CUNI Basque-to-English Submission in IWSLT18

Tom Kocmi *Dušan Variš* *Ondřej Bojar*

Charles University, Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics
Malostranské náměstí 25, 118 00 Prague, Czech Republic
<surname>@ufal.mff.cuni.cz

Abstract

We present our submission to the IWSLT18 Low Resource task focused on the translation from Basque-to-English. Our submission is based on the current state-of-the-art self-attentive neural network architecture, Transformer. We further improve this strong baseline by exploiting available monolingual data using the back-translation technique. We also present further improvements gained by a transfer learning, a technique that trains a model using a high-resource language pair (Czech-English) and then fine-tunes the model using the target low-resource language pair (Basque-English).

1. Introduction

Despite becoming the current dominant approach in the field of machine translation (MT), neural machine translation (NMT) [1] systems still perform poorly in certain scenarios. One of them is learning to translate between language pairs where only a small amount of parallel data is available. Under these circumstances the NMT model quickly overfits and its performance plummets when translating sentences not seen during training. As observed in [2], with small parallel data, NMT performs much worse than the previous approach of phrase-based MT.

There are situations where the ability to learn an MT model of a reasonable quality given only a small amount of training data can be crucial. For example, when a crisis occurs in a region where an under-resourced language is spoken, a quick deployment of an MT system translating from or to that language can make a huge difference in the impact of the provided support [3].

In this paper, we describe the CUNI submission to the IWSLT Low Resource task for translating from Basque-to-English in the domain of TED talks. Our submission is based on the recently introduced self-attentive network architecture called Transformer [4]. We improve the performance of this model by exploiting the English in-domain monolingual data using the back-translation technique [5]. We achieve further improvements via transfer learning. Transfer learning [6, 7] consists of training a "parent" (high-resource) model first and then continuing the training on the "child", low-resource, parallel data as a means of model adaptation. Furthermore, we combine several models saved at training checkpoints by simply averaging the weights ("model averaging") as a substitute of model ensembling.

The structure of the paper is the following. In Section 2, we describe the method of transfer learning followed by the description of back-translation in Section 3. The model description is presented in Section 4 and the dataset overview in Section 5. Section 6 details the results achieved by our systems. Section 7 discusses other works in the area of low-resource translation systems. And finally Section 8 concludes the paper.

2. Transfer learning

Transfer learning is based on the observation that neural machine translation model that is first trained on the parallel data of a high-resource language pair can be adapted to a lower-resource language pair. The two languages can have a linguistic relation, however, transfer learning works even for unrelated languages [7].

The method starts by first training the parent model until it reaches the best possible performance or until a fixed number of gradient updates is performed. This model is then adapted by switching the training dataset from the parent pair to the low-resource child pair. During this transition, we do not change any hyperparameters nor the learning rate.

The transfer learning method does not need any modification of the existing NMT pipeline. The method only relies on a single condition: the vocabulary has to be shared across all the languages in the parent as well as child language pairs.

We construct the shared vocabulary using subword tokens, namely wordpieces [8], instead of words. This way, we are able to handle words not seen during training by splitting them into subwords, which are present in the vocabulary. We learn the subword segmentation using concatenated source and target sides of both the parent and child language pairs. To avoid bias in the vocabulary towards the high-resource language pair, [7] suggest to sample a subset of the sentences from the high-resource language pair that has a size similar to the low-resource language pair dataset, calling this approach "balanced vocabulary". They also showed, that a significant portion of this balanced vocabulary is relevant only for the child model, as it never appears in the parent training data.

Unlike [7], we also experiment with additional vocabulary setups, using either only the parent (or only the child) training data to generate the vocabulary. We call these restricted setups "parent vocabulary" and "child vocabulary", respectively. The idea behind the use of "child vocabulary" is that there will be more child-specific wordpieces which can lead to a better performance of the child model. On the other hand, the reasoning behind the "parent vocabulary" is that we can use only a single parent for the training of several different child models and therefore save the time of training parent models for each child separately.

3. Back-translation

The organizers of IWSLT 2018 provided participants with a vast amount of English monolingual data to use in their system submissions, both in-domain and out-of-domain. We exploit the English in-domain TED talks monolingual data for creation of the synthetic data as described by [5].

The key idea is to use an MT system trained to translate in the opposite direction (English-to-Basque) and use it to translate the monolingual data. These synthetic outputs, when paired with the input monolingual data, can be then used as additional parallel data for the original (Basque-to-English) direction. Even though the source side is noisy, the additional training examples help the decoder to learn a more fluent target side language model.

We use this method to back-translate only the in-domain TED talks data because it is the target domain of the Low Resource task.

