Transfer Learning with Shallow Decoders: BSC at WMT2021's Multilingual Low-Resource Translation for Indo-European Languages Shared Task

Ksenia Kharitonova, Ona de Gibert Bonet, Jordi Armengol-Estapé, Mar Rodríguez i Alvarez, Maite Melero
Text Mining Unit, Barcelona Supercomputing Center
{ksenia.kharitonova,ona.degibert,jordi.armengol,
mar.rodriguez1,maite.melero}@bsc.es

Abstract

This paper describes the participation of the BSC team in the WMT2021's Multilingual Low-Resource Translation for Indo-European Languages Shared Task. The system aims to solve the Subtask 2: Wikipedia cultural heritage articles, which involves translation in four Romance languages: Catalan, Italian, Occitan and Romanian.

The submitted system is a multilingual semi-supervised machine translation model. It is based on a pre-trained language model, namely XLM-RoBERTa, that is later fine-tuned with parallel data obtained mostly from OPUS. Unlike other works, we only use XLM to initialize the encoder and randomly initialize a shallow decoder. The reported results are robust and perform well for all tested languages.

1 Introduction

We present the work carried out by the BSC Team in the context of WMT2021's first edition of the Multilingual Low-Resource Translation Shared Task. The tasks addresses the issue of multilinguality in machine translation (MT) for low-resource languages, focusing on two language families: North Germanic and Romance. We take part in the Subtask 2, which involves translation in four Romance languages: Catalan, Italian, Occitan and Romanian.

2 Background

Machine translation for low-resource languages is characterised by the lack of sufficient parallel data of a given language pair, either because the combination is infrequent or because the languages involved are themselves low-resource. Several works have attempted to overcome this pitfall, using different techniques. A common solution is to employ back-translation (Sennrich et al., 2016), while other research focuses on using other languages as pivots to compensate for the lack of data (Firat et al., 2016; Zoph et al., 2016). Artetxe et al. (2018); Lample et al. (2018) make use of monolingual data only.

Our approach is based on multilinguality. Previous works such as Vergés Boncompte and Ruiz Costa-Jussà (2020); Tubay and Costa-Jussa (2018) have shown that the use of multilingual MT is beneficial, as it generalizes better by sharing parameters among all the languages involved, especially if the languages belong to the same linguistic family. At the same time, training of multilingual MT models from scratch usually requires large parallel corpora and may not be feasible in a low-resource and zero-resource translation scenarios.

Pre-training of large language models from scratch on monolingual data and then fine-tuning them for the specific downstream tasks, has proved to be an extremely successful approach for many natural language processing problems. Cross-lingual language models such as XLM and XLM-RoBERTa (Conneau and Lample, 2019; Conneau et al., 2020), that combine unsupervised (monolingual data) and supervised (parallel data) training objectives, perform especially well both on cross-lingual NLU tasks and in machine translation. The idea of combining the power of pre-trained cross-lingual language models with a multilingual machine translation setting naturally follows from there.

This idea was explored in (Liu et al., 2020) where a denoising seq2seq auto-encoder (mBART) based on BART (Lewis et al., 2020) was pre-trained on extensive monolingual corpora in many languages. A similar approach is implemented in (Lin et al., 2020) where alignment information is used to pre-train a multilingual MT transformer on existing public parallel datasets. Both approaches require either a computationally intensive pre-training on monolingual data or access to extensive large-scale

Proceedings of the Sixth Conference on Machine Translation (WMT), pages 362–367
November 10–11, 2021. ©2021 Association for Computational Linguistics

parallel data.

Initializing an encoder and decoder of a bilingual MT seq2seq transformer with a pre-trained cross-lingual language model was famously proposed in (Conneau and Lample, 2019). A natural next step is to initialize a multilingual MT seq2seq transformer with a shared encoder and a shared decoder by the XLM-like language encoder which was first performed in (Ma et al., 2020).

We reuse this idea by initializing the encoder with a pre-trained XLM-Roberta (Conneau et al., 2020), as in (Ma et al., 2020). However, unlike (Ma et al., 2020), we only initialize the encoder, with the motivation of being able to instantiate a shallower decoder, following previous works that for a given compute budget suggest that it is more efficient to use deeper encoders and shallower decoders (Kasai et al., 2020). The encoder-only initialization was already implemented in Fairseq (Ott et al., 2019).[1]

3 Experimental Framework

3.1 Fine-Tuning Data

To train our MT system, we use all parallel data available in OPUS[2] for the targeted language pairs, namely ca-it, ca-oc, ca-ro.

