Proceedings of SIGUL2022 @LREC2022, pages 78–82
Marseille, 24-25 June 2022

CUNI Submission to MT4All Shared Task

Ivana Kvapilíková and Ondřej Bojar
Institute of Formal and Applied Linguistics,
Faculty of Mathematics and Physics, Charles University,
Prague, Czech Republic,
kvapilikova@ufal.mff.cuni.cz, bojar@ufal.mff.cuni.cz

Abstract

This paper describes our submission to the MT4All Shared Task in unsupervised machine translation from English to Ukrainian, Kazakh and Georgian in the legal domain. In addition to the standard pipeline for unsupervised training (pretraining followed by denoising and back-translation), we used supervised training on a pseudo-parallel corpus retrieved from the provided monolingual corpora. Our system scored significantly higher than the baseline hybrid unsupervised MT system.

1. Introduction

Modern machine translation (MT) systems are trained on large parallel corpora, i.e. collections of sentence-aligned text documents translated by humans. While there are public sources of parallel data for several widely-spoken languages, most language pairs have a very limited access to such data. The same problem is faced by translation in high-resource languages but specific domains, since most training data come from newspaper articles and mixed tests crawled from the web. The MT4All project focuses on such low-resource situations and this shared task encourages participants to create unsupervised MT systems for translation from English to nine languages in three different domains: legal, financial and customer support.

In contrast to the standard MT, unsupervised MT models are trained without any parallel documents, but rather use large monolingual corpora to learn the structure of each language separately. Since monolingual texts are significantly easier to obtain (e.g. by web crawling) than parallel texts, unsupervised techniques have substantial amounts of non-translated text at their disposal, which can be leveraged to build a completely unsupervised translation system. Alternatively, parallel corpus (bitext) mining can be used to expand existing data resources by finding parallel sentences in comparable corpora (e.g. Wikipedia) and train an MT system in a supervised fashion even for low-resource languages.

The shared task organizers asked participants to either add value to existing unsupervised systems by adding monolingual training data or to train an unsupervised MT system from scratch. We chose the latter and trained a new MT system, but we also used an existing pretrained model to mine additional training data for our new model. We only participated in Task 1 which entailed unsupervised machine translation from English into Ukrainian, Kazakh and Georgian in the legal domain.

Section 2 of this paper summarizes related research in unsupervised MT. Section 3 describes the data sources and preprocessing, Section 4 gives more details on the parallel corpus we created. In Section 5 we describe the methodology used to build our system for the shared task and in Section 6 we discuss the results. Section 7 concludes.

2. Related Work

Unsupervised machine translation was pioneered by Artetxe et al. (2018b; Artetxe et al. (2018a) and Lample et al. (2018). They proposed unsupervised training techniques for both the phrase-based statistical machine translation (SMT) model and the neural machine translation (NMT) model to extract all necessary translation information from monolingual data. For the SMT model (Lample et al., 2018; Artetxe et al., 2018a), the phrase table is initialized with an unsupervised n-gram embedding mapping. For the NMT model (Lample et al., 2018; Artetxe et al., 2018b), the system is designed with a shared encoder and it is trained on batches of synthetic sentence pairs generated on-the-fly by denoising auto-encoding (Lample et al., 2018) and by back-translation (Sennrich et al., 2016). Artetxe et al. (2019a) push the translation quality higher by combining the two approaches and hybridizing their phrase based system. They train an NMT system with synthetic parallel data produced by the SMT system and jointly refine both systems by back-translation.

Conneau and Lample (2019) obtain similar results when pretraining the encoder and the decoder with a masked language model objective (Devlin et al., 2018) and fine-tuning for unsupervised MT. Song et al. (2019) pretrain the whole encoder-decoder structure on the task of reconstructing a sentence fragment given the remaining part of the sentence. The state of the art performance was reached in the work of Liu et al. (2020) who also pretrain an encoder-decoder model (mBART) and fine-tune using online back-translation. Tran et al. (2020) iteratively fine-tune mBART on the task of multilingual sentence retrieval as well as unsupervised translation and reach an improvement over vanilla mBART.

