The AST Submission for the CoCo4MT 2023 Shared Task on Corpus Construction for Low-Resource Machine Translation

Steinþór Steingrímsson steinthor.steingrimsson@arnastofnun.is
The Árni Magnússon Institute for Icelandic Studies, Reykjavík, 107, Iceland

Abstract

We describe the AST submission for the CoCo4MT 2023 shared task. The aim of the task is to identify the best candidates for translation in a source data set with the aim to use the translated parallel data for fine-tuning the mBART-50 model. We experiment with three methods: scoring sentences based on n-gram coverage, using LaBSE to estimate semantic similarity and identify misalignments and mistranslations by comparing machine translated source sentences to corresponding manually translated segments in high-resource languages. We find that we obtain the best results by combining these three methods, using LaBSE and machine translation for filtering, and one of our n-gram scoring approaches for ordering sentences.

1 Introduction

Reliable parallel corpora are key to developing useful machine translation (MT) systems. Parallel corpora are commonly compiled by aligning corresponding documents in two or more languages on the sentence level, with the aim of pairing semantically equivalent segments in the languages. This can work well when source texts and translations of these texts are available. In cases where they are not available or not abundant, mining comparable corpora or web-scraped texts can be used to produce useful parallel pairs or to augment available parallel data. For some very low-resource (LR) language pairs or domains, neither parallel documents nor comparable corpora may be available, and in order to build an MT system that generates useful translation, creating a minimum set of training data by manually translating them may be necessary.

In scenarios where corpus creation for machine translation is carried out by translating sentences from a source language into a low-resource target language or in a specialized domain, it is important to use the annotation budget efficiently in order to come up with sentences that are likely to produce high-quality translations. In the CoCo4MT 2023 shared task, participants are to come up with ways to identify the best examples to translate from a high-resource (HR) source language, without any existing data in the target language. Translations of the source data are provided into a number of HR-languages and these translations can be used to help identify the best candidates for translation into the low-resource target language. A parallel corpus of 22,204 lines in English, German, Indonesian and Korean are provided. Participants can select up to 20% of these lines, with the corresponding pairs used to fine-tune an mBART model Liu et al. (2020), namely the mBART-50 (Tang et al., 2021). The winner of the shared task is the team whose instances result in the highest scoring model as measured by chrF++ (Popović, 2017).

Proceedings of the Second Workshop on Corpus Generation and Corpus Augmentation for Machine Translation, pages 33–38
September 5, 2023, Macau SAR, China.

2 Related Work

Parallel corpora compiled by aligning corresponding documents in two or more languages on the sentence level are available for a large number of language pairs, e.g. in the OPUS collection (Tiedemann, 2012). Ramesh et al. (2022) collect a combination of available parallel corpora and web-scraped material in a number of Indian languages. Bañón et al. (2020) collect web-scraped data, predominantly in the languages of the European Union and Bañón et al. (2022) aim to collect corpora for under-resourced European languages. For under-resourced languages unsupervised methods exploiting monolingual corpora have been applied (Lample et al., 2018; Artetxe et al., 2018), but have been found to need abundant monolingual data in similar domains, for the performance not to deteriorate (Marchisio et al., 2020).

Bhatnagar et al. (2022) present a system for choosing source language instances to annotate for MT. They find cross-lingual commonalities in instances that are useful for MT training and use these to identify instances for training a new language pair.

3 System Description

In selecting the best subset of segments using the HR-language pairs we consider sentence length, data diversity, misalignments and semantic equivalence. Our approaches are described in Section 3.2.

Before carrying out our experiments we did some preprocessing of the training, validation and test data sets.

