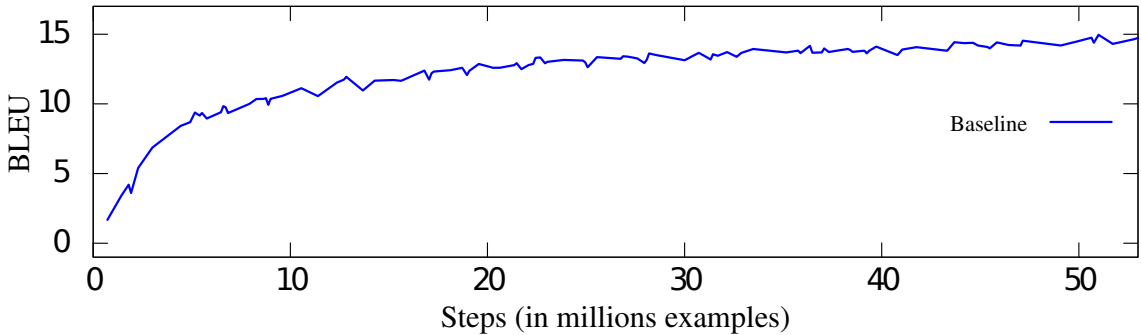


Steps of Transfer Learning



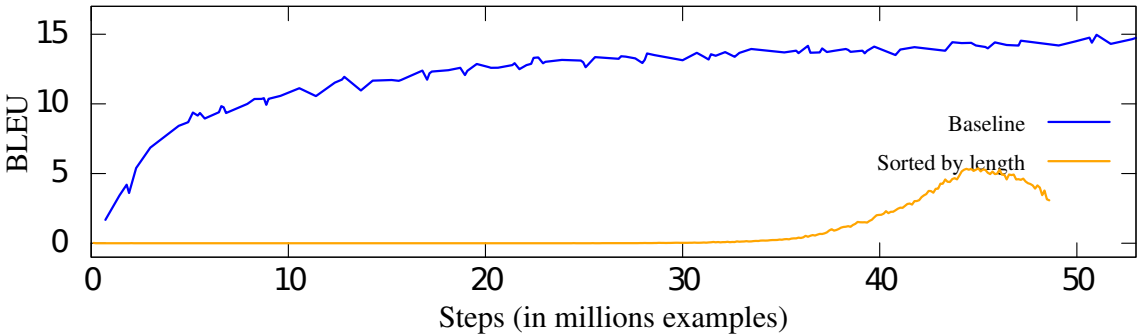
Interlude: Catastrophic Forgetting

- Kocmi and Bojar (2017) explore curriculum learning:
  - Start with simpler sentences first, add complex ones later.



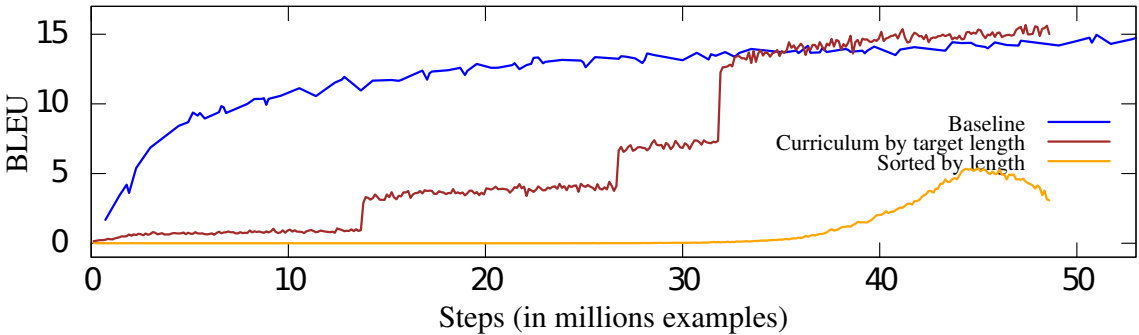
Interlude: Catastrophic Forgetting

- Kocmi and Bojar (2017) explore curriculum learning:
  - Start with simpler sentences first, add complex ones later.
- When "simpler" means "shorter":



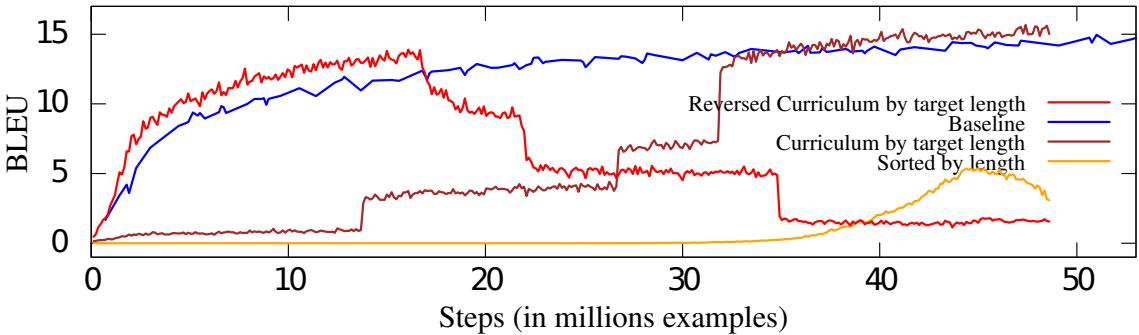
Interlude: Catastrophic Forgetting

- Kocmi and Bojar (2017) explore curriculum learning:
  - Start with simpler sentences first, add complex ones later.
- When "simpler" means "shorter":
  - Clear jumps in score as bins of longer sentences are allowed.



Interlude: Catastrophic Forgetting

- Kocmi and Bojar (2017) explore curriculum learning:
  - Start with simpler sentences first, add complex ones later.
- When "simpler" means "shorter":
  - Clear jumps in score as bins of longer sentences are allowed.
  - Reversed curriculum unlearns to produce long sentences.



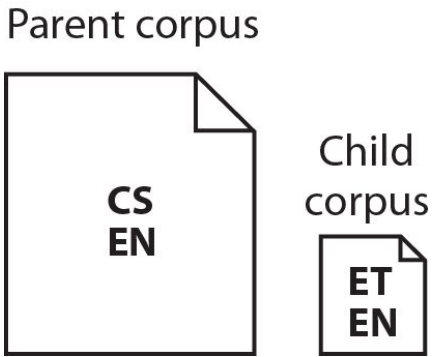
Trivial Transfer Learning

- Early works (Zoph et al., 2016; Nguyen and Chiang, 2017) target one common language (English).
- Kocmi and Bojar (2018) try even unrelated languages.

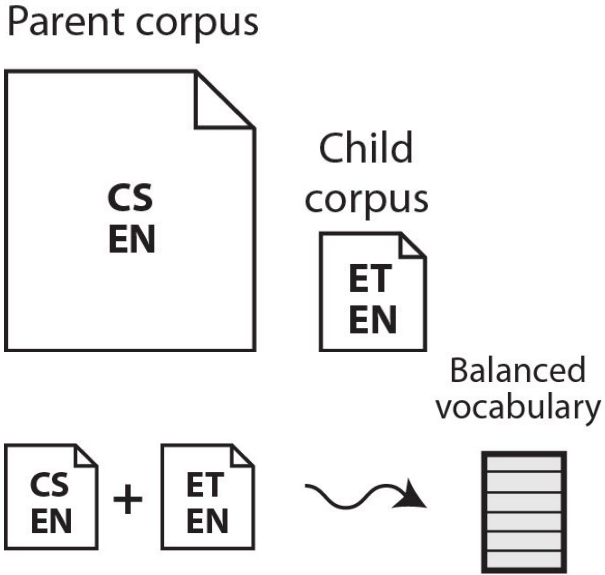
The trivial procedure:

- Train on one pair ("parent"), switch corpus to another ("child").
- The only requirement: joint subword units across all langs.