	en-ka	en-kk	en-uk
monolingual	22.4M x 6.3M	22.4M x 7.6M	22.4M x 9.7M
mined (all)	8.8M	4.7M	21.0M
mined (selected)	400K	300K	600K
mined (cleaned)	230K	169K	496K

Table 1: Final sizes (# of sentences) of cleaned mined parallel corpora in relation to the sizes of monolingual corpora we mined from.

	train (legal)	train (general)	dev	devtest
en	142K	22.4M	997	1,012
ka	264K	6.3M	997	1,012
kk	121K	7.6M	997	1,012
uk	7,601K	9.7M	997	1,012

Table 2: Number of sentences by splits in cleaned monolingual corpora.

3. Data

All provided data sources were monolingual. In addition to domain-specific data sets, the participants were allowed to use any part of the Oscar data set which was primarily intended for pretraining. The Oscar data set is large and we only used a part of it. The details of the data used are summarized in Table 2.

We used the sentence tokenizer from the `nltk` library to split the segments into sentences and we used the `fasttext` language detection model to get rid of sentences which do not appear to be in the desired language. The number of discarded sentences was around 6% of the entire corpus. The resulting size of the clean training corpora is reported Table 2.

Our NMT model processes text segmented into subword units. We used the sentencepiece (Kudo and Richardson, 2018) model trained for mBART50 (Liu et al., 2020) [1] to split the text into subwords and created a shared vocabulary from the most frequent 55k tokens covering 99.90% of the English monolingual training data and 99.97% of the Ukraininan, Kazakh and Georgian monolingual training data. The same vocabulary was used for all our models. The vocabulary size was determined to reasonably cover all training corpora while keeping the final size of the translation model limited.

It was also allowed to use any unsupervised pretrained model available in the Hugging Face Hub. We took the pretrained XLM-100 model and used it to mine parallel sentences from the monolingual corpora as proposed in (Kvapilíková et al., 2020). The details of the mining procedure are given below.

The validation data were taken from the Flores data set which belongs to the general domain. The blind test set was provided by the organizers and came from the legal domain.

4. Parallel Corpus Mining

Pretrained language models produce contextual representations capturing the semantic and syntactic properties of words in their context (Devlin et al., 2018). These representations may be aggregated to represent full sentences and used to assess sentence similarity. Multilingual language models can embed sentences in different languages and these embeddings can be used for parallel corpus mining.

We derive contextualized embeddings from the encoder outputs of the fifth-to-last internal layer of the XLM-100[2] model. It was shown by Kvapilíková et al. (2020) that the representations in the mid layers of the model carry the most multilingual information and are best aligned for the purpose of parallel sentence search.

We use the margin-based approach of (Artetxe and Schwenk, 2019a) to score all candidate sentence pairs rather than simple cosine similarity which cannot deal with the hubness phenomenon of embedding spaces (Artetxe and Schwenk, 2019b). The margin-based score is defined in relative terms to the average cosine similarity between the two sentences and their nearest neighbors, thus reducing the excessive score value of so called *hubs*.

Depending on the total number of retrieved candidates, we selected the top 600,000 sentence pairs for en-uk, 300,000 for en-ka and 400,000 for en-kk. A more careful selection or tuning of the quantities is left to future research. We then used the `clean-corpus-n.perl` script from Moses (Koehn et al., 2007) to get rid of sentences with less than 2 and more than 100 words and sentence pairs with a length ratio higher than 2. The resulting corpus sizes are summarized in Table 1.

An excerpt from the en-uk mined corpus is illustrated in Table 3. Most matched sentences include numerals, special symbols or named entities which probably serve as anchors for the models as they try to represent words in a language-neutral way. However, named entities are also often matched incorrectly. Some sentence pairs have no character overlap indicating that it is not only the identical tokens that drive the parallel sentence search but rather that at least some representations of tokens are properly aligned in the multilingual space. Even though the resulting data set is very noisy with a great number of errors, it seems to be enough to kick

[1] We originally intended to use the pretrained mBART50 model for training.)