We further include a small dataset ca-oc, the Catalan - Occitan Gencat Crawling, specifically obtained for the occasion, by leveraging parallel data from a crawling of the Catalan Government Internet domains and subdomains. We use the CorpusCleaner[3] pipeline to process the WARC files obtained from the crawling. This allows us to maintain the metadata and retrieve the original url per each document. We then extract the content of the same URLs in both languages and align them at document level using vecalign[4]. The final dataset of 503 sentences was obtained by manually reviewing 1,237 automatically aligned sentences. Although smaller than expected, one motivation to crawl this brand new dataset is to contribute to the development of MT resources for Occitan, which is a severely under-resourced language. We are publicly releasing this new dataset with an open license.[5]

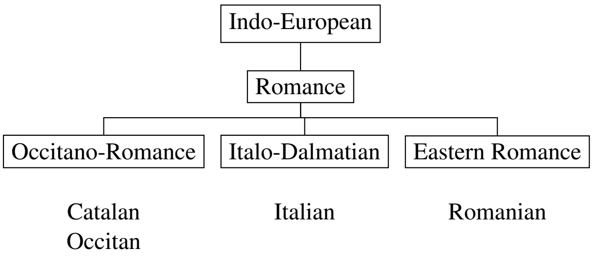

Figure 1: Family tree of Romance languages showing only the languages targeted by the Shared Task.

The resulting statistics of the corpora used to train our system can be seen in Table 1. As expected, the number of aligned sentences is much larger for Italian and Romanian as target languages, since Occitan is such a low-resource language. Nonetheless, we must bear in mind that Catalan and Occitan belong to the same sub-branch in the family tree of the Romance languages, as shown in Figure 1. Thus, their considerable typological closeness makes up for the reduced amount of aligned sentences available for this language pair.

3.2 Preprocessing

We start by preprocessing our data with a filtering and a tokenization step.

To ensure that there is no train-test overlap, we filter all of our training data by removing all sentences from the validation and test sets present in our train set.

To build our system, we use SentencePiece BPE tokenization with the original shared vocabulary of 250,000 tokens of XLM-R model (Conneau et al., 2020; Kudo and Richardson, 2018), and we only keep sentences with a maximum size of 512 tokens.

The final number of parallel sentences used for training is shown in Table 1.

3.3 System Description

We base our system on XLM-RoBERTa (Conneau et al., 2020) and then fine-tune it with the collected parallel data. As described earlier, the seq2seq multilingual transformer with shared encoders, shared decoders and shared embedding tables is initialized by XLM-R BASE pretrained language model on the encoder side, whereas a shallow decoder of 3 layers is initialized randomly. Sharing of embedding tables for all directions (ca-it, it-ca, ca-ro, ro-ca, ca-oc, oc-ca) was initially implemented due

[1] https://github.com/pytorch/fairseq/tree/v0.9.0

[2] https://opus.nlpl.eu/

[3] https://github.com/TeMU-BSC/corpus-cleaner-acl

[4] https://github.com/thompsonb/vecalign

[5] https://github.com/TeMU-BSC/wmt2021-indoeuropean/tree/master/gencat_crawling_ca-oc

Corpus Name	ca-it	ca-oc	ca-ro
EUbookshop v2	2,933	-	769
GlobalVoices v2018q4	6,036	-	468
GNOME v1	2,584	76	2,147
KDE4 v2	140,541	35,416	86,518
MultiCCAligned v1.1	1,335,785	-	890,155
OpenSubtitles v2018	359,798	-	387,044
QED v2.0a	61,013	245	57,279
Tatoeba v2021-03-10	296	-	2
TED2020 v1	49,674	33	46,978
Ubuntu v14.10	6,884	5,764	6,813
WikiMatrix v1	316,208	57,689	110,612
wikimedia v20210402	6,974	11,763	1,064
XLEnt v1.1	590,170	83,982	476,738
Catalan Government	-	503	-
Total	2,878,896	195,471	2,066,587
Cleaned	2,878,422	195,430	2,066,273
Tokenized	2,876,680	195,340	2,064,987

Table 1: Number of aligned sentences per corpus. The last row shows the final number of aligned sentences after the cleaning and tokenization steps.

Language	Tokens (M)
ca	1,752
it	4,983
oc	-
ro	10,354

Table 2: Number of million tokens per language present in the training corpus of XLM-R.

to memory constraints but eventually turned out to work well.

The token indicating the required target language is prepended to the target sentences, thus the model is aware to what language it has to translate to, as in (Wu et al., 2016).

