

# CRISS

Xavier Garcia

July 2019

## 1 Background

## 2 Multilingual MT

The goal of multilingual machine translation is to build a universal machine translation model  $f$  which when given a sentence  $x$  and language  $l$ ,  $f(x, l)$  is the translation of  $x$  in language  $l$ . In theory, not very hard: if you find parallel data of every language with English, then you can translate between any pair of languages either indirectly (via pivoting) or even directly (via zero-shot). In practice, it turns you generally require both high-quality in-domain data as well as large amounts of it. This can be quite difficult for low-resource languages such as Swahili, Yoruba, Afrikaans, etc. Only one resource left: monolingual data.

**Unsupervised MT** Given the limitations of parallel data, people began to think of ways of incorporating monolingual data as well as the extreme case where only monolingual data is available. In general, it was a shitshow until 2019 when people started using pre-training. Most efficient approach came from MASS, which generalized the masked language model to seq2seq tasks. However, MASS didn't explore the confluence between multilingual MT and unsupervised MT.

**M-BART** Facebook's solution was to train a large multilingual model with *only* monolingual data. Instead of MASS, they developed their own objective (BART) and extended it to the multilingual setting. The M-BART objective consists of a denoising autoencoder where the noise function is some combination of masking and shuffling. They fine-tuned the model with either back-translation for a lot of the pairs. They also found (concurrently with me) that leveraging parallel data for auxiliary language pairs also improved performance into English, almost suggesting that you did not require back-translation, but the performance in the reverse direction was never good. Moreover, for m-BART, they had initial issues fine-tuning due to the language output, signaling some issues in the decoding.

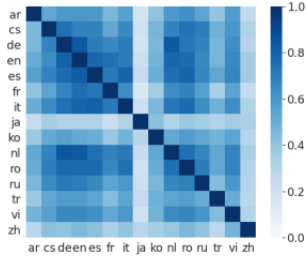


Figure 1: Sentence retrieval accuracy using encoder outputs of mBART

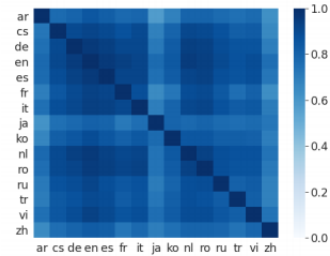


Figure 2: Sentence retrieval accuracy using encoder outputs of mBART finetuned on Japanese-English parallel data

### 3 Cross-Lingual Self-Retrieval

**The dream of language-agnostic representations** Why is there this asymmetry and how does parallel data help? The holy grail of multilingual MT is the idea of *language-agnostic representations*: if you can disentangle meaning from language in the encoder’s output, then the decoder has no choice but to spit out the correct translation from being trained as an autoencoder. You don’t really need disentanglement (it’s a myth anyways), as long as the decoder can override the language information. But is this really true? The authors considered a task where we have multi-way parallel data of English with 57 other languages. Given a pre-train M-BART, can we identify translation pairs based on the encoder’s representations?<sup>1</sup> Refer to Figure 1. Key point: 57% average accuracy. Furthermore, accuracy jumps to 84% when fine-tune with only a single language pair.

**Forsake the decoder** We can conclude from the previous experiment that at the very least the encoder is doing its job and that (given the right language signal) it can encode some notion of language-independent meaning. This raises the question: Do we even need the decoder? Given the successful parallel text mining capabilities, the paper then introduces a simple approach for unsupervised MT: (1) take m-BART and two large datasets of monolingual data (2) use m-BART to mine parallel data and finetune the model on this data. (3) Use this new model to mine higher quality data and re-train a new model from it; (4) Rinse and repeat.

**UNMT Performance** With this approach, they compare against traditional baselines which don’t leverage parallel data. See Figure 2. We see that it

<sup>1</sup>They took the L2-normalized average-pooled embeddings.

Direction	en-de	de-en	en-fr	fr-en	en-ne	ne-en	en-ro	ro-en	en-si	si-en
CMLM [29]	27.9	35.5	34.9	34.8	-	-	34.7	33.6	-	-
XLM [5]	27.0	34.3	33.4	33.0	0.1	0.5	33.3	31.8	0.1	0.1
MASS [32]	28.3	35.2	37.5	34.9	-	-	35.2	33.1	-	-
D2GPO [18]	28.4	35.6	37.9	34.9	-	-	<b>36.3</b>	33.4	-	-
mBART [19]	29.8	34	-	-	4.4	10.0	35.0	30.5	3.9	8.2
CRISS Iter 1	21.6	28.0	27.0	29.0	2.6	6.7	24.9	27.9	1.9	6
CRISS Iter 2	30.8	36.6	37.3	36.2	4.2	12.0	34.1	36.5	5.2	12.9
CRISS Iter 3	<b>32.1</b>	<b>37.1</b>	<b>38.3</b>	<b>36.3</b>	<b>5.5</b>	<b>14.4</b>	35.1	<b>37.6</b>	<b>6.0</b>	<b>13.6</b>