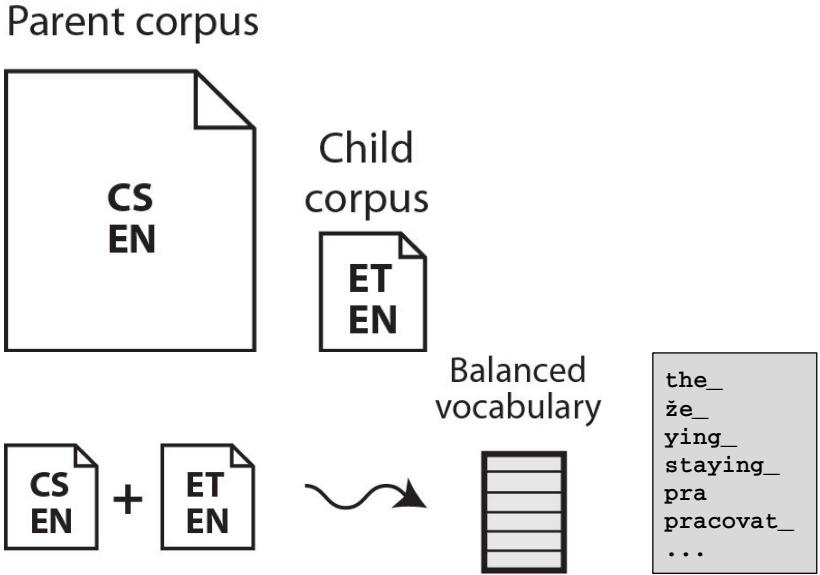
Getting Balanced Vocabulary



Getting Balanced Vocabulary



Getting Balanced Vocabulary



English on Same Side

Child model: Slovak

Parent model	Corpus size difference	Direction	Baseline (BLEU)	Transfer (BLEU)	Δ (BLEU)
Czech	9x	from English	16.13	17.75	1.62 *
Czech	9x	to English	19.19	22.42	3.23 *

* statistically significant

English on Same Side

Child model: Slovak

Parent model	Corpus size difference	Direction	Baseline (BLEU)	Transfer (BLEU)	Δ (BLEU)
Czech	9x	from English	16.13	17.75	1.62 *
Czech	9x	to English	19.19	22.42	3.23 *

* statistically significant

English on Same Side

Child model: Slovak

Parent model	Corpus size difference	Direction	Baseline (BLEU)	Transfer (BLEU)	Δ (BLEU)
Czech	9x	from English	16.13	17.75	1.62 *
Czech	9x	to English	19.19	22.42	3.23 *

Child model: Estonian

Parent model	Corpus size difference	Direction	Baseline (BLEU)	Transfer (BLEU)	Δ (BLEU)
Finnish	3.5x	from English	17.03	19.74	2.71 *
Russian	16x	from English	17.03	20.09	3.06 *
Czech	50x	from English	17.03	20.41	3.38 *
Finnish	3.5x	to English	21.74	24.18	2.44 *
Russian	16x	to English	21.74	23.54	1.80 *

* statistically significant

English on Same Side

Child model: Slovak

Parent model	Corpus size difference	Direction	Baseline (BLEU)	Transfer (BLEU)	Δ (BLEU)
Czech	9x	from English	16.13	17.75	1.62 *
Czech	9x	to English	19.19	22.42	3.23 *

Child model: Estonian

Parent model	Corpus size difference	Direction	Baseline (BLEU)	Transfer (BLEU)	Δ (BLEU)
Finnish	3.5x	from English	17.03	19.74	2.71 *
Russian	16x	from English	17.03	20.09	3.06 *
Czech	50x	from English	17.03	20.41	3.38 *
Finnish	3.5x	to English	21.74	24.18	2.44 *
Russian	16x	to English	21.74	23.54	1.80 *

* statistically significant

English on Same Side

Child model: Slovak

Parent model	Corpus size difference	Direction	Baseline (BLEU)	Transfer (BLEU)	Δ (BLEU)
Related	9x	from English	16.13	17.75	1.62 *
Related	9x	to English	19.19	22.42	3.23 *

Child model: Estonian

Parent model	Corpus size difference	Direction	Baseline (BLEU)	Transfer (BLEU)	Δ (BLEU)
Related	3.5x	from English	17.03	19.74	2.71 *
Cyrillic	16x	from English	17.03	20.09	3.06 *
Biggest	50x	from English	17.03	20.41	3.38 *
Related	3.5x	to English	21.74	24.18	2.44 *
Cyrillic	16x	to English	21.74	23.54	1.80 *

* statistically significant

English on Same Side, Parent Low-Resource

Child model: Finnish

Parent model	Corpus size difference	Direction	Baseline (BLEU)	Transfer (BLEU)	Δ (BLEU)
Estonian	0.3x	from English	19.50	20.07	0.57 *
Estonian	0.3x	to English	24.40	23.95	-0.45

Child model: Czech

Parent model	Corpus size difference	Direction	Baseline (BLEU)	Transfer (BLEU)	Δ (BLEU)
Slovak	0.1x	from English	23.48	22.99	-0.49 *
Slovak	0.1x	to English	29.61	28.20	-1.41 *

English on Same Side, Parent Low-Resource

Child model: Finnish

Parent model	Corpus size difference	Direction	Baseline (BLEU)	Transfer (BLEU)	Δ (BLEU)
Estonian	0.3x	from English	19.50	20.07	0.57 *
Estonian	0.3x	to English	24.40	23.95	-0.45

Child model: Czech

Parent model	Corpus size difference	Direction	Baseline (BLEU)	Transfer (BLEU)	Δ (BLEU)
Slovak	0.1x	from English	23.48	22.99	-0.49 *
Slovak	0.1x	to English	29.61	28.20	-1.41 *

English on the Other Side

Parent model	Child model	Corpus size amplification	Baseline (BLEU)	Transfer (BLEU)	Δ (BLEU)	Parent Aligned Δ
EN - Finnish	Estonian - EN	3.5x	21.74	22.75	1.01 *	2.44 *
EN - Russian	Estonian - EN	16x	21.74	23.12	1.38 *	1.80 *
EN - Czech	Estonian - EN	50x	21.74	22.80	1.06 *	
Finnish - EN	EN - Estonian	3.5x	17.03	18.19	1.16 *	2.71 *
Russian - EN	EN - Estonian	16x	17.03	18.16	1.13 *	3.06 *

