Machine Translation 2: Multilingual MT

Ondřej Bojar
bojar@ufal.mff.cuni.cz
Institute of Formal and Applied Linguistics
Faculty of Mathematics and Physics
Charles University, Prague

Outline

- Motivation for multilingual MT.
- Interesting configurations.
- Dedicated architectures vs. simple data mixing.
- Interlingua?

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.

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.

Strategies for NMT

- Simple data mixing.
- Dedicated architectures.

Simple Data Mixing

. . . 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).

Trivial Transfer Learning

- Multi-lingual resources can be also used for "transfer learning".
  - i.e. improving one task based on another.
- Previous works (Zoph et al., 2016; Nguyen and Chiang, 2017) target one common language (English).

Tom Kocmi's recent work:

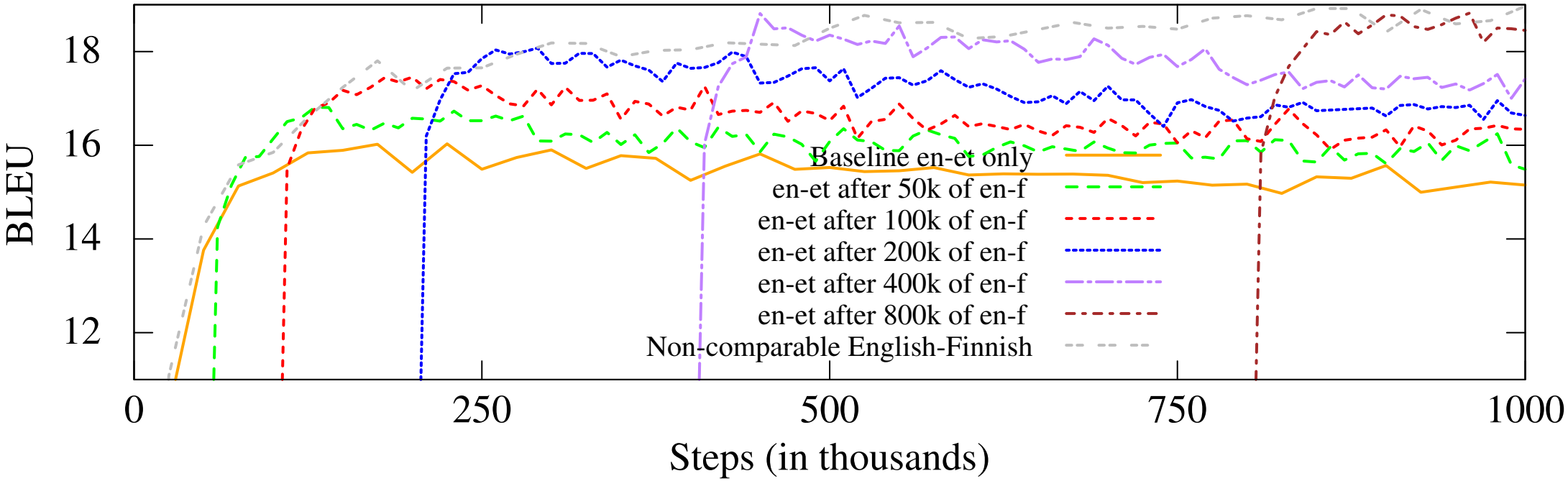
- Train on one pair ("parent"), switch corpus to another ("child").
- The only requirement: joint subword units across all langs.

Trivial Transfer Learning

Parent/Child	ENET	ETEN	ENSK
baseline	17.03	21.74	16.13
parent ENFI	19.74	22.75	-
parent FIEN	18.19	24.18	-
parent ENCS	20.41	-	17.75
parent ENRU	20.09	23.12	-
parent ENET	-	22.04	-
parent ETEN	17.46	-	-

- Uncased BLEU scores for various high-resource parents.
- Each column corresponds to a different child lang. pair.
  - (Scores comparable within columns only.)
- Underlined when neither source nor target side is shared.

