Zero-Shot and Unsupervised Machine Translation

Jonah Philion

April 1, 2018

Overview

1. Traditional Neural Machine Translation
2. Zero-Shot Translation
3. Unsupervised Neural Machine Translation (Artetxe, 2018)

Traditional Neural Machine Translation

Traditional Neural Machine Translation with Attention

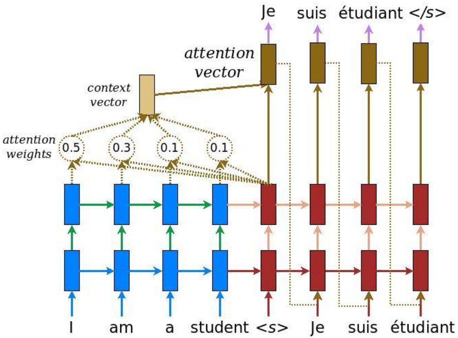


Figure: An NMT model with attention (Bahdanau, 2016).

Zero-Shot Translation

Motivation for Google Zero-Shot Translation (Johnson, 2017)

Problem

- Each pair of languages requires an encoder-decoder pair $\Rightarrow$ space is $O(n^2)$ in number of languages n.
- Google would like to support upwards of 100 languages with sometimes few to no sentence-to-sentence pairs.

Motivation for Google Zero-Shot Translation (Johnson, 2017)

Problem

- Each pair of languages requires an encoder-decoder pair ⇒ space is $O(n^2)$ in number of languages n.
- Google would like to support upwards of 100 languages with sometimes few to no sentence-to-sentence pairs.

Solution

- Universal decoder and encoder for all language pairs.
- Add an artificial token to the input sequence to indicate the required target language.

```
Hello, how are you? -> Hola, ¿cómo estás?

<2es> Hello, how are you? -> Hola, ¿cómo estás?
```

Zero-Shot Architecture

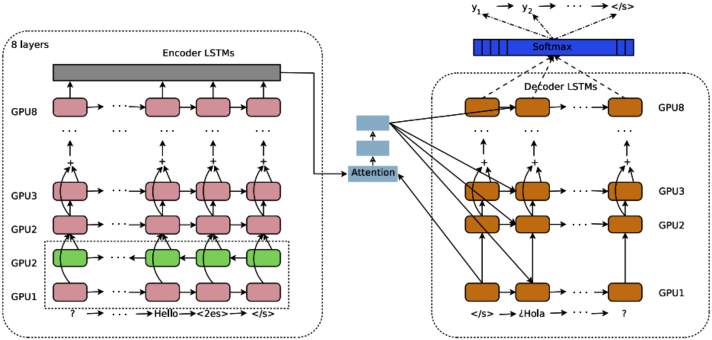
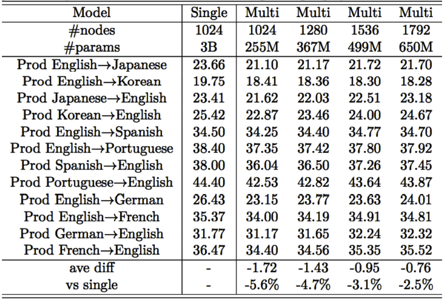


Figure: Google Zero-Shot Architecture (Johnson, 2017). Source sentence is reversed and prepended to the target language codon.

Results

- On the many-to-many task, the model underperforms production translation models by 2.5%.

Table 4: Large-scale experiments: BLEU scores for single language pair and multilingual models.



Model	Single	Multi	Multi	Multi	Multi
#nodes	1024	1024	1280	1536	1792
#params	3B	255M	367M	499M	650M
Prod English→Japanese	23.66	21.10	21.17	21.72	21.70
Prod English→Korean	19.75	18.41	18.36	18.30	18.28
Prod Japanese→English	23.41	21.62	22.03	22.51	23.18
Prod Korean→English	25.42	22.87	23.46	24.00	24.67
Prod English→Spanish	34.50	34.25	34.40	34.77	34.70
Prod English→Portuguese	38.40	37.35	37.42	37.80	37.92
Prod Spanish→English	38.00	36.04	36.50	37.26	37.45
Prod Portuguese→English	44.40	42.53	42.82	43.64	43.87
Prod English→German	26.43	23.15	23.77	23.63	24.01
Prod English→French	35.37	34.00	34.19	34.91	34.81
Prod German→English	31.77	31.17	31.65	32.24	32.32
Prod French→English	36.47	34.40	34.56	35.35	35.52
ave diff	-	-1.72	-1.43	-0.95	-0.76
vs single	-	-5.6%	-4.7%	-3.1%	-2.5%

Figure: When trained on all pairs of three languages, the model achieves comparable BLEU scores. Since the simplest model has 255M instead of 3B parameters, this result is astounding.

Zero-Shot

The model can translate between languages for which there was no parallel data during training.

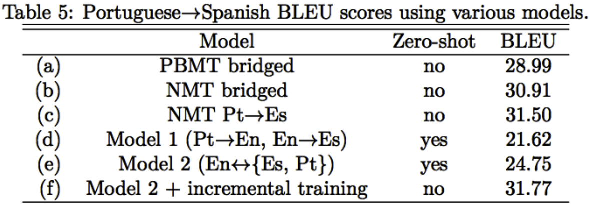


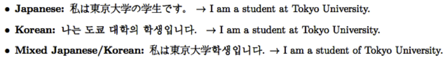
Table 5: Portuguese→Spanish BLEU scores using various models.

	Model	Zero-shot	BLEU
(a)	PBMT bridged	no	28.99
(b)	NMT bridged	no	30.91
(c)	NMT Pt→Es	no	31.50
(d)	Model 1 (Pt→En, En→Es)	yes	21.62
(e)	Model 2 (En↔{Es, Pt})	yes	24.75
(f)	Model 2 + incremental training	no	31.77

Figure: PBMT is a phrase based translation system. NMT bridged is translation from Portuguese to Spanish going through English. NMT Pt→Es is a standard NMT model with attention trained on all available Portuguese to Spanish sentences.

Interlingua

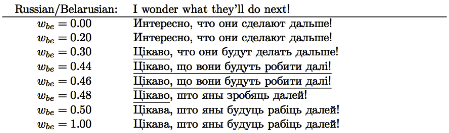
Source Language Code-Switching What if we mix languages in the source sentence?



- **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.

Figure: The model handles mixed input language no problem.

Weighted Target Language Selection What if we feed linear combinations of target codons?



Russian/Belarusian:	I wonder what they'll do next!
$w_{be} = 0.00$	Интересно, что они сделают дальше!
$w_{be} = 0.20$	Интересно, что они сделают дальше!
$w_{be} = 0.30$	Цікаво, что они будут делать дальше!
$w_{be} = 0.44$	Цікаво, що вони будуть робити далі!
$w_{be} = 0.46$	Цікаво, що вони будуть робити далі!
$w_{be} = 0.48$	Цікаво, што яны зробяць далей!
$w_{be} = 0.50$	Цікава, што яны будуць рабіць далей!
$w_{be} = 1.00$	Цікава, што яны будуць рабіць далей!

Figure: The model translates into Ukrainian in the process of translating to Russian and Belarusian.

Unsupervised Neural Machine Translation (Artetxe, 2018)

UNMT Motivation

- **Question** How well can we translate between language l_1 and l_2 given a corpus of text X_1 of language l_1 and a corpus of text X_2 of language l_2?

$\rightarrow$ Strong baseline for supervised NMT

$\rightarrow$ Translation in low-resource contexts

UNMT Motivation

- **Question** How well can we translate between language l_1 and l_2 given a corpus of text X_1 of language l_1 and a corpus of text X_2 of language l_2?

  → Strong baseline for supervised NMT

  → Translation in low-resource contexts

- **Solution** Train universal encoder and language-specific decoders to denoise and backtranslate.

Model

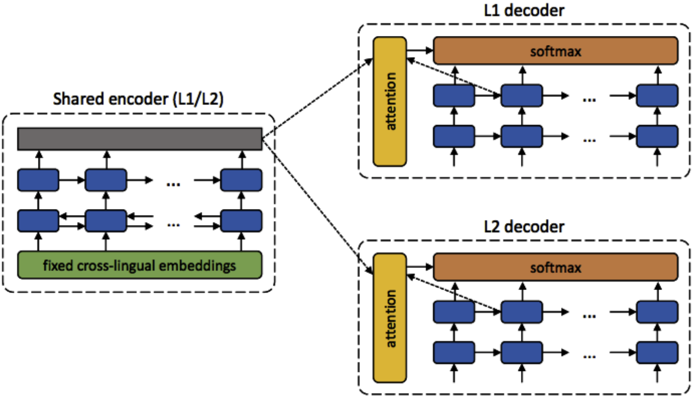


Figure: At test time, a sentence *src* from either l_1 or l_2 is fed to the encoder, and the target sentence decoder is used to predict a translation. At training time, the decoder used is the decoder associated with the language of *src*.

“Differences” from Standard NMT

- **2-step training** Cross-lingual word embeddings are determined in an initial training step (Johnson, 2017). These embeddings are then frozen before training the encoder-decoder system.
- $\mathbf{L}_{denoise} + \mathbf{L}_{backtranslate}$ Training the ecoder-decoder consists of a loss associated with the ability to back-translate and a loss associated with the ability to denoise.