No Language in Common

Child model: Estonian to English

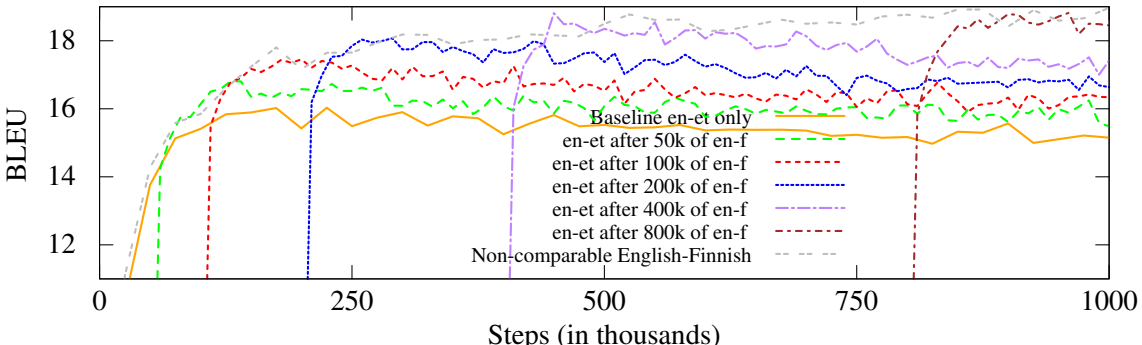
Parent model	Corpus size amplification	Baseline (BLEU)	Transfer (BLEU)	Δ (BLEU)
Arabic - Russian	12x	21.74	22.23	0.49
Spanish - French	12x	21.74	22.24	0.50 *
Spanish - Russian	12x	21.74	22.52	0.78 *
French - Russian	12x	21.74	22.40	0.66 *

No Language in Common

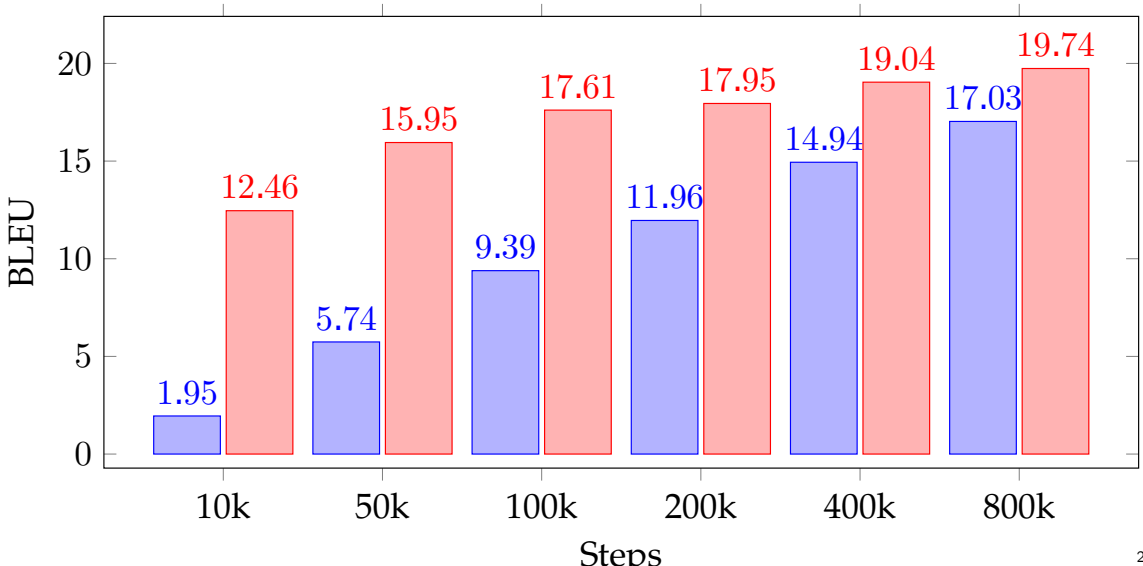
Child model: Estonian to English

Parent model	Corpus size amplification	Baseline (BLEU)	Transfer (BLEU)	Δ (BLEU)
Arabic - Cyrillic	12x	21.74	22.23	0.49
Spanish - French	12x	21.74	22.24	0.50 *
Spanish - Cyrillic	12x	21.74	22.52	0.78 *
French - Cyrillic	12x	21.74	22.40	0.66 *

The Better the Parent, the Better the Child



The Lesser the Child, the Bigger the Gain



Why it Helps? Not Really Vocabulary (1/2)

	Length	BLEU Components	BP
Base ENET	35326	48.1/21.3/11.3/6.4	0.979
ENRU+ENET	35979	51.0/24.2/13.5/8.0	0.998
ENCS+ENET	35921	51.7/24.6/13.7/8.1	0.996

(The reference length in the matching tokenization was 36062.)

- Child models produce longer outputs $\Rightarrow$ lower brevity penalty.
- But n-gram precisions also better.

1-gram present in	ENRU+ENET	ENCS+ENET
Child, Base, Ref	15902 (44.2 %)	15924 (44.3 %)
Child only	9635 (26.8 %)	9485 (26.4 %)
Child, Base	7209 (20.0 %)	7034 (19.6 %)
Child, Ref	**3233 (9.0 %)**	**3478 (9.7 %)**
Total	35979 (100.0 %)	35921 (100.0 %)

- The 3k better toks are regular ET words, not NEs or numbers.

Why it Helps? Not Really Vocabulary (2/2)

Why it Helps? Sentence Lengths Somewhat

	Parent	
Sentence lengths	BLEU	Avg. words
1-10 words	8.57	10.9
10-20 words	16.21	15.4
20-40 words	12.59	21.9
40-60 words	5.76	35.5
1-60 words	22.30	15.3

Why it Helps? Sentence Lengths Somewhat

	Parent		Child	
Sentence lengths	BLEU	Avg. words	BLEU	Avg. words
1-10 words	8.57	10.9	16.57	15.3
10-20 words	16.21	15.4	17.48	15.3
20-40 words	12.59	21.9	17.99	15.3
40-60 words	5.76	35.5	16.80	15.5
1-60 words	22.30	15.3	19.15	15.4

Why it Helps? Sentence Lengths Somewhat

	Parent		Child	
Sentence lengths	BLEU	Avg. words	BLEU	Avg. words
1-10 words	8.57	10.9	16.57	15.3
10-20 words	16.21	15.4	17.48	15.3
20-40 words	12.59	21.9	17.99	15.3
40-60 words	5.76	35.5	16.80	15.5
1-60 words	22.30	15.3	19.15	15.4

Multilingual MT

Multilingual MT Configurations

- Pivot translation (Cascading).
- Multi-lingual source (also called multi-way).
- Multi-lingual multi-source.
- Multi-lingual target.
- Multi-lingual multi-target.
- Both sides multi-lingual.
- (Both sides multi-lingual, multi-source, multi-target. ;-)
- Zero-shot training.
  - i.e. translating an unseen pair when both the source and target langs were covered in the training data in other pairs.
- "Beyond zero-shot" is translating from an unseen language.

Multi-Target and Multi-Source MT

- Multi-Target focus: Efficiency
  - Decrease hardware resources compared using many separate models.
- Multi-Source focus: Resolving ambiguity thanks to existing translation
  - E.g. Translating German "Schloss" to French is easier if we can feed in the English translation ("castle" or "lock").
- Training on: Multi-parallel or bi-parallel multilingual corpora.

