Zero-Shot Multilingual Machine Translation

Tina Buzanis
kristina_buzanis@student.uml.edul

Aria Kieras
aria_kieras@student.uml.edu

Abstract

In this work, we compare bilingual machine translation systems - English to Russian and English to French - with a multilingual system that has been trained on both French↔English and Russian↔English datasets. This is done with fine-tuned variants of Google's mT5-base. Our results demonstrate that translation with an English pivot outperforms direct zero-shot translation with a multilingual model. We additionally compare it with a fully-supervised model trained on Opensubtitles dataset. The code is available on GitHub[1].

1 Introduction

With the rise of the prevalence and importance of Neural Machine Translation (NMT) models, their scale has increased as well. NMT models require massive amounts of data in order to achieve good performance. However, a large portion of this data is English-centric. With the smaller size and relative scarcity of the datasets for lesser-spoken languages, models can be difficult to fine-tune (and therefor difficult to use) for members of these language communities.

A potential solution to this problem is zero-shot multilingual translation (Figure 1). With this approach, massive multilingual language models, such as mT5, have the potential to bridge the gap in accessibility, thus also improving the quality of machine translation systems.

In this work, we compare bilingual machine translation systems - English to Russian and English to French - with a multilingual system that has been trained on both French→English and Russian→English datasets. This is done with fine-tuned variants of Google's mT5-base.

Our results demonstrate that translation with an English pivot outperforms direct zero-shot translation with a multilingual model. We additionally compare it with a fully-supervised model trained on the OpenSubtites CC-Matrix dataset (Lison and Tiedemann, 2016).

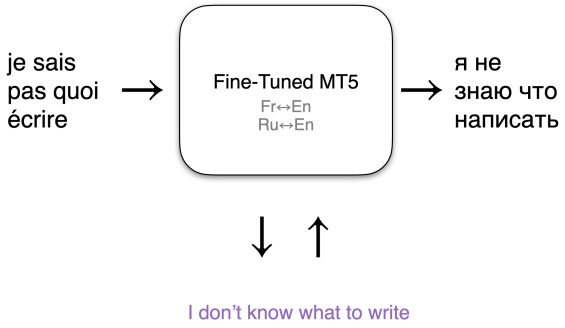

Figure 1: Direct translation vs translation via a pivot language (English).

2 Related Work

Machine Translation As stated earlier, NMT models require a large amount if data to function. Aside from this, they are quite computationally expensive, and with large amounts of data combined with large models, this can cause the NMT to crash (Wu et al., 2016). This is where Googles NMT (GNMT) comes in, in an attempt to address many of the issues with NMT's. Using 8 encoder and 8 decoder layers on a deep LSTM network, GNMT was able to produce competative state of the art results on the wmt14 English-to-French and German-to-English datasets, reducing translation errors by about 60% (Wu et al., 2016).

Zero-Shot Machine Translation For cross-lingual NLP the most widely used approach is to use multilingual pre-trained transformers as features in neural network models. However prior research has indicated that representations learned in context perform much better (McCann et al., 2017).

[1] github.com/tinabuzanis/machine-translation

It is the aim to use representations obtained from a multilingual neural machine translation model, in order to enable cross lingual transfer learning on downstream NLP tasks (Eriguchi et al., 2018). An encoder-classifier model was implemented, where the encoder converts an input sequence to a set of vectors and the classifier predicts the class labels given the encoding of the input sequence (Eriguchi et al., 2018). This method was used to test zero shot classification for the French language, which was assumed to have been unseen by the model, which was used in order to see how well the model could generalize to a previously unseen language. In order to compare the results from the zero shot testing, bridging was utilized; "translating a French test text to English and then applying the English classifier on the translated text" (Eriguchi et al., 2018). It was seen that the zero shot classification was 2% in range of the bridging accuracy when used with the proposed model.

Pre-Trained Multilingual Models Multilingual-BERT (mBERT) was trained on the concatenation of Wikipedia to 104 languages, and has demonstrated that masked language models pre-trained on multilingual exhibit cross-lingual capabilities. Recently, it has been shown to excel in the task of zero-shot translation (Gonen et al., 2020). mT5 is a version of T5 that was pre-trained on multilingual mC4 dataset. We hope that pre-trained multilingual models can help with zero-shot transfer and improve upon daisy-chain baseline.

3 Methods

mT5, a multilingual variant of the original T5 (Text to Text Transfer Transformer), was pre-traned on a multilingual variant of the C4 dataset, called mC4. This dataset encompasses samples from 101 languages, drawn from the public Common Crawl web scrape. The architecture of mT5 is similar to that of T5, however mT5 uses GeGLU nonlinearities, and was pre-trained on unlabeled data only with no dropout.

In our case, we fine-tuned mT5 using two setups: bidirectional translation, and multilingual translation. In order to facilitate this, we prepended a prompt (for example "translate from French to Russian:") to the beginning of each sentence from the given source language. We do this for both multilingual and bilingual models.

We then fine-tuned two bidirectional models: French↔English, and Russian↔English, along with a multilingual model, which was trained on both the French-English and Russian-English datasets.

	FR-EN	RU-EN
Size	269.84MB	269.84MB
Sentence count	500,000	500,000

Table 1: The size of the datasets used in the study.