Trivial Transfer Learning



- The better the parent is, the better the child.

Why it Helps? Not Very Clear.

	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.)

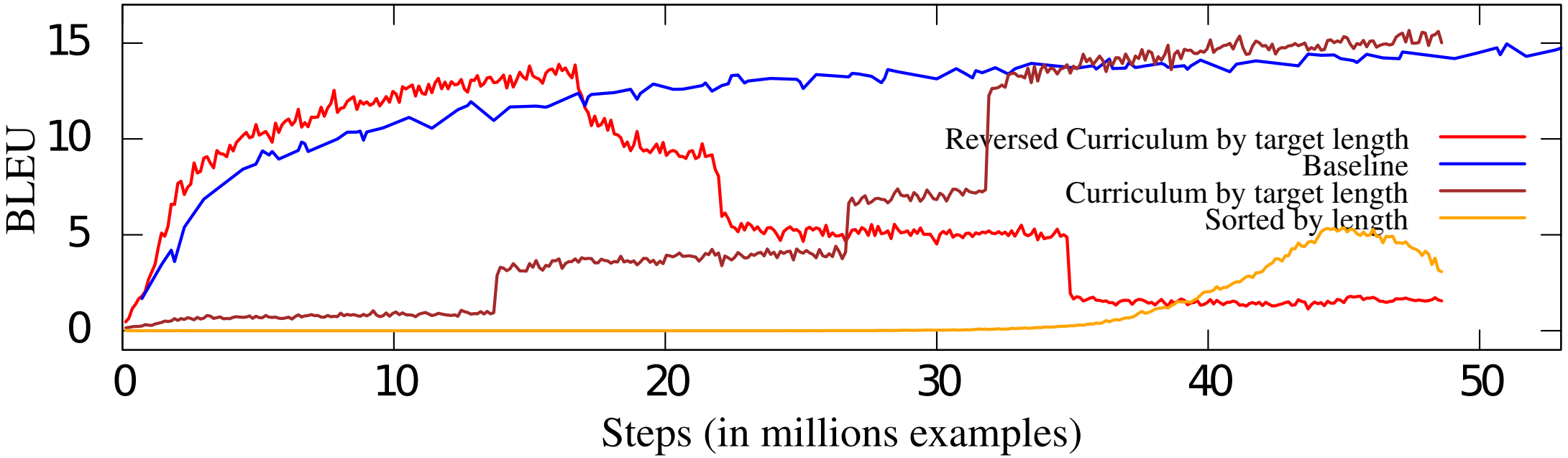
- Child models produce longer outputs ⇒ lower brevity penalty.
- But n-gram precisions also better.

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.

Interlude: Catastrophic Forgetting

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



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**

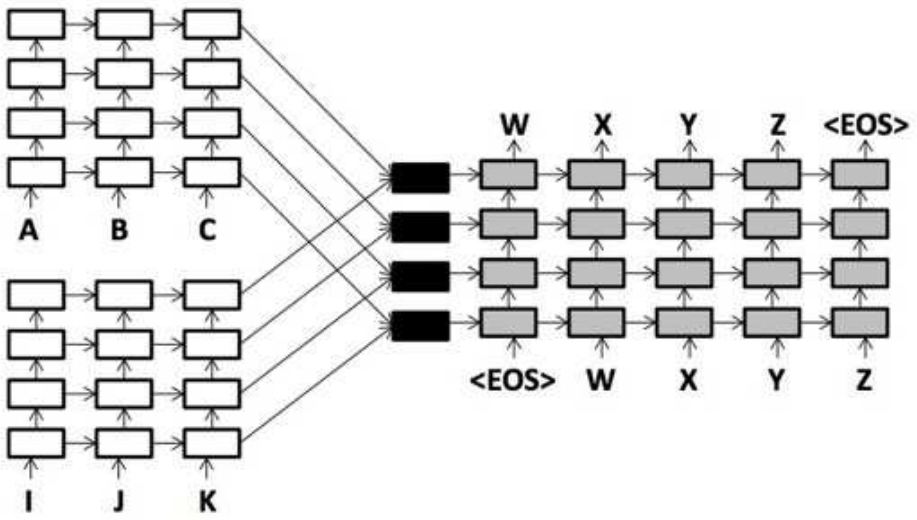
Multi-source translation

- 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?