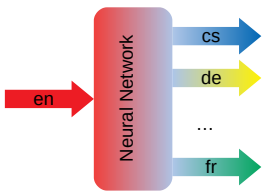


Figure 1: Multi-Target MT

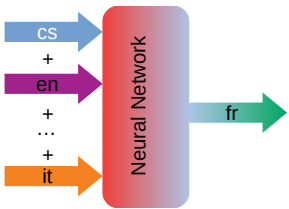


Figure 2: Multi-Source MT

Ideal: Flexible Multi-Lingual MT

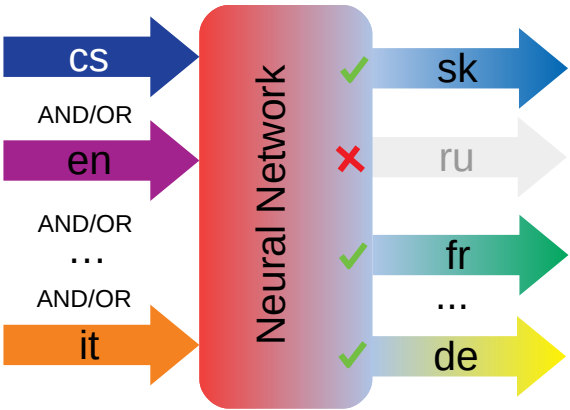


Figure 3: Flexible multilingual MT

Multi-source translation

Quite an old idea (e.g. Och & Ney 2001)

Table 4: Absolute improvements in WER combining two languages using method MAX compared with the best WER obtained by any of the two languages.

	fr	pt	es	it	sv	da	nl
fr	0.0	1.5	1.2	0.5	2.7	1.9	0.8
pt		0.0	2.2	2.1	4.0	3.4	1.3
es			0.0	2.4	3.9	2.6	1.7
it				0.0	3.5	3.2	1.6
sv					0.0	2.7	1.7
da						0.0	4.3
nl							0.0

Table 5: Absolute improvements in WER combining two languages using method PROD compared with the best WER obtained by any of the two languages.

	fr	pt	es	it	sv	da	nl
fr	0.0	0.8	0.1	0.4	1.0	0.8	-0.2
pt		0.0	2.6	2.1	2.6	2.8	-0.1
es			0.0	2.4	3.4	3.7	1.1
it				0.0	1.9	3.0	0.3
sv					0.0	1.8	0.5
da						0.0	1.5
nl							0.0

Table 6: Language combination using method MAX.

languages	WER	PER
fr	55.3	45.3
fr+sv	52.6	43.7
fr+sv+es	**52.0**	**43.2**
fr+sv+es+pt	52.3	43.6
fr+sv+es+pt+it	52.7	44.0
fr+sv+es+pt+it+da	52.5	43.9

Table 7: Language combination using method PROD.

languages	WER	PER
fr	55.3	45.3
fr+sv	54.3	44.5
fr+sv+es	51.0	41.4
fr+sv+es+pt	50.2	40.2
fr+sv+es+pt+it	49.8	39.8
fr+sv+es+pt+it+da	**48.8**	**39.1**

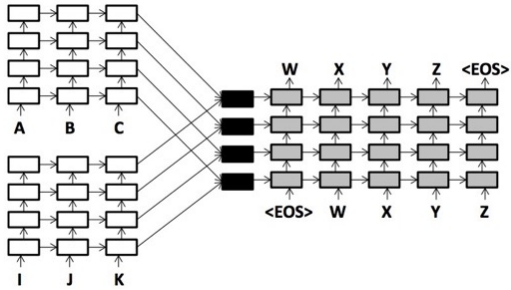
Multi-source translation

- Assorted techniques to do this in IBM-style or phrase-based MT.
- Difficult to model directly due to independence assumptions of these models.
- Usually done as a kind of system combination (merging the output of two MT systems).
- But this introduces other problems, e.g. decoding.
- Fundamentally, it's interpolation of conditional LMs.

Direct multi-source

Zoph & Knight 2016

- Directly learns and uses *p(English|French,German)*
- For attention: two context vectors (uses p-local attention of Luong, et al, but could use other methods).



Multi-way MT

Firat et al. 2016 (two papers)

- Assume only many bilingual parallel corpora.
- For *N* languages: learn *N* encoders and *N* decoders.
- But what about attention?

Multi-way MT

Firat et al. 2016 (two papers)

- Assume only many bilingual parallel corpora.
- For *N* languages: learn *N* encoders and *N* decoders.
- But what about attention?

$$p(f_i|f_{i-1},...,f_1,\mathbf{e}) = g(f_{i-1},s_i,c_i)$$

$$c_i = \sum_{j=1}^{|\mathbf{e}|} \alpha_{ij} h_j \qquad \alpha_{ij} = \frac{\exp(a_{ij})}{\sum_{k=1}^{|\mathbf{e}|} \exp(a_{ik})} \qquad a_{ij} = a(s_{i-1},h_j)$$

Multi-way MT

Firat et al. 2016 (two papers)

- As in Bahdanu et al. (2014), attention mechanism is a feedforward function of both decoder hidden state and encoder context vector.

- **Shared** between all encoders and decoders.

$$p(f_i|f_{i-1}, ..., f_1, \mathbf{e}) = g(f_{i-1}, s_i, c_i)$$

Everything we need is right here!

$$c_i = \sum_{j=1}^{|\mathbf{e}|} \alpha_{ij} h_j$$

$$\alpha_{ij} = \frac{\exp(a_{ij})}{\sum_{k=1}^{|\mathbf{e}|} \exp(a_{ik})}$$

$$a_{ij} = a(s_{i-1}, h_j)$$

Multi-way MT

Firat et al. 2016 (two papers)

	Size	Single	Single+DF	Multi
En→Fi	100k	5.06/3.96	4.98/3.99	6.2/**5.17**
	200k	7.1/6.16	7.21/6.17	8.84/**7.53**
	400k	9.11/7.85	9.31/8.18	11.09/**9.98**
	800k	11.08/9.96	11.59/10.15	12.73/**11.28**
De→En	210k	14.27/13.2	14.65/13.88	16.96/**16.26**
	420k	18.32/17.32	18.51/17.62	19.81/**19.63**
	840k	21/19.93	21.69/20.75	22.17/**21.93**
	1.68m	23.38/23.01	23.33/22.86	23.86/**23.52**
En→De	210k	11.44/11.57	11.71/11.16	12.63/**12.68**
	420k	14.28/14.25	14.88/15.05	15.01/**15.67**
	840k	17.09/17.44	17.21/17.88	17.33/**18.14**
	1.68m	19.09/19.6	19.36/20.13	19.23/**20.59**

Table 2: BLEU scores where the target pair's parallel corpus is constrained to be 5%, 10%, 20% and 40% of the original size. We report the BLEU scores on the development and test sets (separated by /) by the single-pair model (Single), the single-pair model with monolingual corpus (Single+DF) and the proposed multi-way, multilingual model (Multi).

Low-resource **simulation** (using high-resource European languages)

Multi-way MT

Firat et al. 2016 (two papers)

			Fr (39m)		Cs (12m)		De (4.2m)		Ru (2.3m)		Fi (2m)	
		Dir	→ En	En →	→ En	En →	→ En	En →	→ En	En →	→ En	En →
(a) BLEU	Dev	Single	27.22	26.91	21.24	15.9	24.13	20.49	21.04	18.06	13.15	9.59
		Multi	26.09	25.04	21.23	14.42	23.66	19.17	21.48	17.89	12.97	8.92
	Test	Single	27.94	**29.7**	20.32	**13.84**	24	**21.75**	22.44	**19.54**	12.24	**9.23**
		Multi	**28.06**	27.88	**20.57**	13.29	**24.20**	20.59	**23.44**	19.39	**12.61**	8.98
(b) LL	Dev	Single	-50.53	-53.38	-60.69	-69.56	-54.76	-61.21	-60.19	-65.81	-88.44	-91.75
		Multi	-50.6	-56.55	-54.46	-70.76	-54.14	-62.34	-54.09	-63.75	-74.84	-88.02
	Test	Single	-43.34	**-45.07**	-60.03	**-64.34**	-57.81	**-59.55**	-60.65	-60.29	-88.66	-94.23
		Multi	**-42.22**	-46.29	**-54.66**	-64.80	**-53.85**	-60.23	**-54.49**	**-58.63**	**-71.26**	**-88.09**

Table 3: (a) BLEU scores and (b) average log-probabilities for all the five languages from WMT'15.