3.1 Data Sets and Preprocessing

Participants in the shared task were provided access to parallel data from the JHU Bible corpus (McCarthy et al., 2020), in HR-languages to use for instance selection, and in LR-languages for evaluation of the data selection algorithms. The HR-languages were English, German, Korean and Indonesian, and the LR-languages were Gujarati, Burmese and French. Training, development and test splits for all languages, as well as baselines in the form of selected English instances, are made available in a GitHub repository for the shared task.[1]

The training files contain 22,204 lines, the test files are 8,708 lines and the development data comprise 3,919 lines. Upon inspection of the data we found that some lines contain empty strings in one or more of the HR-languages. We removed these empty lines from the English source file and deduplicated it, which gave us 19,718 lines to choose from. We also prepared our test and development data by removing all lines that were empty in one or more of the four HR-languages. This resulted in development files containing 3,410 lines and test files containing 7,622 lines. In our experiments, the development data are used for validation when fine-tuning the mBART-50 model on our selected training data sets and the test data for evaluating the usefulness of the data used in each experiment.

Data sets selected using two baseline methods are provided by the shared task organizers, one based on length and the other a random selection of sentences.

3.2 Experiments

Aiming to identify the instances in the training data that have the greatest potential to correctly inform an NMT system on how to correctly translate, we experimented with a number of approaches. Using the available HR-data, we evaluated our approaches on two translation directions: En→De and En→Id. We compare the results of our experiments to translations obtained by fine-tuning on the two baseline data sets. Our code is available on GitHub.[2]

[1] `https://github.com/ananyaganesh/coco4mt-shared-task/`

[2] `https://github.com/steinst/coco4MT`

Sentence length and diversity: In the task we are allowed a maximum number of sentences. Longer sentences should generally contain more information than shorter ones and should therefore be likely to give better results. One of the baselines is indeed a set of the 20% longest sentences in the source language, English. Instead of opting for a simple count of characters or tokens, we tokenize all source language sentences using the BPE-tokenization model used with mBART-50 and devise a greedy algorithm to try to order the sentences based on both length and diversity. The algorithm considers unigrams, bigrams and trigrams in the tokenized sentences and counts different such n-grams. In each round the highest scoring sentence is selected and removed from the pool of sentences. When previously selected sentences contain an n-gram for a set maximum number of times, it stops counting towards the score in the remaining sentences. Simplified pseudocode is given in Algorithm 1. We conduct three experiments, each having different number of allowed repetitions of each n-gram, with 1, 2 or 3 repetitions allowed.

Algorithm 1 Greedy Algorithm Selecting based on sentence length and n-gram diversity

```
$remaining\_lines = all\_source\_language\_lines$
$max\_ngrams = \{\}$
$allowed\_repetitions = n$
while $remaining\_lines \geq 1$ do
    for $line$ in $remaining\_lines$ do
        Count ngrams where (ngrams not in $max\_ngrams$ or $allowed\_repetitions \leq n$)
    end for
    $max\_ngrams \leftarrow$ ngrams from highest scoring line
    Remove highest scoring line from $remaining\_lines$
    Yield highest scoring line
end while
```

Semantic Similarity: When training MT systems we want the training data to contain accurate translations. LaBSE (Feng et al., 2022) is a model trained and optimized to produce similar representations for bilingual sentence pairs. It has been used for retrieving bitexts from parallel corpora. Feng et al. (2022) use it on its own for that purposes while Steingrímsson et al. (2021) use it as a part of a system that combines multiple approaches for the same purpose. It has been shown to be useful for scoring sentence pairs to identify possible faulty pairs that should be filtered out of an MT training set (Steingrímsson et al., 2023) and as a scoring mechanism for sentence alignment (Steingrímsson, 2023). In our experiments, we use LaBSE in combination with other approaches and remove all lines that obtain a LaBSE score under a given threshold for any of the three HR-language pairs with English is a source language. Feng et al. (2022) suggest that sentence pairs obtaining higher scores than 0.6 when mining comparable corpora can be useful for MT training. (Steingrímsson et al., 2023) experiment with using the scoring mechanism for multiple datasets and find that when working with data derived from parallel corpora, a lower threshold can be set. In our experiments we set a threshold to 0.5, with all sentence pairs obtaining lower scores being discarded.

