The University of Helsinki submissions to the IWSLT 2018 low-resource translation task

Yves Scherrer

Department of Digital Humanities
University of Helsinki, Finland
yves.scherrer@helsinki.fi

Abstract

This paper presents the University of Helsinki submissions to the Basque–English low-resource translation task. Our primary system is a standard bilingual Transformer system, trained on the available parallel data and various types of synthetic data. We describe the creation of the synthetic datasets, some of which use a pivoting approach, in detail. One of our contrastive submissions is a multilingual model trained on comparable data, but without the synthesized parts. Our bilingual model with synthetic data performed best, obtaining 25.25 BLEU on the test data.

1. Introduction

The University of Helsinki has participated in the IWSLT low-resource translation task on Basque-to-English translation with one primary and two contrastive systems. Our experiments mainly focused on creating synthetic training data for classical supervised neural machine translation models. In particular, we show that a bilingual system trained on partly synthetic data performs better than a multilingual system that includes the same data in their original, non synthetic form. Our best submitted system obtained a BLEU score of 25.25.

Section 2 describes the available Basque–English parallel datasets at the basis of our systems, as well as a baseline system trained on these parallel datasets alone. In Section 3, we present additional datasets that contain either Basque or English text, but not both. We discuss several strategies for synthetically creating parallel Basque–English datasets out of these sources, and show the impact of these synthetic datasets on translation quality. In Section 4, we present a contrastive system that uses the additional datasets in their original state, without the synthesized parts. Section 5 summarizes our submissions and details the post-processing steps carried out at prediction time.

2. Parallel Basque–English data

The IWSLT organizers released an in-domain data set for Basque-to-English translation containing 64 TED talks for training and 10 TED talks for development [1]. Another 10 TED talks have been held out for testing.

The only allowed out-of-domain data source containing parallel Basque–English datasets is OPUS [2, 3]: it contains computer program localization files (repositories GNOME, KDE4 and Ubuntu), crowd-sourced translations (Tatoeba) and film subtitles (OpenSubtitles2018). We only selected OpenSubtitles2018 as the largest and most domain-similar dataset for our experiments. Table 1 summarizes the available parallel data.[1]

Source	Talks	Lines	EU tokens	EN tokens
TED train	64	5623	97k	128k
OST	—	806k	4.8M	6.5M
TED dev	*10*	*1140*	*20k*	*27k*

Table 1: Basque–English parallel data.

2.1. Baseline system

We trained a baseline system using only the parallel data mentioned in the previous section. Data were tokenized and truecased using the Moses scripts [4]; no effort was spent on adapting the tokenization tools to Basque. Following the good results on various typologically diverse language pairs, we used the Transformer model setup [5] as implemented in Marian-NMT [6] (see Appendix). We used an initial setting of 20 000 BPE units [7] shared across both languages with tied embeddings. Training of this model converged after about 20 hours on a single-GPU node, obtaining a BLEU score of 15.40 on the development set (see first line of Table 5).[2]

3. Synthetic data

Backtranslation has proven to be an effective way of improving the performance of neural machine translation systems by taking advantage of existing monolingual datasets for the target language [8]. Monolingual data of the target language is translated to the source language using a target-to-source

[1] In all tables, validation and test sets are displayed in italics, whereas the translation output of the described system is displayed in bold (if applicable).

[2] All BLEU scores were computed using the *multi-bleu-detok.perl* script of the Moses distribution.

translation system. The resulting bilingual dataset, whose source is noisy, is then used as additional training data for the source-to-target translation system.

In our setting, direct backtranslation would amount to translating English data to Basque, but such an English-to-Basque system would have to be trained on the same small dataset as the baseline system presented above. Therefore, we experimented with other ways of creating synthetic data, exploiting the larger Spanish–Basque and Spanish–English datasets and using Spanish as a pivot language [9].[3] The different data augmentation strategies are discussed in Sections 3.1 to 3.3, whereas the Basque-to-English systems trained on these synthetic datasets are presented in Section 3.4 and Table 5.

