Multilingual Neural Machine Translation

Yerin Han, Su Park, Malaika Vijay
Language Technologies Institute, Carnegie Mellon University
Pittsburgh, PA
yerinh,suminpar,mvijay@cs.cmu.edu

1 Multilingual Machine Translation

Machine Translation in low-resource languages has remained a challenge despite the ever-continuing innovation of deep learning models, due to the lack of availability of parallel data between the languages. For this assignment, we successfully reproduced the results on the provided baseline models and conducted two experiments to 1) improve the bilingual baseline by augmenting the data with copied target side monolingual data 2) improve multilingual transfer by selecting a second transfer language (Korean for Azerbaijan and Ukrainian for Belarusian) which resulted in a 345 percent increase in BLEU over the baseline for bel-eng.

2 Experiments

2.1 Data Augmentation by Copying Monolingual Data

This set of experiments studies the effect of replicating target side monolingual data to the source side to create additional pairs of training data. The premise of this modification is that the translation model will learn how to simply copy things such as numbers and named entities (Currey et al., 2017) that must remain the same across the source and target sentence. We also hypothesize that adding an autoencoder objective of reconstructing the source sentence will lead to richer encodings of the source sentence. In this experiment, we first obtain monolingual data for each target language. We then create a bitext between the source and target language by replicating the monolingual data to the source side. We use 10,000 monolingual sentences from the Leipzig Corpora Collection (dat, a) (dat, b) (dat, c) to generate this bitext, yielding a ratio of monolingual to original parallel data of 3:5 for Azerbaijani-English and 1:2 for Belarusian-English. This bitext is then mixed with the original parallel corpus and shuffled. We then experiment with two variations of training the Byte Pair Encoding (BPE) model. In the first experiment, we train the BPE model using the original parallel corpus only as in (Currey et al., 2017). In the second set of experiments we train the BPE model on the augmented parallel corpus. We then train the baseline bilingual model on this augmented data.

2.2 Choosing a Second Transfer Language

LangRank (Lin et al., 2019) is a framework that ranks the top languages for each low resource (LR) language to perform transfer from for various NLP tasks. It applies a gradient boosting model on the following dataset-dependent and data-independent features: the ratio of the dataset size, type-token ratio (the ratio between the number of unique words and the number of tokens), word overlap and subword overlap, geographic distance, genetic distance (genealogical distance of the languages derived from the Glottolog language tree), inventory distance (cosine distance between the phonological feature vectors from the PHOIBLE database), syntactic distance (cosine distance between syntactic feature vectors from the WALS database), and featural distance (the cosine distance between feature vectors combining all 5 mentioned above).
The LangRank model selected Turkish and Korean as the top two transfer languages for aze-eng, while for bel-eng, the ranking model recommended Ukranian and Russian, for which the exact statistics are listed in GitHub. We added the recommended second transfer language to the baseline multilingual training pipeline and experiment with upsampling as the data set size ratio and word overlap are crucial for effective machine translation.

3 Results and Analysis

3.1 Data Augmentation by Copying Monolingual Data

Table 1 presents the results of augmenting the original parallel corpus with replicated target-side monolingual data. The MC-PBPE (Monolingual Copy-

Model	Baseline		MC-PBPE		MC-ABPE		FLORES	
	BLEU	COMET	BLEU	COMET	BLEU	COMET	BLEU	COMET
aze-eng	2.83	-1.1681	3.56	-1.0712	4.38	-0.9833	13.67	-0.0495
eng-aze	1.54	-1.3101	2.61	-1.0968	3.20	-1.0631	8.04	0.0323
bel-eng	1.61	-1.3568	4.97	-1.1367	7.17	-1.0016	20.26	-0.0306
eng-bel	1.27	-1.4116	5.22	-1.1037	4.29	-1.1914	14.65	0.0327