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

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

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

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

Multi-way MT

Firat et al. 2016 (two papers)

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

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

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)

- 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)

		Dir	Fr (39m) → En	Fr (39m) En →	Cs (12m) → En	Cs (12m) En →	De (4.2m) → En	De (4.2m) En →	Ru (2.3m) → En	Ru (2.3m) En →	Fi (2m) → En	Fi (2m) 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)

		Dir	Fr (39m) → En	Fr (39m) En →	Cs (12m) → En	Cs (12m) En →	De (4.2m) → En	De (4.2m) En →	Ru (2.3m) → En	Ru (2.3m) En →	Fi (2m) → En	Fi (2m) 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)	√		20.64	20.4

Table 4: Zero-resource translation from Spanish (Es) to French (Fr) *without* finetuning. When pivot is √, 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)

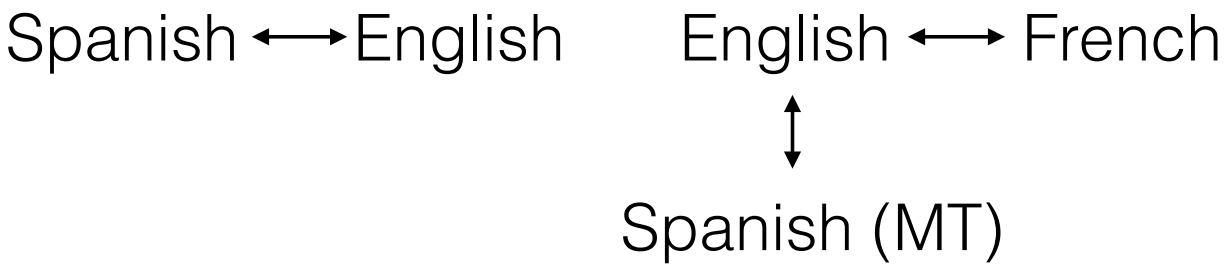
- *Finetuning*: what if we use a small amount of parallel data in this setting?
- Q: Where would we get this data? **Backtranslation**

Spanish $\longleftrightarrow$ English English $\longleftrightarrow$ French

Zero-shot MT

Firat et al. 2016 (two papers)

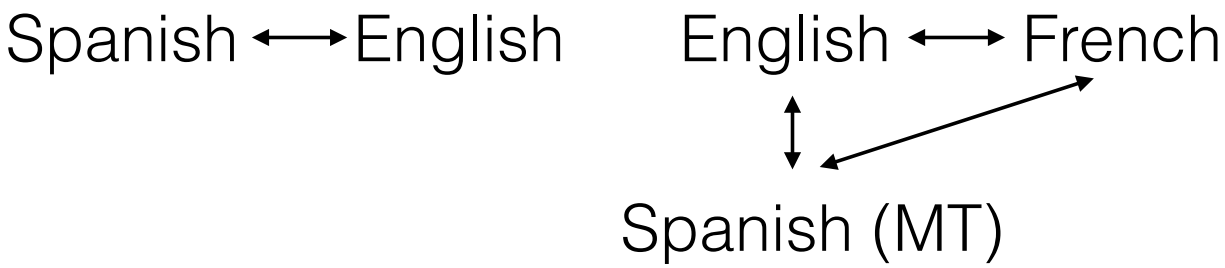
- *Finetuning*: what if we use a small amount of parallel data in this setting?
- Q: Where would we get this data? **Backtranslation**



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

Zero-shot MT

Johnson et al. 2016 (Google)

- Do we really need *N* encoders and *N* decoders?
- Can we just learn a single function parameterized by the desired output language?
  - Implementation: add a token indicating desired output language to input.
- Why is this a nice solution (for Google)?

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

Zero-shot MT

Johnson et al. 2016 (Google)

- Incremental training: add a small amount of (true) parallel data in the language pair of interest.

Table 5: Portuguese→Spanish BLEU scores using various models.

	Model	BLEU
(a)	PBMT bridged	28.99
(b)	NMT bridged	30.91
(c)	NMT Pt→Es	31.50
(d)	Model 1 (Pt→En, En→Es)	21.62
(e)	Model 2 (En↔{Es, Pt})	24.75
(f)	Model 2 + incremental training	31.77

Zero-shot MT

Johnson et al. 2016 (Google)

Table 6: BLEU scores for English↔{Belarusian, Russian, Ukrainian} models.

	Zero-Shot	From-Scratch	Incremental
English→Belarusian	16.85	17.03	16.99
English→Russian	22.21	22.03	21.92
English→Ukrainian	18.16	17.75	18.27
Belarusian→English	25.44	24.72	25.54
Russian→English	28.36	27.90	28.46
Ukrainian→English	28.60	28.51	28.58
Belarusian→Russian	56.53	82.50	78.63
Russian→Belarusian	58.75	72.06	70.01
Russian→Ukrainian	21.92	25.75	25.34
Ukrainian→Russian	16.73	30.53	29.92

trained on parallel data

Zero-shot MT

Johnson et al. 2016 (Google)

Table 6: BLEU scores for English↔{Belarusian, Russian, Ukrainian} models.

	Zero-Shot	From-Scratch	Incremental
English→Belarusian	16.85	17.03	16.99
English→Russian	22.21	22.03	21.92
English→Ukrainian	18.16	17.75	18.27
Belarusian→English	25.44	24.72	25.54
Russian→English	28.36	27.90	28.46
Ukrainian→English	28.60	28.51	28.58
Belarusian→Russian	56.53	82.50	78.63
Russian→Belarusian	58.75	72.06	70.01
Russian→Ukrainian	21.92	25.75	25.34
Ukrainian→Russian	16.73	30.53	29.92

zero-shot + small parallel data

Zero-shot MT

Johnson et al. 2016 (Google)

Table 6: BLEU scores for English↔{Belarusian, Russian, Ukrainian} models.

	Zero-Shot	From-Scratch	Incremental
English→Belarusian	16.85	17.03	16.99
English→Russian	22.21	22.03	21.92
English→Ukrainian	18.16	17.75	18.27
Belarusian→English	25.44	24.72	25.54
Russian→English	28.36	27.90	28.46
Ukrainian→English	28.60	28.51	28.58
Belarusian→Russian	56.53	82.50	78.63
Russian→Belarusian	58.75	72.06	70.01
Russian→Ukrainian	21.92	25.75	25.34
Ukrainian→Russian	16.73	30.53	29.92

actual zero-shot experiment

Zero-shot MT

Johnson et al. 2016 (Google)

code-switching in the input language:

Japanese: 私は東京大学の学生です。 → I am a student at Tokyo University.

Korean: 나는 도쿄 대학의 학생입니다. → I am a student at Tokyo University.

Mixed Japanese/Korean: 私は東京大学학생입니다. → I am a student of Tokyo University.

code-switching in the output language:

Spanish/Portuguese:	Here the other guinea-pig cheered, and was suppressed.
$w_{pt} = 0.00$	Aquí el otro conejillo de indias animó, y fue suprimido.
$w_{pt} = 0.30$	Aquí el otro conejillo de indias animó, y fue suprimido.
$w_{pt} = 0.40$	Aquí, o outro porquinho-da-índia alegrou, e foi suprimido.
$w_{pt} = 0.42$	Aqui o outro porquinho-da-índia alegrou, e foi suprimido.
$w_{pt} = 0.70$	Aqui o outro porquinho-da-índia alegrou, e foi suprimido.
$w_{pt} = 0.80$	Aqui a outra cobaia animou, e foi suprimida.
$w_{pt} = 1.00$	Aqui a outra cobaia animou, e foi suprimida.