[2] `https://huggingface.co/xlm-mlm-100-1280`

	uk	en
1	Encyclopædia Britannica, англ.	Encyclopædia Britannica, Inc.
2	№ 538.	Number 538.
3	Це 100%.	This 100%.
4	Все життя.	Nice life.
5	Їй було 35.	He was 31.
6	Свій!	Sure!
7	І він відмінно працює!	It works perfectly!
8	Це надзвичайно цікава історія.	It's an extraordinarily beautiful work.
9	Ці компанії раніше вже були включені [. . .]	Search features have been added into [. . .]
10	Одним з таких є приватна медична практика.	One of those is analytic continuation.
11	6 місяців назад вона народила дитину.	And two years ago she had another healthy baby boy.
12	Сьогодні стає зрозуміло, що боротьба з COVID-19 триватиме не один рік.	Today it is clear that the fight against COVID-19 will last more than one year.
13	Людське тіло містить від 55% до 78% води.	Human beings are made up of 50 – 86% .

Table 3: A sample from the en-uk mined parallel corpus. The translations are of differing quality, e.g. #12 is accurate, #11 and #14 have mistranslated numerals, #10 matches only in the first four words, #4 matches only in the second word.

off the training of an otherwise unsupervised MT system.

5. Training Methodology

We used the unsupervised training pipeline proposed by (Conneau and Lample, 2019). We first pretrained a cross-lingual masked language model (XLM) jointly on all data in English, Ukrainian, Kazakh and Georgian from scratch. The languages with a lower corpus size were upsampled to match the larger corpora. We used the pretrained model to initialize both the encoder and the decoder of an NMT model and fine-tuned with

1. standard MT objective using the mined parallel corpus,
2. online back-translation from monolingual data,
3. denoising from monolingual data.

After reaching convergence, we continued training using only denoising and online back-translation as we suspected that the translation quality of the trained MT system already surpassed the quality of the noisy corpus. After reaching convergence again, we further fine-tuned the model using online back-translation only on the texts from the legal domain.

All models were trained using the XLM toolkit.[3] on 4 GPUs with 16GB of RAM and delayed update of 2 to simulate training on 8 GPUs. The inference was performed with a beam size of 6. Selected training parameters are listed below

```
--tokens_per_batch 3450
--batch_size 30  #for back-translation
--accumulate_gradients 2
--amp 1
--fr16 True
--optimizer adam_inverse_sqrt,beta1=0.9,
beta2=0.98,lr=$LR,
```

		en-ka	en-kk	en-uk
1	MT-BT-DN	1.9	1.7	6.7
2	XLM + BT-DN	1.6	1.4	4.5
3	XLM + MT-BT-DN	3.4	2.7	9.0
4	(3) + BT	4.1	3.7	10.6
5	(4) + legal BT	4.1	3.8	8.8

Table 4: Validation results of our models on the Flores dev set. XLM - crosslingual masked LM pretraining; MT - supervised NMT fine-tuning on mined corpus; BT - online back-translation; DN - denoising.

		en-ka	en-kk	en-uk
1	MT-BT-DN	-	-	-
2	XLM + MT-BT-DN	-	-	
3	(2) + BT	**13.8**	7.7	27
4	(3) + legal BT	**13.8**	**9.4**	**28.1**
	Baseline	12	6.4	20.8

Table 5: Results of the submitted models. on the blind test set XLM - crosslingual masked LM pretraining; MT - supervised NMT fine-tuning; BT - online back-translation; DN - denoising.

6. Results

All results are measured on the detokenized data using `sacrebleu` (Post, 2018). In Table 4 we compare different techniques for NMT pretraining and its influence on the final translation quality.

To assess the impact of pretraining on the noisy parallel training data, we measured the performance of an unsupervised MT system trained according to the methodology of Conneau and Lample (2019) and we see an improvement of between 1.3 (en-kk) and 4.5 (en-uk) BLEU points caused by adding the mined parallel corpus. We also measured the effect of XLM pretraining

[3] https://github.com/facebookresearch/XLM

and we can conclude that for the language pairs in question, XLM pretraining significantly helps while also offering the flexibility of pretraining one multilingual model and using it to initialize all bilingual translation models.

When measured on the general Flores dev set, the effect of domain-specific fine-tuning is negative and leads to a decrease of up to 1.8 BLEU. However, when measured on the domain-specific test set, the fine-tuning adds up to 7.3 BLEU points (en-uk).