To create the synthetic parallel data from the English monolingual corpus, we used a Transformer model and transfer learning. We first trained on the English-to-Czech corpus and then adapted the model using English-to-Basque corpus. This was based on our previous experiments where transfer learning resulted in a model with a better translation performance and therefore a better quality of synthetic data.

4. Model description

We use the self-attentive neural network architecture called Transformer [4]. We chose this network architecture due to its reported state-of-the-art results [9, 10],[1] making it a strong baseline for our experiments.

The architecture follows the encoder-decoder paradigm where the encoder creates hidden representations of the source language tokens and the decoder outputs the target sequence conditioned on that source language representations and the representations of the already decoded tokens.

The self-attentive encoder contains several layers each consisting of two sublayers: the first one applies a self-attention and the second one a feed-forward network. The decoder is similar, including an additional attention-over-encoder layer between its own self-attention and the feedforward layers. The self-attention layer is the key component of the Transformer architecture, effectively modeling the context of each token and thus substituting other methods such as the recurrent hidden units [1, 11] or convolutional networks [12]. The absence of recurrent units makes the training much faster due to a possible parallelism while requiring a lower number of layers when compared to the convolutional network.

We use the Transformer implemented in Tensor2Tensor [13],[2] version 1.4.2. Our models are based on the "big single GPU" configuration as defined in the paper. We use the default setup, only changing the batch size to 2300 and a maximum sentence length to 100 wordpieces in order to fit the model to our GPUs (NVIDIA GeForce GTX 1080 Ti with 11 GB RAM).

We use Noam learning rate decay scheme with the starting learning rate of 0.2 and 32000 warm-up steps. The decoding uses the beam size of 8 and length normalization penalty is set to 0.8.

Dataset	Sentences	Tokens EN	Tokens CS/EU
Genuine EN-EU	0.9 M	7.0 M	5.1 M
Genuine EN-CS	40.1 M	563.4 M	490.5 M
Synthetic EN-EU	0.3 M	5.3 M	3.6 M

Table 1: Sizes of the parallel corpora. The "synthetic" have Basque side back-translated from English.

5. Dataset

For Basque-English, we used all the available data that were allowed by the organizers of IWSLT 2018. The parallel corpora consist of only around 5,600 in-domain (TED) sentence pairs and around 940,000 out-of-domain sentence pairs.

In addition to the resources suggested by the organizers, we also used data from OPUS and WMT, which were also allowed. Specifically, corpora PaCo2 English-Basque and QTLeap Batches 1-3 from WMT.[3]

For English-Czech, we use all parallel data available for WMT 2018 except of the Paracrawl. The majority of the data is part of the CzEng 1.7 corpus (the filtered version, [14]).[4]

We also created synthetic Basque-English data using back-translation. We generate them by translating all English sentences from the TED talks data gathered across all language pairs provided for IWSLT 2018. The data do not contain sentences from talks in test set.

From all training sentences, we dropped sentence pairs shorter than 4 words or longer than 75 words on either source or target side. This results in a speedup of the training by allowing a larger batch size. A similar setup was used in [15] where the authors argue that in their experiments, the

[1] http://www.statmt.org/wmt18/translation-task.html

[2] https://github.com/tensorflow/tensor2tensor

[3] http://www.statmt.org/wmt16/it-translation-task.html

[4] https://ufal.mff.cuni.cz/czeng/czeng17

Vocabulary	CS to EN (BLEU)	EU to EN (BLEU)
Child only	24.93	22.92
Parent only	27.81	23.29
Balanced	27.93	23.63
Baseline	–	19.09

Table 2: The results of transfer learning. The first column shows the performance of the parent model, the second column is the child model based on the corresponding parent. The baseline does not use transfer learning. The results are reported on the development set. Scores are comparable only within columns.

performance is not negatively influenced by the reduction of training data.

To evaluate the models during training we used the development data provided by IWSLT 2018 (Basque-English) and development data available for WMT (Czech-English), namely WMT 2011 Newstest.

6. Results

In this section we first compare results obtained when using the three types of vocabulary and then describe our systems submitted to the IWSLT evaluation.

6.1. Effect of vocabulary

We experiment with three types of shared vocabulary as described in Section 2. All setups use the exact same data (and a same layout of the transfer learning); they differ only in the vocabulary. First, we trained three models from Czech-to-English with different vocabularies for 1M steps and then we continued with the transfer learning of the child Basque-to-English models until their performance on the validation set stopped improving.

As seen in Table 6.1, the transfer learning for Basque-to-English improves the model performance significantly over the baseline, gaining over 4 BLEU points (19.09 vs. 23.63).

When we look at the parent-only or child-only vocabulary setups, both performed worse than the balanced vocabulary. With the balanced vocabulary, we obtain the best result on the Basque-to-English translation. We suppose that the same holds for other language pairs too, since there is no language specific restriction.