Multi-way MT

Firat et al. 2016 (two papers)

			Fr (39m)		Cs (12m)		De (4.2m)		Ru (2.3m)		Fi (2m)	
		Dir	→ En	En →	→ En	En →	→ En	En →	→ En	En →	→ En	En →
(a) BLEU	Dev	Single	27.22	26.91	21.24	15.9	24.13	20.49	21.04	18.06	13.15	9.59
		Multi	26.09	25.04	21.23	14.42	23.66	19.17	21.48	17.89	12.97	8.92
	Test	Single	27.94	**29.7**	20.32	**13.84**	24	**21.75**	22.44	**19.54**	12.24	**9.23**
		Multi	**28.06**	27.88	**20.57**	13.29	**24.20**	20.59	**23.44**	19.39	**12.61**	8.98
(b) LL	Dev	Single	-50.53	-53.38	-60.69	-69.56	-54.76	-61.21	-60.19	-65.81	-88.44	-91.75
		Multi	-50.6	-56.55	-54.46	-70.76	-54.14	-62.34	-54.09	-63.75	-74.84	-88.02
	Test	Single	-43.34	**-45.07**	-60.03	**-64.34**	-57.81	**-59.55**	-60.65	-60.29	-88.66	-94.23
		Multi	**-42.22**	-46.29	**-54.66**	-64.80	**-53.85**	-60.23	**-54.49**	**-58.63**	**-71.26**	**-88.09**

Table 3: (a) BLEU scores and (b) average log-probabilities for all the five languages from WMT'15.

ok, but what about multi-source?

Multi-way multi-source MT

Firat et al. 2016 (two papers)

- Still assumes only many bilingual parallel corpora.
- What to do if there are multiple input sentences?
- Early averaging (average context vectors). $\mathbf{c}_t = \frac{\mathbf{c}_t^1 + \mathbf{c}_t^2}{2}.$
- Late averaging (aka linear interpolation).

$$P(w_i|\boldsymbol{c}) = \sum_{k=1}^{K} \lambda_k(\boldsymbol{c}) P_k(w_i|\boldsymbol{c})$$

Early and late averaging are orthogonal, can be combined.

Multi-way multi-source MT

Firat et al. 2016 (two papers)

	Src	Trgt	Multi Test	Single Test
(a)	Es	En	28.32	27.48
(b)	Fr	En	27.93	27.21
(c)	En	Es	28.41	28.90
(d)	En	Fr	23.41	24.05

Table 2: One-to-one translation qualities using the multi-way, multilingual model and four separate single-pair models.

Multi-way multi-source MT

Firat et al. 2016 (two papers)

	Src	Trgt	Multi Test	Single Test
(a)	Es	En	28.32	27.48
(b)	Fr	En	27.93	27.21

Table 2: One-to-one translation qualities using the multi-way multilingual model and four separate single-pair models.

		Multi		Single	
		Dev	Test	Dev	Test
(a)	Early	31.89	31.35	–	–
(b)	Late	32.04	31.57	32.00	31.46
(c)	E+L	32.61	31.88	–	–

Table 3: Many-to-one quality (Es+Fr→En) using three translation strategies. Compared to Table 2 (a–b) we observe a significant improvement (up to 3+ BLEU), although the model was never trained in these many-to-one settings. The second column shows the quality by the ensemble of two separate single-pair models.

Multi-way multi-source MT

Firat et al. 2016 (two papers)

	Src	Trgt	Multi Test	Single Test
(a)	Es	En	28.32	27.48
(b)	Fr	En	27.93	27.21

Table 2: One-to-one translation qualities using the multi-way multilingual model and four separate single-pair models.

		Multi		Single	
		Dev	Test	Dev	Test
(a)	Early	31.89	31.35	–	–
(b)	Late	32.04	31.57	32.00	31.46
(c)	E+L	32.61	31.88	–	–

Table 3: Many-to-one quality (Es+Fr→En) using three translation strategies. Compared to Table 2 (a–b) we observe a significant improvement (up to 3+ BLEU), although the model was never trained in these many-to-one settings. The second column shows the quality by the ensemble of two separate single-pair models.

Zero-shot MT

Firat et al. 2016 (two papers)

- Suppose our bilingual parallel data include a pair of languages for which we have no parallel data.

Spanish ⟷ English English ⟷ French

- Q: Can we use the multi-way encoder-decoder system to translate Spanish into French?

Zero-shot MT

Firat et al. 2016 (two papers)

- Suppose our bilingual parallel data include a pair of languages for which we have no parallel data.
- Q: Can we use the multi-way encoder-decoder system to translate Spanish into French?

	Pivot	Many-to-1	Dev	Test
(a)			< 1	< 1
(b)	$\surd$		20.64	20.4

Table 4: Zero-resource translation from Spanish (Es) to French (Fr) *without* finetuning. When pivot is $\surd$, English is used as a pivot language.

A: Not really

Must pivot (explicitly) through English

Zero-shot MT

Firat et al. 2016 (two papers)

- *Finetuning*: what if we use a small amount of parallel data in this setting?
- Q: Where would we get this data?

	Pivot	Many-to-1	Dev	Test
(a)			< 1	< 1
(b)	$\surd$		20.64	20.4

Table 4: Zero-resource translation from Spanish (Es) to French (Fr) *without* finetuning. When pivot is $\surd$, English is used as a pivot language.

Zero-shot MT

Firat et al. 2016 (two papers)

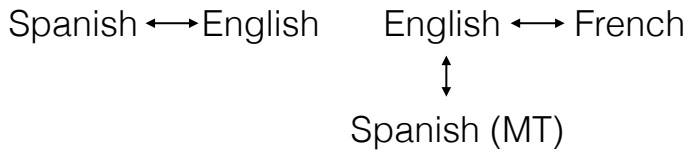
- *Finetuning*: what if we use a small amount of parallel data in this setting?
- Q: Where would we get this data? **Backtranslation**

Spanish ⟷ English English ⟷ French

Zero-shot MT

Firat et al. 2016 (two papers)

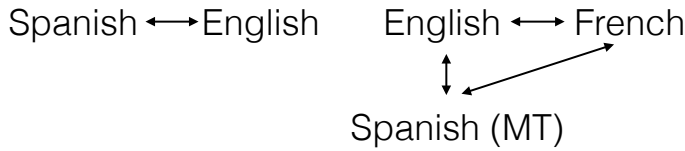
- *Finetuning*: what if we use a small amount of parallel data in this setting?
- Q: Where would we get this data? **Backtranslation**



Zero-shot MT

Firat et al. 2016 (two papers)

- *Finetuning*: what if we use a small amount of parallel data in this setting?
- Q: Where would we get this data? **Backtranslation**



Zero-shot MT

Firat et al. 2016 (two papers)

- *Finetuning*: what if we use a small amount of parallel data in this setting?

Pivot	Many-to-1		Pseudo Parallel Corpus 1k	10k	100k	1m
Single-Pair Models		Dev	–	–	–	–
		Test	–	–	–	–
√	No Finetuning		Dev: 20.64, Test: 20.4			
		Dev	0.28	10.16	15.61	17.59
		Test	0.47	10.14	15.41	17.61

Zero-shot MT

Firat et al. 2016 (two papers)

- *Finetuning*: what if we use a small amount of parallel data in this setting?