4 Experiments

One of the inital challenges we faced with training this model was with memory constraints. The mT5 tokenizer is large, with a default vocabulary size of 250,112, and 2.2GB. To mitigate this issue, we explored several methods related to optimization. We began by swapping our AdamW optimizer for Adafactor, which worked, however the results appeared to be unstable. Next, we tried a combination of Facebook Research's 8-bit Adam, alongside Huggingface's Accelerate. Finally, we removed Accelerate, and settled on PyTorch's Adam optimizer.

Lastly, we experimented with adapting mT5 for the specific languages we were working with. Beginning with the full-size model, and with English, French and Russian datasets from the Leipzig corpora collection, we discovered that the top 20,000 tokens constitute roughly 96 percent of the corpus for each language. We thus decided to reconstruct the vocabulary as follows: the top 1000 tokens of our original tokenizer, the top 10000 tokens of the English vocabulary, the top 20000 tokens of both the French and Russian vocabularies, along with mT5's special tokens. This gave us an updated vocabulary size of 51,100, about twenty percent of the original size. We then updated both the model and tokenizer accordingly. Unfortunately, due to time and cost constraints, we were unable to complete this experiment, however the initial results appeared promising, and perhaps warrant further investigation.

Datasets The datasets in use for this project were derived from the WMT14 - the FR-EN and RU-EN subsets (Bojar et al., 2014).The size of these datasets can be seen in table 1. Both of these datasets consisted of sentence pairs in order to be used with the task of training a translation model.

Training Setup for Multilingual Model We trained our models on a Google Cloud VM, on

Russian	А зарплата? - Примерно 10 долларов раз в 4 года.
Pivot	And the salary? - Approximately 10 dollars per 4 year
French Translation	Et le salaire? 10 dollars par quatre ans.
French Label	"Combien je serai payé?" "Dix dollars tous les quatre a

Figure 2: Daisy-chain translation from Russian to French. Notice that the quality of the translation is acceptable, however BLEU score is low due to punctuation.

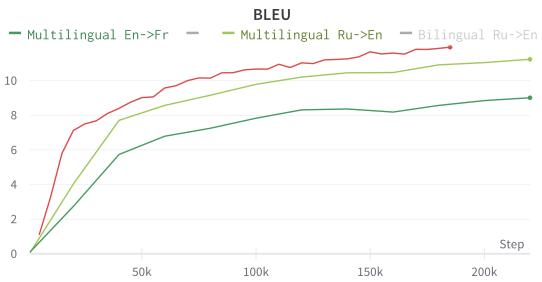

Figure 3: Validation performance of models during training.

one 16GB Tesla V100 GPU. Due to time and cost constraints, we also trimmed our original datasets, resulting in a total of 500,000 sentence pairs, evenly split between four language pairs. For training, the following parameters were used:

- learning rate 3e-4
- batch size 4
- gradient accumulation 16
- weight decay 0.0
- dropout rate 0.1
- linear scheduler
- warmup steps 5000
- epochs 1

Training Setup for Bilingual Models The setup for bilingual models was the same as for multilingual, noting that the dataset size for a single language pair is 250,000. The models are trained in both directions simultaneously We still trained for one epoch because of the computation constraints.

5 Results

	RU→EN	FR→EN	RU→FR
Bilingual	11.93	9.91	11.67
Multil	11.23	9.01	1.00*
Daisy			11.25*

Table 2: BLEU scores on both bilingual and multilingual models. Bilingual RU→FR model is trained on Opensubtitles. Zero-shot results are marked with a star*.

We evaluated the results of all three models on the FR→EN and RU→EN datasets from WMT14 (Table **??** and Figure 3). The difference in performance between the bilingual and multilingual models is noticeable, although not as large as anticipated. The zero-shot scores, however, are quite low, and need improvement.

Despite the low BLEU scores, a quick look at some of the examples revealed that BLEU scores were lower than they should be due to excess punctuation, as can be seen in Figure 2.

6 Conclusion

In this work, we have compared a bilingual machine translation system that uses an English pivot to direct zero-shot translation with a multilingual system using mT5 model. Our results demonstrated that the English pivot translation outperforms direct zero-shot translation by a large margin. We compared this with a fully-supervised model trained on the OpenSubtitles CC-Matrix dataset (Bojar et al., 2014) and observed similar performance to the pivot translation method.

References

Ondrej Bojar, Christian Buck, Christian Federmann, Barry Haddow, Philipp Koehn, Johannes Leveling, Christof Monz, Pavel Pecina, Matt Post, Herve Saint-Amand, Radu Soricut, Lucia Specia, and Ales Tamchyna. 2014. Findings of the 2014 workshop on statistical machine translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 12–58, Baltimore, Maryland, USA. Association for Computational Linguistics.

Akiko Eriguchi, Melvin Johnson, Orhan Firat, Hideto Kazawa, and Wolfgang Macherey. 2018. Zero-shot cross-lingual classification using multilingual neural machine translation.

Hila Gonen, Shauli Ravfogel, Yanai Elazar, and Yoav Goldberg. 2020. It's not greek to mbert: Inducing word-level translations from multilingual bert. *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*.

Pierre Lison and Jörg Tiedemann. 2016. Opensubtitles2016: Extracting large parallel corpora from movie and tv subtitles.

Bryan McCann, James Bradbury, Caiming Xiong, and Richard Socher. 2017. Learned in translation: Contextualized word vectors.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Łukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation.