Multilingual Translation

Evan Williams
Carnegie Mellon University
emwillia@cmu.edu

Charity King
Carnegie Mellon University
csking@andrew.cmu.edu

Abstract

This paper evaluates neural machine translation benchmarks for low-resource languages using COMET and BLEU metrics. The bilingual framework uses parallel data from one high-resource language English and two low-resource languages Azerbaijani and Belarus. Our multilingual model boosts the performance of low-resource languages by using parallel datasets from similar languages. This framework supplements our Azerbaijani (aze) dataset with Turkish (tur) and Belorussian (bel) with Russian (rus) to have increased COMET and BLEU performance for our low-resource languages.

1 Introduction

In this paper, we discuss the performance of bilingual machine translation baselines and multilingual machine translation baselines as well as several bilingual experiments that we ran for homework 2 of 11-737. Machine translation is a difficult task as syntax, morphology, and grammar can vary substantially across languages. Translation can result in the loss of context in languages where some words contain information not present in a target language, e.g., languages with words for older and younger brother but no word for brother. Additionally, high-quality machine translation models are extremely data-intensive, which results in low-quality translations for the vast majority of languages where data is limited to the Bible or simply not available (Siddhant et al., 2022). In this work, we explore the quality, or lack thereof, of our baseline bilingual predictions, assessing English $\rightarrow$ Azerbaijani and Azerbaijani $\rightarrow$ English as well as English-Belorussian translations in both directions. We then leverage cross-lingual transfer to improve the models. We append Turkish data to Azerbaijani data to improve Azerbaijani $\longleftrightarrow$ English translations and we append Russian data to Belorussian data to improve English $\longleftrightarrow$ Belorussian translations.

2 Bilingual Baseline Results

The results of our bilingual baseline model are displayed in Table 1. Each of the baselines was run using the transformer_iwslt_de_en architecture available in fairseq. One might assume that a model with this name would be pretrained on a de $\rightarrow$ de IWSLT task (Menezes, 2014), but it turns out that is not the case [1]. Rather, the difference between transformer_iwslt_de_en and a vanilla transformer architecture is that transformer_iwslt_de_en has 1024 hidden dimensions in its encoder forward function whereas the vanilla transformer (transformer_wmt_en_de) has 2048. We observe that in both models, translating to English performed slightly better than translating from English. We speculate that as the majority of the TED talks were likely delivered in English, the translations may vary in quality or implicitly follow the original English syntax and word order. In other words, translations out of English may not sound "natural." This may make some translations easier to map back to English, whereas results could be more mixed in the other direction.

	BLEU	COMET
AZE-ENG	**1.96**	**-1.1737**
ENG-AZE	1.54	-1.3101
BEL-ENG	**1.39**	**-1.3867**
ENG-BEL	1.29	-1.3987

Table 1: BLEU and COMET scores for Bilingual Azerbaijani and English models and Belorussian and English models.

[1] There currently is no documentation for this architecture on fairseq, but we verified this in the source code.

	BLEU	COMET
AZE⊕TUR → ENG	**11.97**	**-0.206**
ENG → AZE⊕TUR	6.05	-0.0913
BEL⊕RUS → ENG	**17.47**	**-0.3419**
ENG → BEL⊕RUS	9.91	-0.4414

Table 2: Cross-lingual (AZE BLEU and Comment scores. ⊕ is the concatenation operation

3 Cross-Lingual Training

To improve the two baseline model pairs, we leverage the similarity of Azerbaijani to Turkish and the Similarity of Belorussian to Russian. We append Azerbaijani data to Turkish data and append Belorussian data to Russian data and retrain our models. The results can be seen in Table 2. We observe a substantial increase in BLEU scores, and COMET scores all become less negative. Cross-lingual transfer increased the performance of each translation model.

In these models, the difference between source languages → English and English → source languages is even more pronounced. Translations from source languages into English are substantially better than the inverse. The two models with English as a target both achieve almost double the BLEU score of the models where English is the source. Once again, we speculate that this may be a result of the fact that many TED talks were likely translated from English to other languages, and so English artefacts may exist in the translations with which we train these models.

4 Qualitative Evaluation

Reading through a subset of the original English test data, and the BEL → ENG and the AZE → ENG. Even when word overlap existed, often the text would still be incomprehensible as displayed below. We also observe a large amount of repetition of filler words in the English translations. As TED talks are transcribed from speech and speakers often use "like", "uh", "or", etc., it may be decoded as the most likely token in all positions. An example of this behavior can be seen below in the AZ → EN translation.

Original English: “You know we’re interested, in , like, you know – (laughter) – an awkward interaction, or a smile, or a contemptuous glance, or maybe a very awkward wink, or like a handshake.”

BEL → EN: “(Laughter) (Applause) EM: Well, I’ve got a lot of them.”

AZE → EN: “(Laughter) Yeah, I’m going to see it’s going to see it or”

The AZE → EN repeats "or" over and over again. This might be an issue with the beam search, as ’or’ is a very common word in the corpus and will frequently be a likely token. It’s also worth noting that the original English sentence is not grammatically correct, and may present challenges in translation, especially if the translator attempts to keep the same style and grammatical errors as the original speaker. This further reinforces our hypothesis that there may be artifacts of translation from English that cause models from source languages to English to outperform models that use English as a source language.

5 Conclusion

In conclusion, we find that augmenting data with language data from similar languages can improve machine translation systems. We also find that the direction of the translations in the training data can have a substantial impact on model performance. This implies that there would likely be value, in terms of improving MT systems, in creating corpora translated from a non-English languages into English.

References

Arul Menezes. 2014. Speech translation for everyone – breaking down the barriers. In *Proceedings of the 11th International Workshop on Spoken Language Translation: Keynotes*, Lake Tahoe, California.

Aditya Siddhant, Ankur Bapna, Orhan Firat, Yuan Cao, Mia Xu Chen, Isaac Caswell, and Xavier Garcia. 2022. Towards the next 1000 languages in multilingual machine translation: Exploring the synergy between supervised and self-supervised learning. *arXiv preprint arXiv:2201.03110*.