$L_{denoise}$

Given a sequence of length N, make $N/2$ transpositions iteratively. The model is trained with the standard word-by-word loss. Let $C(x)$ return our noise model applied to sentence x.

$$L_{denoise} = \mathbb{E}_{x \sim X_l, \hat{x} \sim d(e(C(x)),l)} \Delta(\hat{x}, x)$$

An example denoising case

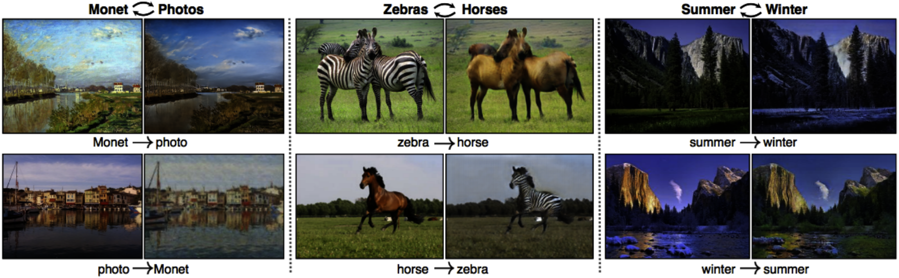
```
 this is example an sentence denoising for presentation my

-> this is an example denoising sentence for my presentation
```

$L_{backtranslate}$

Given an input sentence in one language, use the system in inference mode with greedy decoding to translate it to the other language. In this way, we obtain pseudo-parallel sentence pairs. Let $x' = d(e(x), l')$.

$$L_{backtranslate} = \mathbb{E}_{x \sim X_l, \hat{x} \sim d(e(x'), l)} \Delta(\hat{x}, x)$$

In the semi-supervised model, we also mix in parallel sentence examples to this loss. The above loss has had some success in unsupervised image-to-image translation (Zhu, 2018).



Results

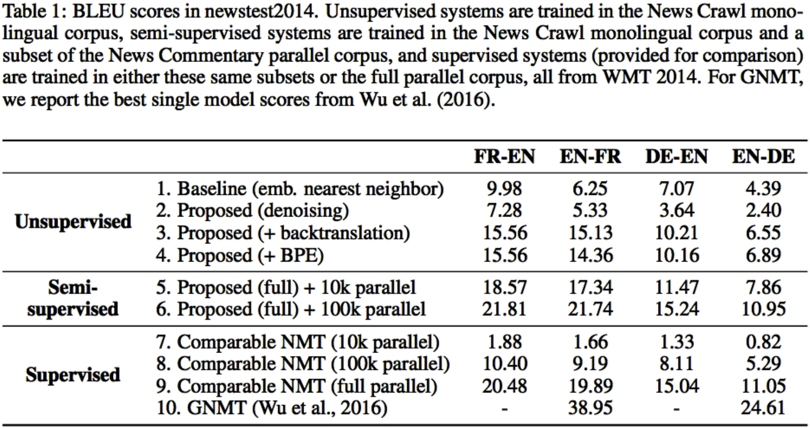


Table 1: BLEU scores in newstest2014. Unsupervised systems are trained in the News Crawl monolingual corpus, semi-supervised systems are trained in the News Crawl monolingual corpus and a subset of the News Commentary parallel corpus, and supervised systems (provided for comparison) are trained in either these same subsets or the full parallel corpus, all from WMT 2014. For GNMT, we report the best single model scores from Wu et al. (2016).

		FR-EN	**EN-FR**	**DE-EN**	**EN-DE**
Unsupervised	1. Baseline (emb. nearest neighbor)	9.98	6.25	7.07	4.39
	2. Proposed (denoising)	7.28	5.33	3.64	2.40
	3. Proposed (+ backtranslation)	15.56	15.13	10.21	6.55
	4. Proposed (+ BPE)	15.56	14.36	10.16	6.89
Semi-supervised	5. Proposed (full) + 10k parallel	18.57	17.34	11.47	7.86
	6. Proposed (full) + 100k parallel	21.81	21.74	15.24	10.95
Supervised	7. Comparable NMT (10k parallel)	1.88	1.66	1.33	0.82
	8. Comparable NMT (100k parallel)	10.40	9.19	8.11	5.29
	9. Comparable NMT (full parallel)	20.48	19.89	15.04	11.05
	10. GNMT (Wu et al., 2016)	-	38.95	-	24.61

Figure: The model performs “surprisingly” well. However, there is discussion of whether this model is actually interesting given the lack of success of models 7., 8., and 9. in a fully supervised setting.

The End. Questions?