Misalignments in the training data: Misalignments or partial misalignments in the training data can potentially have detrimental effects on the performance of MT systems (see e.g. Khayrallah and Koehn (2018)). To try to identify the most prominent misalignments we use the mBART-50 model, without any fine-tuning, to translate the English sentence pairs to German and Indonesian. For the HR-language pairs we expect to obtain translations that give a decent representation of the source sentences. We then calculate Chrf++ scores for each sentence pair

BLEU scores for two HR-language pairs		
Approach	EN→DE	EN→ID
Baseline: Longest Sentences	18.9	23.1
Baseline: Random Sentences	18.8	23.1
Greedy Algorithm - ngrams only count the first time (GA1)	17.4	23.4
Greedy Algorithm - First 2 occurrences of an ngram count (GA2)	18.4	26.0
Greedy Algorithm - First 3 occurrences of an ngram count (GA3)	18.2	25.9
GA1 + LaBSE and Chrf++ scores used for filtering LaBSE	18.4	23.0
GA2 + LaBSE and Chrf++ scores used for filtering LaBSE	21.3	26.2
GA3 + LaBSE and Chrf++ scores used for filtering LaBSE	17.9	23.6

Table 1: BLEU scores for different approaches.

and if the Chrf++ score is very low we assume there is a mismatch between the source and target sentences and remove these from our training data. On the other hand, if we achieve very high Chrf++ scores, we assume the contents of the sentence pairs are well represented in the training data and thus opt not to use these sentence pairs for training. We set the minimum threshold to 20.0 and maximum to 60.0.

Combinations of different methods: Finally, we try to combine our three approaches in various ways. The next section discusses the results for each of our approaches

4 Results and Discussion

We fine-tuned the mBART-50 model on the different datasets and selected the set of lines that obtained the highest BLEU score (Papineni et al., 2002) for English→German and English→Indonesian as evaluated on the cleaned test-set. Table 1 gives the results for the different approaches. We find that for English→German we only manage to beat the baseline instances once, while for English→Indonesian we do it all but once. There may be various reasons for this, one of them might be that the mBART-50 model is trained on large quantities of data in German, over 45M sentences Tang et al. (2021), and that our small parallel set is not enough to improve the translations to any extent. A lot less Indonesian data is used for training mBART-50, only 84k sentences, and thus a careful selection of parallel sentence pairs for fine-tuning may be more important in that case.

Our highest-scoring method uses the Chrf++ evaluation of translated sentences as well as a LaBSE threshold score to filter the datasets and remove lines that we deem more likely to be detrimental than others. We then order the remaining lines from highest to lowest scores obtained by running our greedy scoring algorithm and select the top 20% of the original number of lines, resulting in training sets of 4,440 lines for each language pair. This list of lines was submitted as our optimal list of instances for corpus construction.

5 Conclusion

We have described our approach to selecting 20% of the lines in a source language training set, for pairing with translations in other languages in order to obtain the optimal training data set from what is available. We find that our approach, as measured by BLEU to evaluate HR-language translation models, increases translation quality on the language that is not as well represented in the multilingual language model being fine-tuned. This may indicate that using only a low amount of data to fine-tune a model such as mBART-50 is more effective when the language is not well represented in the training data for the model.

References

Artetxe, M., Labaka, G., Agirre, E., and Cho, K. (2018). Unsupervised neural machine translation. In *International Conference on Learning Representations*.

Bañón, M., Chen, P., Haddow, B., Heafield, K., Hoang, H., Esplà-Gomis, M., Forcada, M. L., Kamran, A., Kirefu, F., Koehn, P., Ortiz Rojas, S., Pla Sempere, L., Ramírez-Sánchez, G., Sarrías, E., Strelec, M., Thompson, B., Waites, W., Wiggins, D., and Zaragoza, J. (2020). ParaCrawl: Web-Scale Acquisition of Parallel Corpora. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4555–4567, Online. Association for Computational Linguistics.