Pivot	Many-to-1		Pseudo Parallel Corpus 1k	10k	100k	1m	True Parallel Corpus 1k	10k	100k	1m
Single-Pair Models		Dev	–	–	–	–	–	–	11.25	21.32
		Test	–	–	–	–	–	–	10.43	20.35
√	No Finetuning		Dev: 20.64, Test: 20.4				–			
		Dev	0.28	10.16	15.61	17.59	0.1	8.45	16.2	20.59
		Test	0.47	10.14	15.41	17.61	0.12	8.18	15.8	19.97

Simple Data Mixing

Do we really need separate encoders and decoders?
... simply feed in various language pairs:

Source Sent 1 (De)	**2en** versetzen Sie sich mal in meine Lage !
Target Sent 1 (En)	put yourselves in my position .
Source Sent 2 (En)	**2nl** I flew on Air Force Two for eight years .
Target Sent 2 (Nl)	ik heb acht jaar lang met de Air Force Two gevlogen .

- The model of the same size will learn both pairs.
- Hopefully benefiting from various similarities.
- Risk of catastrophic forgetting.

See Johnson et al. (2016) or Ha et al. (2017).

Multi-source MT

Johnson et al. 2016 (Google)

- Sanity check: must not make things worse.

Table 1: Many to One: BLEU scores on various data sets for single language pair and multilingual models.

Model	Single	Multi	Diff
WMT German→English (oversampling)	30.43	30.59	+0.16
WMT French→English (oversampling)	35.50	35.73	+0.23
WMT German→English (no oversampling)	30.43	30.54	+0.11
WMT French→English (no oversampling)	35.50	36.77	+0.27
Prod Japanese→English	23.41	23.87	+0.46
Prod Korean→English	25.42	25.47	+0.05
Prod Spanish→English	38.00	38.73	+0.73
Prod Portuguese→English	44.40	45.19	+0.79

Multi-*target* MT

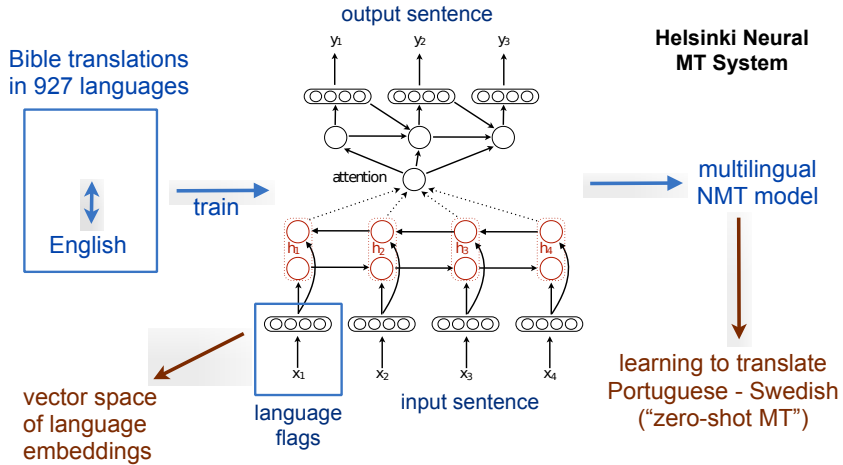
Johnson et al. 2016 (Google)

- Sanity check: must not make things worse.

Table 2: One to Many: BLEU scores on various data sets for single language pair and multilingual models.

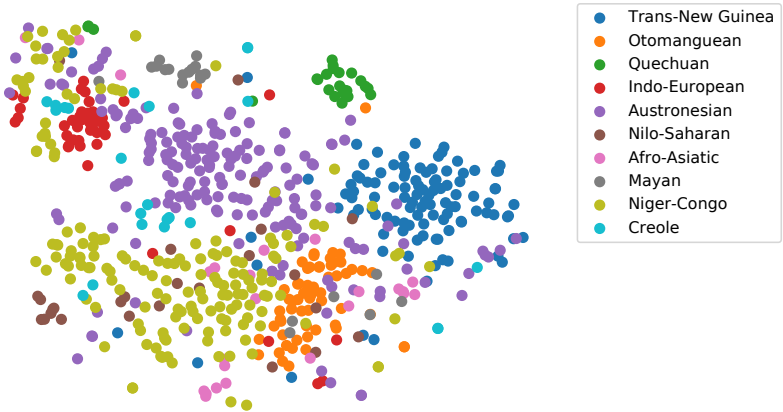
Model	Single	Multi	Diff
WMT English→German (oversampling)	24.67	24.97	+0.30
WMT English→French (oversampling)	38.95	36.84	-2.11
WMT English→German (no oversampling)	24.67	22.61	-2.06
WMT English→French (no oversampling)	38.95	38.16	-0.79
Prod English→Japanese	23.66	23.73	+0.07
Prod English→Korean	19.75	19.58	-0.17
Prod English→Spanish	34.50	35.40	+0.90
Prod English→Portuguese	38.40	38.63	+0.23

“Language Embeddings” from 927 Bibles



Tiedemann (2018)

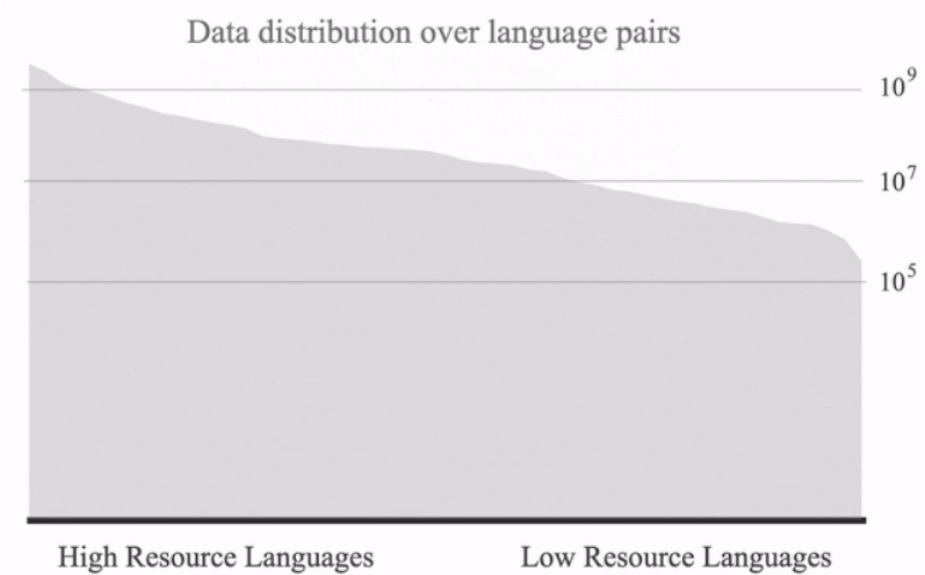
“Language Embeddings” from 927 Bibles



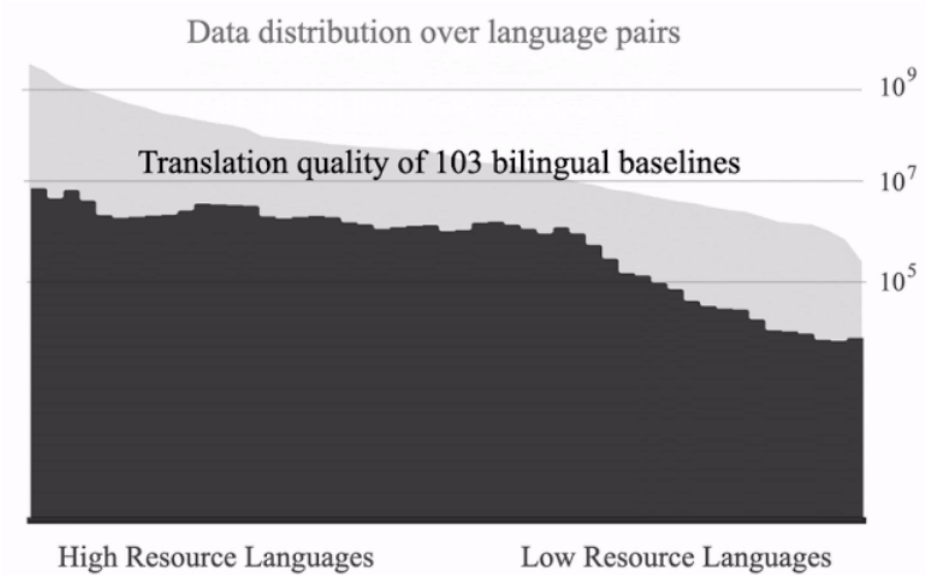
t-SNE of the language-embedding vectors, colored by language family.