We were the only participants to this shared task so we cannot compare ourselves to other candidates but our models scored significantly higher than the baseline provided by the organizers. The baseline is a hybrid model trained from the bilingual word embeddings using the methodology of Artetxe et al. (2019b).

7. Conclusion

The performance of unsupervised models has significantly increased since the first attempts of Artetxe et al. (2018b) and Lample et al. (2018). We were able to train MT systems of reasonable quality for languages and domains where finding genuine parallel data is extremely difficult. We showed that adding a noisy parallel corpus mined from monolingual corpora to the training pipeline helps the final translation quality.

We submitted two unsupervised MT systems to the MT4All shared task, one of which was specifically fine-tuned for translation in the legal domain. Both systems scored significantly higher that the baseline (up to 7.3 BLEU points on the test set) but a comparison with the state-of-the-art mBART model remains for future work.

8. Acknowledgements

This work has received funding from the grant 19-26934X (NEUREM3) of the Czech Science Foundation, and support from the project "Grant Schemes at CU" (reg. no. CZ.02.2.69/0.0/0.0/19_073/0016935). This research was partially supported by SVV project number 260 575.

9. Bibliographical References

Artetxe, M. and Schwenk, H. (2019a). Margin-based parallel corpus mining with multilingual sentence embeddings. In *Proceedings of the 57th Annual Meeting of the ACL*, Florence, Italy. Association for Computational Linguistics.

Artetxe, M. and Schwenk, H. (2019b). Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610, Mar.

Artetxe, M., Labaka, G., and Agirre, E. (2018a). Unsupervised statistical machine translation. In *Proceedings of the 2018 Conference on EMNLP*, Brussels, November. Association for Computational Linguistics.

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

Artetxe, M., Labaka, G., and Agirre, E. (2019a). An effective approach to unsupervised machine translation. In *Proceedings of the 57th Annual Meeting of the ACL*, pages 194–203, Florence, July. Association for Computational Linguistics.

Artetxe, M., Labaka, G., and Agirre, E. (2019b). An effective approach to unsupervised machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 194–203, Florence, Italy, July. Association for Computational Linguistics.

Conneau, A. and Lample, G. (2019). Cross-lingual language model pretraining. In H. Wallach, et al., editors, *Advances in Neural Information Processing Systems 32*, pages 7059–7069. Curran Associates, Inc.

Devlin, J., Chang, M., Lee, K., and Toutanova, K. (2018). BERT: pre-training of deep bidirectional transformers for language understanding. *arXiv [e-Print archive]*, abs/1810.04805.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., and Herbst, E. (2007). Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the ACL*, pages 177–180, Prague, June. Association for Computational Linguistics.

Kudo, T. and Richardson, J. (2018). SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium, November. Association for Computational Linguistics.

Kvapilíková, I., Artetxe, M., Labaka, G., Agirre, E., and Bojar, O. (2020). Unsupervised multilingual sentence embeddings for parallel corpus mining. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 255–262, Online, July. Association for Computational Linguistics.

Lample, G., Ott, M., Conneau, A., Denoyer, L., and Ranzato, M. (2018). Phrase-based & neural unsupervised machine translation. In *Proceedings of the 2018 Conference on EMNLP*, pages 5039–5049.

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.

Post, M. (2018). A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on*

Machine Translation: Research Papers, pages 186–191, Brussels, Belgium, October. Association for Computational Linguistics.

Sennrich, R., Haddow, B., and Birch, A. (2016). Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the ACL*, pages 1715–1725, Berlin, August. Association for Computational Linguistics.

Song, K., Tan, X., Qin, T., Lu, J., and Liu, T.-Y. (2019). MASS: Masked sequence to sequence pre-training for language generation. In Kamalika Chaudhuri et al., editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 5926–5936. PMLR, 09–15 Jun.

Tran, C., Tang, Y., Li, X., and Gu, J. (2020). Cross-lingual retrieval for iterative self-supervised training. In H. Larochelle, et al., editors, *Advances in Neural Information Processing Systems*, volume 33, pages 2207–2219. Curran Associates, Inc.