Bañón, M., Esplà-Gomis, M., Forcada, M. L., García-Romero, C., Kuzman, T., Ljubešić, N., van Noord, R., Sempere, L. P., Ramírez-Sánchez, G., Rupnik, P., Suchomel, V., Toral, A., van der Werff, T., and Zaragoza, J. (2022). MaCoCu: Massive collection and curation of monolingual and bilingual data: focus on under-resourced languages. In *Proceedings of the 23rd Annual Conference of the European Association for Machine Translation*, pages 303–304, Ghent, Belgium. European Association for Machine Translation.

Bhatnagar, R., Ganesh, A., and Kann, K. (2022). CHIA: CHoosing instances to annotate for machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 7299–7315, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Feng, F., Yang, Y., Cer, D., Arivazhagan, N., and Wang, W. (2022). Language-agnostic BERT Sentence Embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891, Dublin, Ireland. Association for Computational Linguistics.

Khayrallah, H. and Koehn, P. (2018). On the Impact of Various Types of Noise on Neural Machine Translation. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pages 74–83, Melbourne, Australia. Association for Computational Linguistics.

Lample, G., Conneau, A., Denoyer, L., and Ranzato, M. (2018). Unsupervised machine translation using monolingual corpora only. In *International Conference on Learning Representations*.

Liu, Y., Gu, J., Goyal, N., Li, X., Edunov, S., Ghazvininejad, M., Lewis, M., and Zettlemoyer, L. (2020). Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Marchisio, K., Duh, K., and Koehn, P. (2020). When does unsupervised machine translation work? In *Proceedings of the Fifth Conference on Machine Translation*, pages 571–583, Online. Association for Computational Linguistics.

McCarthy, A. D., Wicks, R., Lewis, D., Mueller, A., Wu, W., Adams, O., Nicolai, G., Post, M., and Yarowsky, D. (2020). The Johns Hopkins University Bible corpus: 1600+ tongues for typological exploration. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2884–2892, Marseille, France. European Language Resources Association.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). BLEU: A Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania. Association for Computational Linguistics.

Popović, M. (2017). chrF++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation*, pages 612–618, Copenhagen, Denmark. Association for Computational Linguistics.

Ramesh, G., Doddapaneni, S., Bheemaraj, A., Jobanputra, M., AK, R., Sharma, A., Sahoo, S., Diddee, H., J, M., Kakwani, D., Kumar, N., Pradeep, A., Nagaraj, S., Deepak, K., Raghavan, V., Kunchukuttan, A., Kumar, P., and Khapra, M. S. (2022). Samanantar: The Largest Publicly Available Parallel Corpora Collection for 11 Indic Languages. *Transactions of the Association for Computational Linguistics*, 10:145–162.

Steingrímsson, S., Loftsson, H., and Way, A. (2023). Filtering matters: Experiments in filtering training sets for machine translation. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 588–600, Tórshavn, Faroe Islands. University of Tartu Library.

Steingrímsson, S., Lohar, P., Loftsson, H., and Way, A. (2021). Effective Bitext Extraction From Comparable Corpora Using a Combination of Three Different Approaches. In *Proceedings of the 14th Workshop on Building and Using Comparable Corpora (BUCC 2021)*, pages 8–17, Online (Virtual Mode). INCOMA Ltd.

Steingrímsson, S. (2023). *Effectively compiling parallel corpora for machine translation in resource-scarce conditions*. PhD thesis, Reykjavik University.

Tang, Y., Tran, C., Li, X., Chen, P.-J., Goyal, N., Chaudhary, V., Gu, J., and Fan, A. (2021). Multilingual translation from denoising pre-training. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3450–3466, Online. Association for Computational Linguistics.

Tiedemann, J. (2012). Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218, Istanbul, Turkey. European Language Resources Association (ELRA).