Massively Multi-Lingual Models

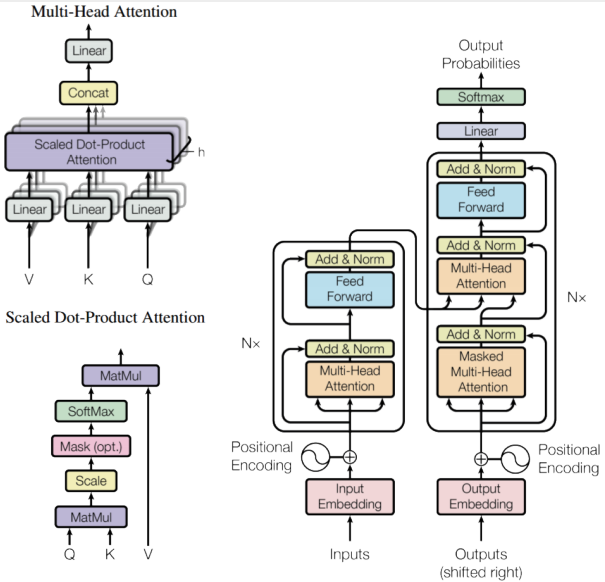
Available Data for EN↔100+ Langs



Translation Quality of Bilingual MT



Standard Transformer Model



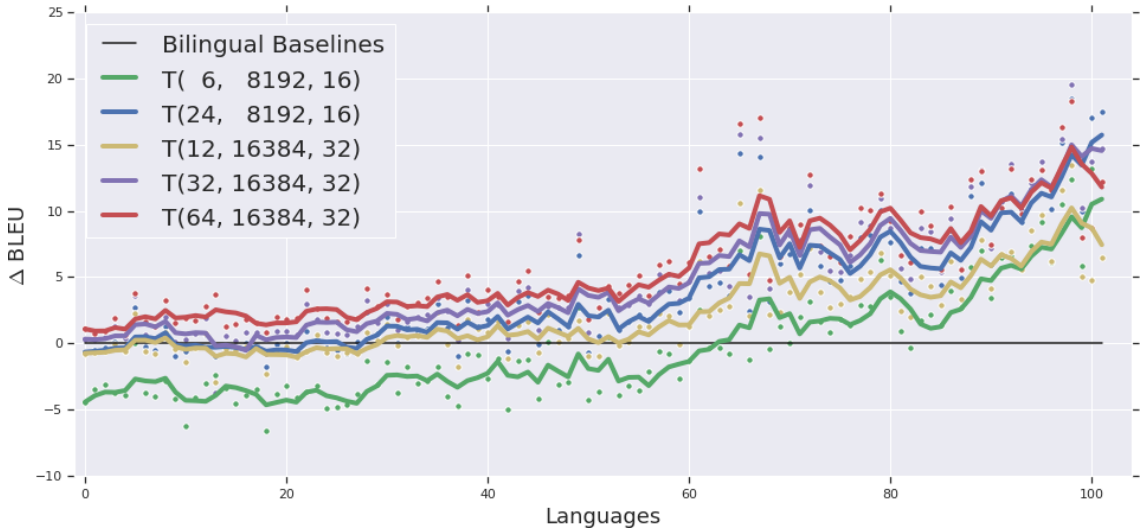
Google Transformer Sizes

GPipe (Huang et al., 2019) introduces microbatches for faster training of deep models across multiple GPUs.

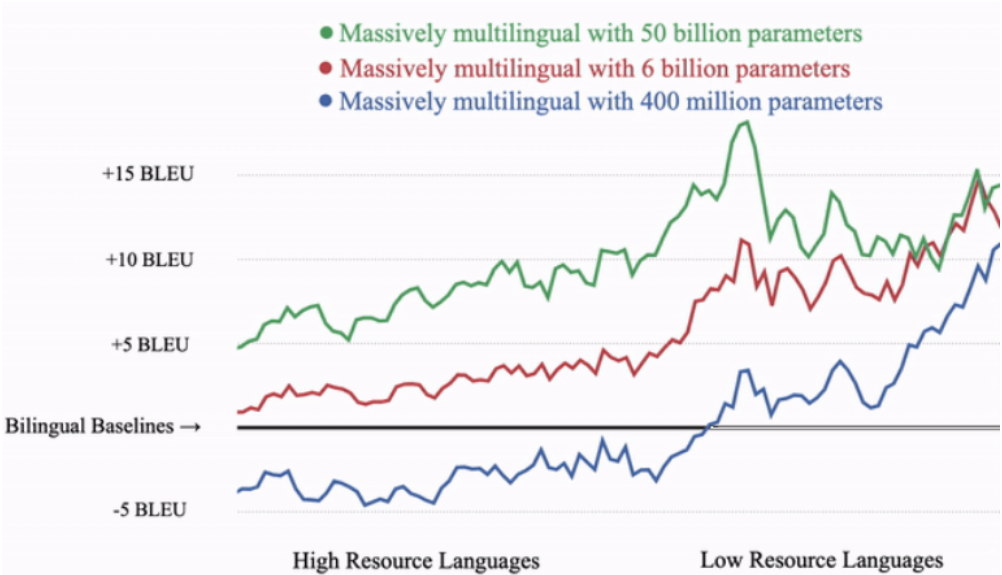
Enc/Dec Depth	FF Dim	Heads	Total Parameters	GPUs Used	
6	8192	16	400M	1 default	
12	16384	32	1.3B	2	“wide”
24	8192	16	1.3B	4	“deep”
32	16384	32	3.0B	8	
64	16384	32	6.0B	16	

- “Deep” better than “wide” on low-resource languages.
  - Indicates better generalization.
- Further tricks needed to keep the training stable.