Table 1: Results of the bilingual baseline model when trained without monolingual augmentation (Baseline),with monolingual augmentation and BPE model trained only on original data (MC-PBPE), and with monolingual augmentation and BPE model trained on the augmented data

Translation	GT	ORIG	AUG
aze-eng	167	3	99
eng-aze	176	1	157
bel-eng	95	2	62
eng-bel	98	2	138

Table 2: Word Overlap between the source sentence and the ground truth (GT), prediction without augmentation (ORIG) and prediction with augmentation (AUG)

ing - Parallel Byte Pair Encoding) model refers to the model where the BPE model is trained on the original parallel bitext only. Every language pair and direction of translation sees an increase in BLEU as well as COMET score. The MC-ABPE (Monolingual Copying - Augmented Byte Pair Encoding) model refers to the model where the BPE model is trained on the augmented parallel bitext. Interestingly, we see that the performance of models that translate into English benefit from the augmented BPE training, while models that translate from English do not.

In an effort to understand why monolingual copying boosts performance, we analyzed the amount of lexical overlap between the source and target sentences. We measure lexical overlap by counting the number of words that appear in the source and target sentence and sum these counts up across the entire corpus. We hypothesize that monolingual copying helps by teaching the model which words to translate unaltered, such as named entities and numbers. Table 2 presents the results of this analysis. We see that monolingual copying dramatically increases the number of words that get copied to the predicted sentence unaltered, bringing them closer to the ground truth predictions.

3.2 Choosing a Second Transfer Language

Table 3 shows that adding a second transfer language to each model resulted in an improvement in at least one metric for all models except aze-tur-kor-eng. The only improvement that adding Korean contributed to was from eng to aze (eng-kor-tur-aze), which resulted in a COMET score of -0.0505. This (44.68%) does exceed the increase that eng-bel saw by adding ukr and rus (32.17%). Applying upsampling resulted in improvements in both metrics for all of the four models. eng-kor-tur-aze-upsampled saw the highest improvement in COMET by 73.27%, followed by eng-ukr-rus-bel-upsampled with 57.25%, while the latter saw the highest increase in BLEU and the former the second highest. It's interesting to note that both metrics indicate that adding a transfer language and upsampling the LR benefit the translation for English to LR languages to a greater degree than from LR to English.

We again evaluated the lexical overlap between the ground truth and our predictions as a proxy metric to further estimate the quality of the produced translations. In GitHub, we've shown that adding the second transfer languages resulted in a marginally higher word-level overlap for aze-tur-kor-eng and eng-ukr-rus-bel, the trend of which coincides with LangRank's evaluation of the word and subword-level overlap between the LR's and the second transfer languages (0.47 on average for tur and kor with aze and 0.42 on average for ukr and rus with bel). We hypothesize that adding kor and ukr in the multilingual models enriched the produced translations as they share genetic and word-overlap-wise similarities with the sources and the first transfer languages, despite kor ranking as the second best transfer language after tur.

3.3 Error Analysis

Aside from analyzing the translation results using language-specific features as well as typological information, we also wanted to investigate whether there is a relationship between data-dependent fea-

Model	Baseline		2nd Transfer		2nd Transfer with Up	
	BLEU	COMET	BLEU	COMET	BLEU	COMET
aze-tur-eng / aze-tur-kor-eng	12.20	-0.2256	11.61	-0.2333	**13.04**	**-0.1671**
eng-tur-aze / eng-kor-tur-aze	6.05	-0.0913	5.89	**-0.0505**	**6.89**	**-0.0244**
bel-rus-eng / bel-rus-ukr-eng	17.47	-0.3419	**18.00**	-0.3508	**18.88**	**-0.2349**
eng-rus-bel / eng-ukr-rus-bel	9.91	-0.4414	**10.20**	**-0.2994**	**12.31**	**-0.1887**

Table 3: Results of the multilingual baseline models compared with the models with second transfer languages (2nd Transfer), and models with upsampled LR and second transfers (2nd Transfer with Up)