It is important to note that the data used to train XLM-RoBERTa does not contain any Occitan text, as can be seen in Table 2. Thus the only knowledge that the multilingual transformer has about the language directions including Occitan comes from the XLM-R language model being pre-trained on text in related languages, such as Catalan.

We use default Fairseq parameters for fine-tuning, first of all, the Adam optimizer (Kingma and Ba, 2017) with $\beta_1 = 0.9$ and $\beta_2 = 0.98$. The polynomial decay learning rate schedule starts from $5\text{e}{-04}$, warmed up to over 1000 updates and gradually decays to 0 over around 60k updates. The model was fine-tuned for 2 days on 4 NVIDIA V100 GPUs. The final learning rate was around $3\text{e}{-04}$ with a batch size of $3,072$ sentences.

During inference we use the beam search generation algorithm with a beam size of 5. Since the languages between which we are translating are typologically close, we do not assign any length penalty, and we use the best checkpoint for generating.

4 Results

Here we report the official evaluation.[6] We submitted our results a bit later than the deadline due to some bottlenecks in our in-house computational resources. Table 3 reports the results obtained by our system on the evaluation test set, together with the two official baselines provided by the organisers (M2M-100 and mt5-devFinetuned).

Out of 7 competing systems and 2 baselines, our system was ranked 5th in Average, 3rd in the Catalan-to-Occitan direction, 4th in the Catalan-to-Romanian direction and 6th in the Catalan-to-Italian direction.

4.1 Human Evaluation

The organizers of the workshop have recently released the results of a human evaluation for the

[6] http://statmt.org/wmt21/multilingualHeritage-translation-results.html

Model	ca-it	ca-oc	ca-ro	Avg.
Ours	42.00	**57.10**	24.90	**49.77**
M2M-100	**46.75**	40.24	**33.06**	40.02
mT5-dev-ft.	30.38	40.14	17.33	29.28

Table 3: Official BLEU scores for the evaluation of the final test set

language pairs ca-it and ca-oc.[7] A sentence level evaluation has been performed taking into account the document as context. Each sentence is evaluated in a Likert-like scale [1,5] answering the question of direct assessments. A second human evaluation is performed where 60 selected terms (mostly named entities, dates and locations) are annotated as being either well translated, not translated or mistranslated, by majority voting among the annotators. The results can be found in Tables 4 and 5, respectively.

5 Discussion

5.1 Little sisters over big cousins

As seen in Table 3 the average results of our system are above both baselines, although it is the results for Catalan-Occitan that give the greater leverage, because in the other two scenarios M2M has a higher BLEU. Actually, the score for Catalan-Occitan is substantially higher than the score obtained for the other two pairs, although the fine-tuning data used in this model is, at least, ten times smaller than the data used in the other two models. These results are replicated in most of the other competing systems[8]. The reason for this apparent anomaly is clearly due to the linguistic similarity between the Catalan and Occitan, which in medieval times were practically one and the same language. This result confirms the intuition that when two languages are similar enough, less data is needed.

That said, we also hypothesize a positive impact of the curated dataset (Catalan - Occitan Gencat Crawling) added to the rest of parallel data obtained in the OPUS repository, but there is no definitive proof of it. Furthermore, we can also hypothesize that the presence of Spanish in the multilingual corpus, being a high-resource language and also linguistically close to Catalan and Occitan (more so than to the other two Romance languages involved in the task), has a beneficial impact on the results. Indeed, low-resource languages can greatly benefit from their similarity to other languages present in the multilingual training. In such scenarios, less data can lead to satisfactory results, and with a smaller carbon print, since the models use less computational power for training.

5.2 Human Evaluation results

The human evaluation on two of the test sets shown in in table 4 validates the relative position of our system in the global ranking. Interestingly, human scores correlate well with BLEU for Catalan-Italian, and less well for Catalan-Occitan. In the latter case, human scores tend to be lower than the corresponding BLEU. The reason for this may again have to do with linguistic similarity between Catalan and Occitan: "Catalanish" Occitan may be deemed acceptable by subword-based BLEU, but not by human evaluators. The performance of our system as evaluated for term translation, shown in in table 5 is consistent with the other evaluations regarding the position of the system in the overall ranking.

5.3 Vocabulary

One of the shortcomings of our approach is the big vocabulary size (250k tokens), inherited from XLM. This big vocabulary size was required by XLM to cover a very diverse set of languages. However, this makes it sometimes challenging to fit the embedding tables in memory, which is especially inefficient taking into account that a large proportion of tokens are not used (since we focus on a tiny subset of languages). Thus, either pruning the vocabulary, or using pre-trained models specifically trained for the Romance languages family (with a reduced vocabulary size) would be better alternatives.