Still, it would be interesting to know whether the data ratio 50:50 is the best possible setup or whether other ratios could improve the results. We plan to investigate this in our future work. We assume that the exact ratio might be language specific, however, in general, using the balanced approach with an equal representation of the languages might still be an effective option.

Run	Transfer	Back-translation	Genuine	BLEU	NIST	TER
Primary	✓	✓	✓	22.86	6.01	60.31
Contrastive 1	–	–	✓	16.13	4.98	66.55
Contrastive 2	✓	✓	–	22.26	6.00	63.89
Contrastive 3	✓	–	✓	21.11	5.84	62.34

Table 3: Results of our submissions. Official evaluation on the test set.

6.2. Final results

We submitted several contrastive models for the final IWSLT evaluation. All our systems use the same balanced vocabulary. The synthetic data were generated by the English-to-Basque system. All final models are averaged over the last 5 checkpoints.

The primary system uses the transfer learning: the parent model is trained for 1M steps on Czech-to-English, followed by transfer learning using only the synthetic data for 405k steps, and completed by 60k steps on the genuine, original parallel data.

The run labelled "Contrastive 1" in Table 3 is the baseline trained only on the official parallel Basque-to-English data.

"Contrastive 2" uses transfer learning on the parent model Czech-to-English trained for 1M steps, followed by training on only the synthetic English-to-Basque data, without the use of genuine parallel data.

Finally, "Contrastive 3" also uses transfer from Czech-to-English as the primary, followed by genuine parallel English-to-Basque data, without the use of any synthetic data.

As clearly confirmed by three automatic metrics, the combination of the back-translation and transfer learning leads to the best performance.

7. Related work

In [16], Firat et al. propose zero-resource multi-way multilingual systems, with the main goal of reducing the total number of parameters needed to train multiple source and target languages. To prevent the network from forgetting the previously learned language pairs, they implement a special training schedule.

Another multilingual approach is proposed by [8] where Johnson et al. simply use all translation pairs at the same time and the choice of the target language happens at runtime by special token at the end of the input sentence. This forces the model to learn to translate between many languages, including language pairs without available 'direct' parallel data.

The lack of sufficient amounts of parallel data can be also tackled by unsupervised translation [17, 18]. The general idea is to train monolingual embeddings using large amounts of monolingual data and finding a projection from the source to target words that preserves the structure of embedding vector spaces [19]. Using these shared fixed bilingual embeddings an architecture with a shared encoder [18] or both

shared encoder and decoder [17] is then trained using multiple training objectives.

Aside from the common back-translation [5], simple copying of the target monolingual data back to the source-side has also been shown to improve translation quality in the low-resource setting [20].

The transfer learning we used could be also seen as a variant of the so-called "curriculum learning" [21, 22], where the training data are ordered from foreign out-of-domain to the in-domain training examples to speed up the training convergence.

8. Conclusion

In this paper, we presented our systems for IWSLT 2018 low-resource Basque-to-English translation task. We reached a significant improvement using transfer learning and back-translation. We compared three types of vocabularies used for the transfer learning and concluded, that the balanced vocabulary is the best option.

9. Acknowledgements

This study was supported in parts by the grants SVV 260 453, GAUK 8502/2016, GAUK 1140218, and 18-24210S of the Czech Science Foundation. This work has been using language resources and tools stored and distributed by the LINDAT/CLARIN project of the Ministry of Education, Youth and Sports of the Czech Republic (projects LM2015071 and OP VVV VI CZ.02.1.01/0.0/0.0/16 013/0001781).

10. References

[1] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, ser. NIPS'14. Cambridge, MA, USA: MIT Press, 2014, pp. 3104–3112. [Online]. Available: http://dl.acm.org/citation.cfm?id=2969033.2969173

[2] P. Koehn and R. Knowles, "Six Challenges for Neural Machine Translation," in *Proceedings of the First Workshop on Neural Machine Translation*. Vancouver: Association for Computational Linguistics, August 2017, pp. 28–39. [Online]. Available: http://www.aclweb.org/anthology/W17-3204

[3] W. Lewis, R. Munro, and S. Vogel, "Crisis mt: Developing a cookbook for mt in crisis situations," in *Proceedings of the Sixth Workshop on Statistical Machine Translation*. Association for Computational Linguistics, July 2011. [Online]. Available: https://www.microsoft.com/en-us/research/publication/crisis-mt-developing-a-cookbook-for-mt-in-crisis-situations/

[4] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems 30*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds. Curran Associates, Inc., 2017, pp. 6000–6010. [Online]. Available: http://papers.nips.cc/paper/7181-attention-is-all-you-need.pdf

[5] R. Sennrich, B. Haddow, and A. Birch, "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)*. Berlin, Germany: Association for Computational Linguistics, August 2016, pp. 86–96. [Online]. Available: http://www.aclweb.org/anthology/P16-1009

[6] B. Zoph, D. Yuret, J. May, and K. Knight, "Transfer learning for low-resource neural machine translation," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Austin, Texas: Association for Computational Linguistics, November 2016, pp. 1568–1575. [Online]. Available: https://aclweb.org/anthology/D16-1163

[7] T. Kocmi and O. Bojar, "Trivial Transfer Learning for Low-Resource Neural Machine Translation," in *Proceedings of the 3rd Conference on Machine Translation (WMT)*, Brussels, Belgium, Nov. 2018.