Table 1: Unsupervised machine translation. CRISS outperforms others in 9 out of 10 directions

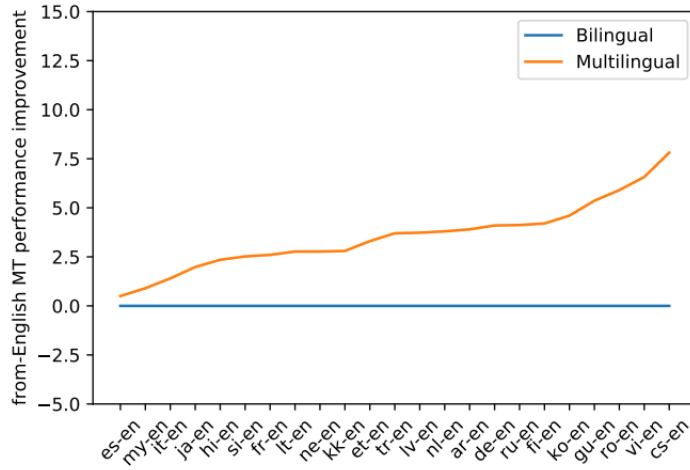


Figure 4: Bilingual versus Multilingual: x-En

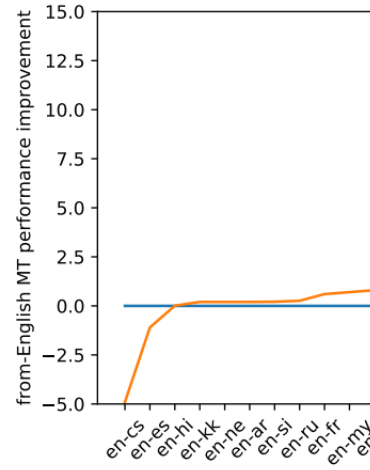


Figure 5: Bilingual versus Multilingual: y-En

actually outperforms all the baselines by a significant amount. In fact, they can even contest early-versions of M-UNMT 1, despite not leveraging parallel data.

**Effects of multilinguality beyond text mining** It is known that multilinguality can help for low-resource language. Using that logic, it should be true that if you mined multiple parallel directions, then this should overall improve the model’s performance in finetuning. See figure 3. They test finetuning on 24 language pairs at once, and compare with a bilingual baseline. There are two patterns emerging: (1) there is always positive transfer in to-English direction. (2) Mixed results in the from-English direction. Given our initial question regarding the universal machine translation model, one might ask, ”how many

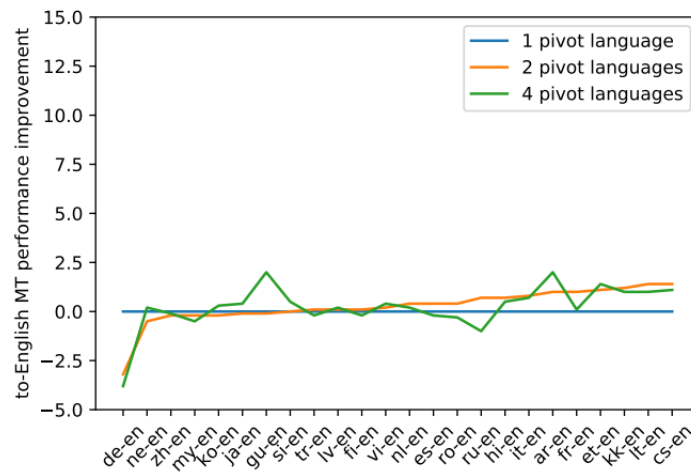


Figure 6: Pivot languages ablation: x-En MT

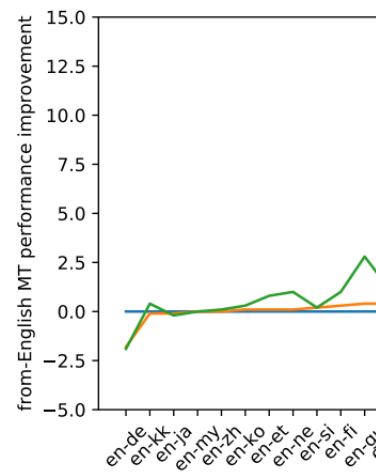


Figure 7: Pivot language

language pairs do you need?" See figure 4.