5.4 Shallow decoders and transfer learning

While recent works have suggested that allocating more computation to (deeper) encoders (Kasai et al., 2020) at the expense of allocating less computation to (shallower) decoders is more efficient, this approach is not yet standard in the machine translation literature, especially when applying transfer learning. This method has the advantage of not reusing pre-trained weights for the decoder, although a middle ground is perhaps worth exploring.

[7] http://www.statmt.org/wmt21/multilingualHeritage-translation-manual.html

[8] http://statmt.org/wmt21/multilingualHeritage-translation-results.html

	ca-it		ca-oc	
Model	z-score	raw	z-score	raw
Human	0.8±0.4	4.8±0.6	0.8±0.7	4.0±1.0
Ours	-0.1±0.8	3.7±1.1	0.3±0.9	3.4±1.2
M2M-100	0.4±0.7	4.2±1.0	-0.7±0.8	2.0±1.0
mT5-dev-ft.	-1.2±0.9	2.3±1.2	-1.0±0.7	1.7±0.9

Table 4: Official human evaluation scores at sentence level

	ca-it				ca-oc			
Model	well	mis	no	$\sum$	well	mis	no	$\sum$
Human	53	0	3	56	40	0	2	42
Ours	27	7	5	39	33	4	0	37
M2M-100	33	2	6	41	26	9	0	35
mT5-dev-ft.	20	17	10	47	25	11	4	40

Table 5: Official human evaluation scores for 60 selected terms

Namely, use just some of the pre-trained weights to initialize the decoder layers. For example reuse the first N layers of XLM in the decoder, even if there is no 1-to-1 mapping between layers because there are less in the fine-tuned model.

6 Conclusions

We have showed that our approach is a simple, yet effective method for multilingual machine translation between linguistically similar languages. The encoder-only initialization allows for having a shallow decoder, which is computationally wise. As future work, we plan to further explore transfer learning techniques in the context of shallow decoders as well as applying different vocabulary pruning techniques.

Code availability

We release[9] with an open license the scripts used for this work for the sake of reproducibility.

Acknowledgements

This work was funded by the MT4All CEF project.[10]

References

Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho. 2018. Unsupervised neural machine translation. In *International Conference on Learning Representations*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Édouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, pages 7059–7069.

Orhan Firat, Baskaran Sankaran, Yaser Al-onaizan, Fatos T. Yarman Vural, and Kyunghyun Cho. 2016. Zero-resource translation with multi-lingual neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 268–277, Austin, Texas. Association for Computational Linguistics.

Jungo Kasai, Nikolaos Pappas, Hao Peng, James Cross, and Noah Smith. 2020. Deep encoder, shallow decoder: Reevaluating non-autoregressive machine translation. In *International Conference on Learning Representations*.

Diederik P. Kingma and Jimmy Ba. 2017. Adam: A method for stochastic optimization.

Taku Kudo and John Richardson. 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. Association for Computational Linguistics.

[9] https://github.com/TeMU-BSC/wmt2021-indoeuropean

[10] https://ec.europa.eu/inea/en/connecting-europe-facility/cef-telecom/2019-eu-ia-0031

Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018. Unsupervised machine translation using monolingual corpora only. In *International Conference on Learning Representations*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Zehui Lin, Xiao Pan, Mingxuan Wang, Xipeng Qiu, Jiangtao Feng, Hao Zhou, and Lei Li. 2020. Pre-training multilingual neural machine translation by leveraging alignment information. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2649–2663, Online. Association for Computational Linguistics.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8(0):726–742.

Shuming Ma, Jian Yang, Haoyang Huang, Zewen Chi, Li Dong, Dongdong Zhang, Hany Hassan Awadalla, Alexandre Muzio, Akiko Eriguchi, Saksham Singhal, Xia Song, Arul Menezes, and Furu Wei. 2020. Xlm-t: Scaling up multilingual machine translation with pretrained cross-lingual transformer encoders.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.

Brian Tubay and Marta R Costa-Jussa. 2018. Neural machine translation with the transformer and multi-source romance languages for the biomedical wmt 2018 task. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 667–670.

Pere Vergés Boncompte and Marta Ruiz Costa-Jussà. 2020. Multilingual neural machine translation: Case-study for catalan, spanish and portuguese romance languages. In *EMNLP 2020, Fifth Conference on Machine Translation: November 19-20, 2020, online: proceedings of the conference*, pages 447–450. Association for Computational Linguistics.

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, Lukasz 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. *CoRR*, abs/1609.08144.

Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. 2016. Transfer learning for low-resource neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1568–1575, Austin, Texas. Association for Computational Linguistics.