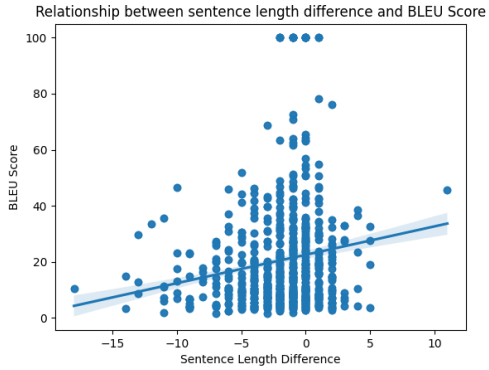

Figure 1: Relationship between sentence length difference and translation quality

tures such as sentence length with the quality of translation. We used the multilingual-bel-rus-ukr-up model which received the highest BLEU score out of all the other models for this analysis. We calculated the differences in sentence length in number of words between the source language and the target predictions (source length - target length) then measured the quality of each target sentence using BLEU. We found a more or less positive correlation between the two metrics as shown in Figure 1 where the greater the difference in sentence length between the source sentence and the target sentence, the more accurate the produced translation. The correlation graph also shows that the highest BLEU scores are concentrated in the region where the sentence length difference is close to zero.

We looked at three different examples of varying lengths to inspect the minute differences in the quality of translations for our models. Multilingual models resulted in the most fluent and accurate, if not even more natural than the ground truth sentences, by translating the ground truth sentences "Today I'm going to talk about unexpected discoveries. Now I work in the solar technology industry." into "Today I'm going to talk about unexpected discoveries. I'm now working with solar technology." For longer sentences, upsampling resulted in a seemingly enriched vocabulary and fluency with a display of metaphors ("And I'm just going to say, let's explore the power of your door, your eyes, your devices, to see the great ideas inside of you."), despite the awkward word choice. In contrast, Flores produced the shortest and the choppiest sentences ("I feel the surface in my body. Thank you." and "I'm going to talk about unexpected discovery today. Now, I'm working in solar technology industry.").

Bilingual Monolingual model produced the worst model across the three examples, where it completely omitted the predicate ("work") and failed to distinguish between "I am working in the industry" to "I am the industry" and mistranslated "light" to "right." Byte Pair Encoding used in the bilingual augmentation experiments resulted in a critical grammatical error and childish repetitions of words where the sentence "So I urge you to shut your eyes and discover the great ideas that lie inside us, to shut your engines and discover the power of sleep." was converted into "You're going to be a little bit of the world that you can're going to know that you can're going to be a lot of the world." for Augmentation, and " There's a little little little little bit of them, and you can see that you can see the little little bit of them, but you can see that you can see them." for Baseline. Whether this occurs for commonly used words such as "little" for all bilingual translations is a topic worthy of investigating further.

4 Limitations

Future studies can further hone in on understanding why adding the second transfer language resulted in a much higher performance going from English to LR than from LR to English.

References

a. Leipzig corpora collection (2020): Azerbaijani newspaper corpus based on material from 2020. leipzig corpora collection.

b. Leipzig corpora collection (2020): Belarusian newspaper corpus based on material from 2020. leipzig corpora collection.

c. Leipzig corpora collection (2020): English newspaper corpus based on material from 2020. leipzig corpora collection.

Anna Currey, Antonio Valerio Miceli Barone, and Kenneth Heafield. 2017. Copied monolingual data improves low-resource neural machine translation. In *Proceedings of the Second Conference on Machine Translation*, pages 148–156, Copenhagen, Denmark. Association for Computational Linguistics.

Yu-Hsiang Lin, Chian-Yu Chen, Jean Lee, Zirui Li, Yuyan Zhang, Mengzhou Xia, Shruti Rijhwani, Junxian He, Zhisong Zhang, Xuezhe Ma, et al. 2019. Choosing transfer languages for cross-lingual learning. *arXiv preprint arXiv:1905.12688*.