Zero-shot MT

Johnson et al. 2016 (Google)

Portuguese informant: “we decided it's impossible to judge the correctness of the translation without context (but it's likely wrong). After finding the context (Alice in Wonderland) we can conclude it's wrong.”

code-switching in the output language:

Spanish/Portuguese:	Here the other guinea-pig cheered, and was suppressed.
$w_{pt} = 0.00$	Aquí el otro conejillo de indias animó, y fue suprimido.
$w_{pt} = 0.30$	Aquí el otro conejillo de indias animó, y fue suprimido.
$w_{pt} = 0.40$	Aquí, o outro porquinho-da-índia alegrou, e foi suprimido.
$w_{pt} = 0.42$	Aqui o outro porquinho-da-índia alegrou, e foi suprimido.
$w_{pt} = 0.70$	Aqui o outro porquinho-da-índia alegrou, e foi suprimido.
$w_{pt} = 0.80$	Aqui a outra cobaia animou, e foi suprimida.
$w_{pt} = 1.00$	Aqui a outra cobaia animou, e foi suprimida.

Interlingua?

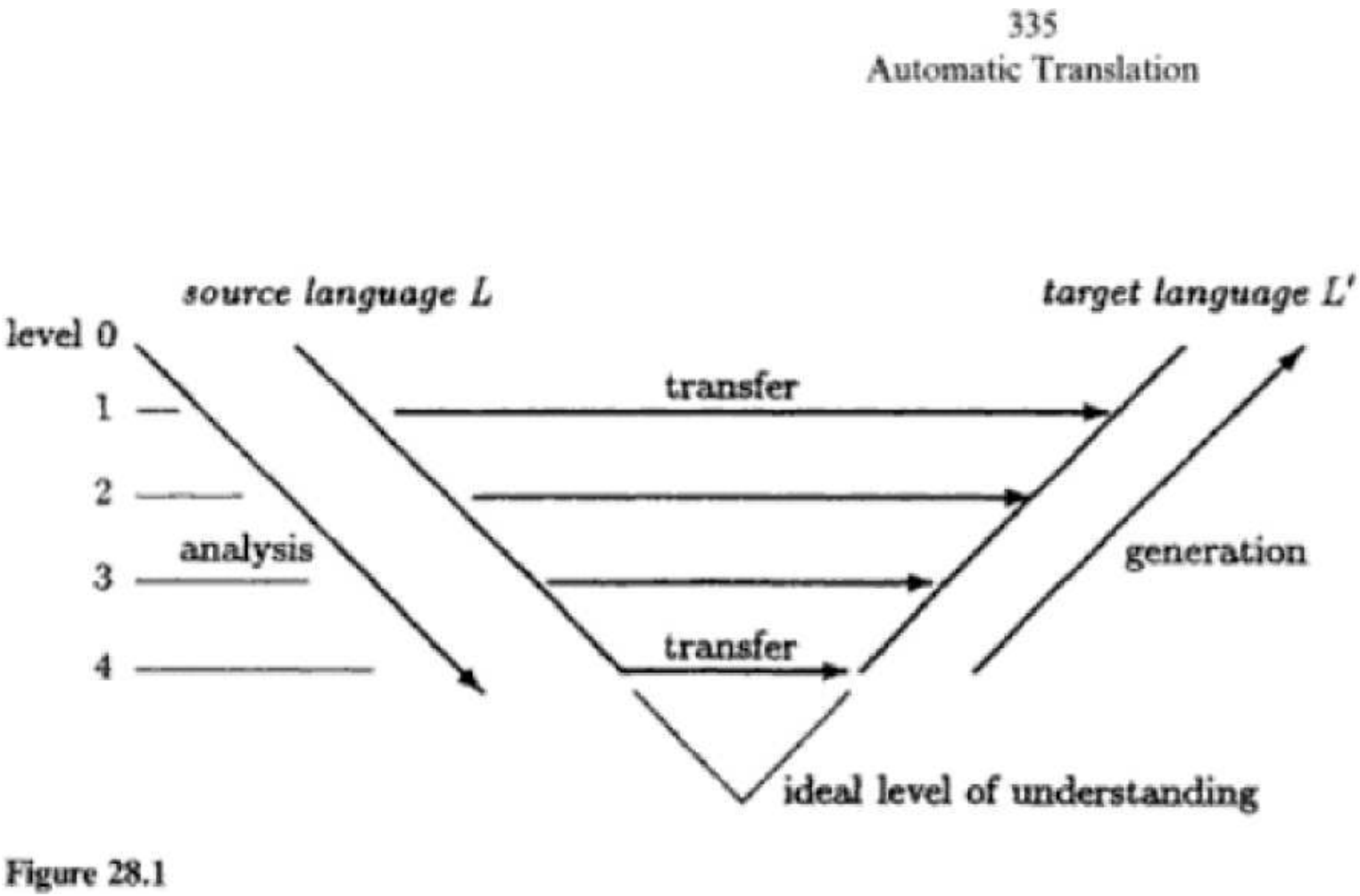


Figure 28.1

From Vauquois (1968), reproduced by Adam Lopez.

Interlingua?

- Theoretically, a very inspiring concept.
- Need for $2N$ instead of n^2 systems.
- Sceptical view:
  - Need to capture all distinctions in word meanings:

    `https://en.wikipedia.org/wiki/Eskimo_words_for_snow`
  - Text form underspecifies the meaning, formally captured content underspecifies the form (Lampert, 2001).
  - Interannotator agreement decreases as we proceed along layers of linguistic analysis (Dorr et al., 2010).

Interlingua?

- Optimistic/wishful view:
  - Molto-Project (EU FP7, 2011-2013), among others:
    `http://www.molto-project.eu/`

    Isn't interlingua an unrealistic dream? Yes, it is, if we want to have a universal interlingua working for everything. This is why we don't believe we can ever translate newspapers with MOLTO techniques. However, domain-specific interlinguas have proved quite feasible. Notice that this move is similar to what has happened in ontologies: they have moved from universal ontologies to domain ontologies.

Google Interlingua (Johnson et al., 2016)