3.1. Direct backtranslation of TED talks

The provided in-domain data contains a total 2683 English TED talks. Excluding those that already have Basque translations (for training, development or testing) and excluding those that do not have a Spanish translation (to provide comparability with the experiment below), 2576 English TED talks can be backtranslated to Basque.

Source	Talks	Lines	EN tokens	EU tokens
TED train	64	5623	128k	97k
OST	—	806k	6.5M	4.8M
TED dev	*10*	*1140*	*27k*	*20k*
TED direct-BT	2576	271k	6.2M	**3.9M**

Table 2: Basque–English data used to train the backtranslation model (above the line) and monolingual English data backtranslated with this model (below the line, backtranslation output in bold).

In this first experiment, we train an English-to-Basque system analogously to the baseline system above, using the same training data, parameter settings (20k joint BPE units) and development set for validation, obtaining a BLEU score of 8.65 on the English-to-Basque development set.[4] This low score confirmed our initial reservations about direct backtranslation. We nevertheless translate the monolingual English TED talks to Basque with this system. Table 2 summarizes the data of this experiment.

[3]Note that we employ the term *pivot language* in the context of a data augmentation strategy, not of a machine translation model *per se*. We take a parallel corpus of languages $\langle X, Y \rangle$ and translate its X side to language Z using a $X \rightarrow Z$ machine translation system, yielding a corpus of languages $\langle Z, Y \rangle$. This approach is simpler than the common acceptation of pivot-based translation, where two (more or less independent) translation models are trained, and the output of the first serves as the input of the second one. Examples of recent work in this area include [10, 11].

[4]The BLEU score of a Basque-to-English system including these backtranslations is 21.04, as shown in the second row of Table 5.

3.2. Pivot-based backtranslation of TED talks

We hypothesize that the direct backtranslation approach would not be particularly effective, as the system used to generate them would suffer from the same data sparsity issues as the baseline system (trained with the same data, but in the other direction). In order to take advantage of the other datasets provided by the organizers, we follow a pivot-based approach along the lines of [9]: we take all TED talks available in both English and Spanish (but not Basque), translate the Spanish version to Basque, and align the Basque side with the English side to constitute additional Basque–English data. In this setting, the backtranslation model needs to be trained on Spanish-to-Basque data; using the same 64+10 TED talks for training and validation, as well as the out-of-domain Open Data Euskadi (ODE) dataset and the Basque–Spanish OpenSubtitles (OST), we create a Transformer model with the same parameters as the baseline model. At the end of training, this system obtained a BLEU score of 14.52 on the Spanish-to-Basque development set.

The resulting data consists thus of the same English target sentences as above but different Basque source sentences. Details on the setup are given in Table 3. It is striking that the Basque sentences translated via Spanish are considerably longer than those translated directly from English (4501k total tokens in Table 3 vs. 3886 total tokens in Table 2). The experiments described below will show which of the two datasets improves translation most, and whether the two datasets are complementary or not.

Source	Talks	Lines	ES tokens	EU tokens
TED train	64	5546	124k	98k
OST	—	794k	5.8M	4.8M
ODE	—	927k	23.1M	17.5M
TED dev	*10*	*1122*	*26k*	*20k*
TED pivot-BT	2576	271k	EN 6.2M	**4.5M**

Table 3: Basque–Spanish data used to train the backtranslation model (above the line) and monolingual Spanish data backtranslated to Basque and aligned with English (below the line).

3.3. Pivot-based translation of Open Data Euskadi

Whereas backtranslation yields datasets with noisy source sides and clean target sides, we also wanted to explore the impact of a corpus with clean source side and noisy target side. This approach is not generally used in standard high-resource settings, but could yield additional improvements in low-resource settings. The Open Data Euskadi corpus is a good candidate for this approach. It is rather large and contains Basque–Spanish parallel data. In order to create a Basque–English version of this corpus, we proceed by translating the Spanish version to English and aligning it with the existing Basque one.