[8] M. Johnson, M. Schuster, Q. Le, M. Krikun, Y. Wu, Z. Chen, N. Thorat, F. a. Vigas, M. Wattenberg, G. Corrado, M. Hughes, and J. Dean, "Google's multilingual neural machine translation system: Enabling zero-shot translation," *Transactions of the Association for Computational Linguistics*, vol. 5, pp. 339–351, 2017. [Online]. Available: https://transacl.org/ojs/index.php/tacl/article/view/1081

[9] M. Popel and O. Bojar, "Training Tips for the Transformer Model," *The Prague Bulletin of Mathematical Linguistics*, vol. 110, no. 1, pp. 43–70, 2018. [Online]. Available: https://content.sciendo.com/view/journals/pralin/110/1/article-p43.xml

[10] O. Bojar, C. Federmann, M. Fishel, Y. Graham, B. Haddow, M. Huck, P. Koehn, and C. Monz, "Findings of the 2018 conference on machine translation (WMT18)," in *Proceedings of the Third Conference on Machine Translation, Volume 2: Shared Task Papers*. Brussels, Belgium: Association for Computational Linguistics, October 2018.

[11] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *CoRR*, vol. abs/1409.0473, 2014. [Online]. Available: http://arxiv.org/abs/1409.0473

[12] J. Gehring, M. Auli, D. Grangier, D. Yarats, and Y. N. Dauphin, "Convolutional sequence to sequence learning," in *ICML*, ser. Proceedings of Machine Learning Research, vol. 70. PMLR, 2017, pp. 1243–1252.

[13] A. Vaswani, S. Bengio, E. Brevdo, F. Chollet, A. Gomez, S. Gouws, L. Jones, L. Kaiser, N. Kalchbrenner, N. Parmar, R. Sepassi, N. Shazeer, and J. Uszkoreit, "Tensor2Tensor for Neural Machine Translation," in *Proceedings of the 13th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Papers)*. Boston, MA: Association for Machine Translation in the Americas, March 2018, pp. 193–199. [Online]. Available: http://www.aclweb.org/anthology/W18-1819

[14] O. Bojar, O. Dušek, T. Kocmi, J. Libovický, M. Novák, M. Popel, R. Sudarikov, and D. Variš, "Czeng 1.6: enlarged czech-english parallel corpus with processing tools dockered," in *International Conference on Text, Speech, and Dialogue*. Springer, 2016, pp. 231–238.

[15] T. Kocmi, oman Sudarikov, and O. Bojar, "CUNI Submissions in WMT18," in *Proceedings of the 3rd Conference on Machine Translation (WMT)*, Brussels, Belgium, Nov. 2018.

[16] O. Firat, K. Cho, and Y. Bengio, "Multi-Way, Multilingual Neural Machine Translation with a Shared Attention Mechanism," in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. San Diego, California: Association for Computational Linguistics, June 2016, pp. 866–875. [Online]. Available: http://www.aclweb.org/anthology/N16-1101

[17] M. Artetxe, G. Labaka, E. Agirre, and K. Cho, "Unsupervised neural machine translation," in *Proceedings of the Sixth International Conference on Learning Representations*, April 2018.

[18] A. Conneau, G. Lample, M. Ranzato, L. Denoyer, and H. Jégou, "Word translation without parallel data," *CoRR*, vol. abs/1710.04087, 2017.

[19] M. Artetxe, G. Labaka, and E. Agirre, "Learning bilingual word embeddings with (almost) no bilingual data," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Vancouver, Canada: Association for Computational Linguistics, July 2017, pp. 451–462. [Online]. Available: http://aclweb.org/anthology/P17-1042

[20] A. Currey, A. V. M. Barone, and K. Heafield, "Copied monolingual data improves low-resource neural machine translation," in *Proceedings of the Second Conference on Machine Translation*, 2017, pp. 148–156.

[21] Y. Bengio, J. Louradour, R. Collobert, and J. Weston, "Curriculum learning," in *Proceedings of the 26th annual international conference on machine learning*. ACM, 2009, pp. 41–48.

[22] T. Kocmi and O. Bojar, "Curriculum Learning and Minibatch Bucketing in Neural Machine Translation," in *Recent Advances in Natural Language Processing 2017*, Sept. 2017.