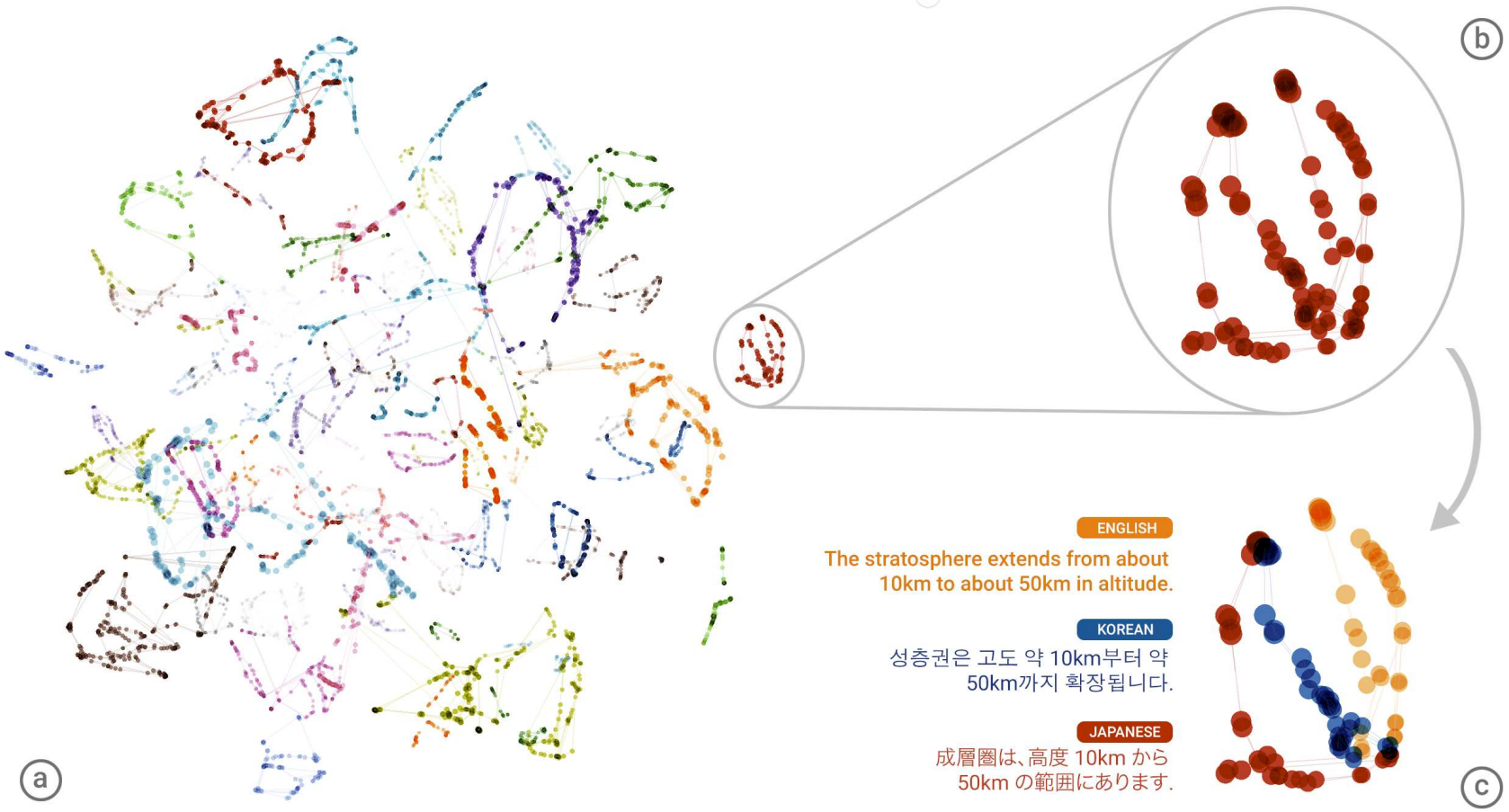


Figure 2: A t-SNE projection of the embedding of 74 semantically identical sentences translated across all 6 possible directions, yielding a total of 9,978 steps (dots in the image), from the model trained on English↔ Japanese and English↔ Korean examples. (a) A bird's-eye view of the embedding, coloring by the index of the semantic sentence. Well-def ned clusters each having a single color are apparent. (b) A zoomed in view of one of the clusters with the same coloring. All of the sentences within this cluster are translations of "The stratosphere extends from about 10km to about 50km in altitude." (c) The same cluster colored by source language. All three source languages can be seen within this cluster.

Summary

- Machine translation is multilingual from the beginning.
- Very big improvements in low-resource conditions.
  - With as simple techniques as continued training.
- Multi-way/multi-source and/or multi-target very promising.

  - Gains very likely but still rather unclear.
- Good progress towards zero-shot or fully unsupervised MT.

References

Bonnie j. Dorr, Rebecca j. Passonneau, David Farwell, Rebecca Green, Nizar Habash, Stephen Helmreich, Eduard Hovy, Lori Levin, Keith j. Miller, Teruko Mitamura, Owen Rambow, and Advaith Siddharthan. 2010. Interlingual annotation of parallel text corpora: A new framework for annotation and evaluation. Nat. Lang. Eng., 16(3):197–243.

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

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).

Andrew Lampert. 2001. Interlingua in Machine Translation. http://sgi.nu/nlp/content/pdf/InterlinguaInMachineTra September.

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.

Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. 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, November. Association for Computational Linguistics.