Massively Multilingual Models



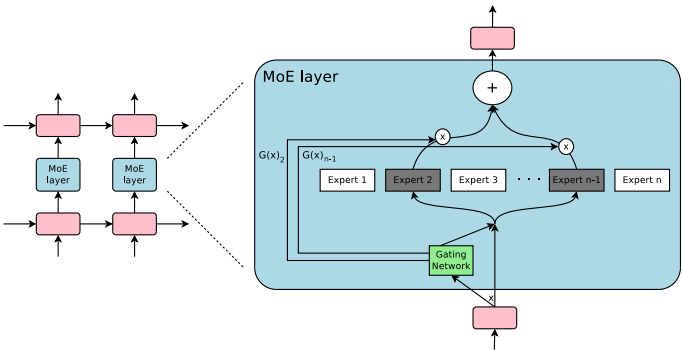
Massive Massively Multilingual Models



Google-Sized Experiment

The recent 50 billion parameters Transformer needed further trick:

- sparsely-gated mixture of experts (Shazeer et al., 2017):

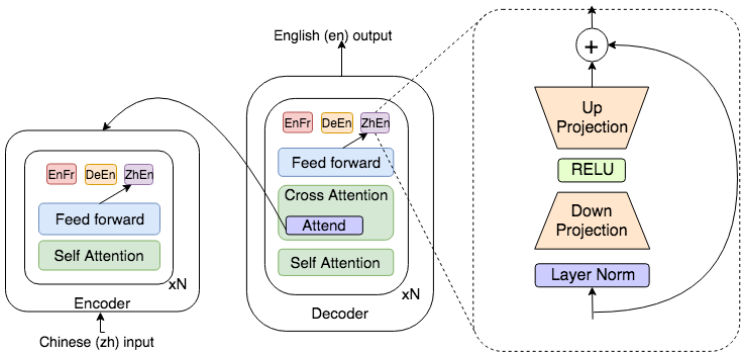


$\Rightarrow$ BLEU on 100 langs re-gained and improved by 125x larger model.

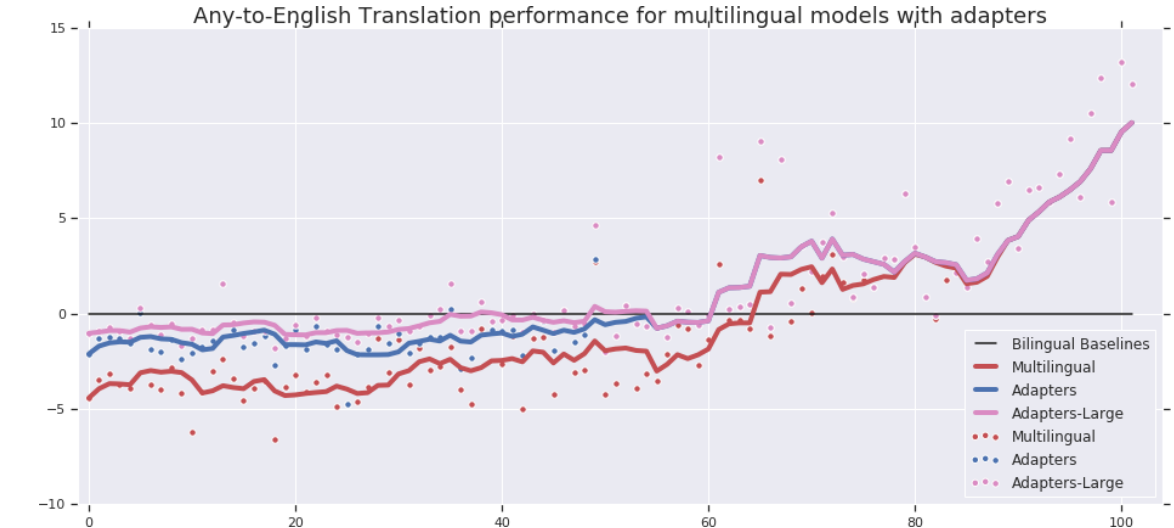
`https://ai.googleblog.com/2019/10/exploring-massively-multilingual.html`

Domain Adapters to Recover Practical Sizes

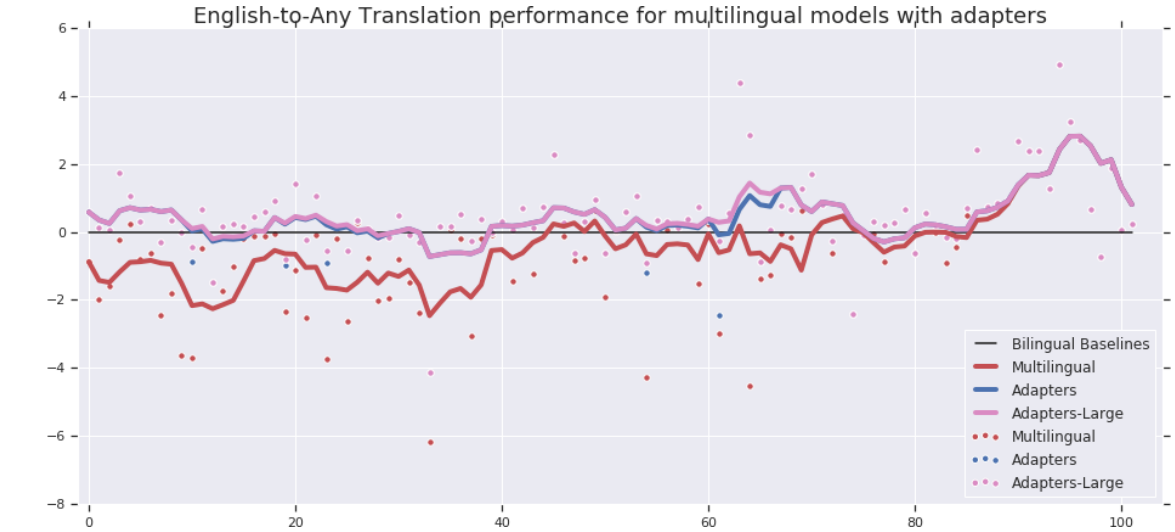
- Bapna and Firat (2019) propose tiny tunable "adapter" layers.
  1. Pretrain on a large mixed-language corpus.
  2. Inject adapter layers.
  3. Finetune adapter layers for each of the target tasks.



Domain Adapters into English



Domain Adapters from English



Summary

- Transfer learning in NMT works.
  - $\Rightarrow$ NMT can exploit more and less related data.
    - Trivial Transfer: Parent just has to be larger.
    - Even unrelated language pairs can help.
    - *Probably* better initialization.
- Multi-source, multi-target, ..., flexible multi-lingual setups.
- Language families emerge in language token embedding.
- Model capacity is the bottleneck.
  - Models 125x large for 100 languages in one model allow gains on high-resource languages, too.
  - With tiny adaptors instead of mixture of experts model sizes can decrease again.

References

Ankur Bapna and Orhan Firat. 2019. Simple, scalable adaptation for neural machine translation. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 1538–1548, Hong Kong, China, November. Association for Computational Linguistics.

Thanh-Le Ha, Jan Niehues, and Alexander H. Waibel. 2017. Effective strategies in zero-shot neural machine translation. CoRR, abs/1711.07893.

Yanping Huang, Youlong Cheng, Ankur Bapna, Orhan Firat, Dehao Chen, Mia Chen, HyoukJoong Lee, Jiquan Ngiam, Quoc V Le, Yonghui Wu, and zhifeng Chen. 2019. Gpipe: Efficient training of giant neural networks using pipeline parallelism. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, Advances in Neural Information Processing Systems 32, pages 103–112. Curran Associates, Inc.

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda B. Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's multilingual neural machine translation system: Enabling zero-shot translation. CoRR, abs/1611.04558.

Tom Kocmi and Ondřej Bojar. 2017. Curriculum Learning and Minibatch Bucketing in Neural Machine Translation. In Proceedings of Recent Advances in NLP (RANLP 2017).

Tom Kocmi and Ondřej Bojar. 2018. Trivial Transfer Learning for Low-Resource Neural Machine Translation. In Proceedings of the Third Conference on Machine Translation, Volume 1: Research Papers, volume 1, pages 244–252, Stroudsburg, PA, USA. Association for Computational Linguistics, Association for Computational Linguistics.

Toan Q. Nguyen and David Chiang. 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. Asian Federation of Natural Language Processing.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc V. Le, Geoffrey E. Hinton, and Jeff Dean. 2017.