The Spanish–English system is trained using most of the parallel data that was made available in WMT 2013, the last year in which Spanish–English featured as a WMT news translation language pair (see Table 4) [12]. In particular, we use the CommonCrawl, Europarl V7, NewsCommentary V12 and UN datasets for training,[5] the NewsTest 2008-2012 corpora for validation and NewsTest 2013 for testing. We did not use OpenSubtitles as we did not find it helpful for translating the legal and news domain documents present in Open Data Euskadi. Due to the larger training corpora sizes, we increased the vocabulary to 40k joint BPE units, but kept the same Transformer architecture and parameters otherwise. This system obtained a BLEU score of 29.69 on the development set and 31.45 on the test set, slightly surpassing the best systems participating in WMT 2013.[6] The figures of the resulting Basque–English Open Data Euskadi corpus are shown on the last line of Table 4.

Source	Lines	ES tokens	EN tokens
CommonCrawl	1845k	49.5M	46.9M
Europarl	1965k	57.0M	54.5M
NewsCommentary	292k	8.5M	7.5M
UN	11196k	366.1M	320.0M
News dev	*13k*	*357k*	*336k*
News test	*3k*	*70k*	*64k*
ODE pivot-T	927k	EU 17.3M	**21.5M**

Table 4: Spanish–English data used to train the translation model (above the line) and monolingual Spanish data translated to English and aligned with Basque (below the line).

3.4. Bilingual systems using synthetic data

We trained various Basque-to-English systems with different combinations of the synthetic datasets described above. All experiments use the same Transformer model architecture, but slightly different vocabulary sizes (see below).

For some experiments, we introduce variants with domain labels [14, 15]. Tars et al. have found domain labeling useful to teach the model about possible domain mismatches in the training data. In our experiments, we use four labels, distinguishing text sources (TED, OST, OPD) and methods of corpus construction (TED-parallel and TED-BT). The validation and test instances are labeled as TED-parallel. Domain labels were included as the first tokens of each sentence. Table 5 summarizes these experiments.

Table 5 shows that any additional synthetic dataset helps in the given low-resource setting. The direct TED backtranslations are surprisingly helpful despite their low quality, although the pivot-based TED backtranslations are much more useful, presumably due to the higher quality of the system that generated them. The impact of the ODE synthetic dataset is less remarkable, but still improves BLEU scores by 2-3 absolute points. Interestingly, the direct and pivot-based TED backtranslations are somewhat complementary, yielding slight improvements compared to using just the pivot-based ones.

Training data	Domain labels	BPE	BLEU
Parallel (= TED train + OST)	No	20k	15.40
+ TED direct-BT	No	20k	21.04
+ TED pivot-BT	No	20k	23.20
+ TED direct-BT + ODE pivot-T	No	30k	23.20
	Yes	30k	23.84
+ TED pivot-BT + ODE pivot-T	No	30k	24.22
	Yes	30k	24.52
+ TED direct-BT + TED pivot-BT	No	30k	24.39
+ ODE pivot-T	Yes	30k	**25.06**

Table 5: Experiments with different combinations of training data.

On the basis of the *Parallel + TED pivot-BT* model (third line of Table 5), we performed a grid search to find the best subword encoding scheme. We used various sizes of joint BPE vocabularies with tied embeddings (10k, 15k, 20k, 25k, 30k, 35,) and various sizes of language-specific BPE vocabularies in conjunction with distinct embeddings (10k, 15k, 20k, 25k, 30k, 35k per language). The difference between the worst and best setting lay at 1.5 BLEU points. The best results were achieved with joint vocabularies and tied embeddings and a total of 25k-30k subword units. The final submissions were made with a joint vocabulary of 30k units, like most experiments presented in Table 5.

Domain labels show consistent improvements of about 0.5 BLEU points. As mentioned above, we labeled the validation data with TED-parallel. Additional experiments using other domain labels at test time have shown the following results: TED-BT +0.04 BLEU, OST -2.83 BLEU, OPD -4.70 BLEU, no label -1.71 BLEU. This experiment shows that the TED-parallel and TED-BT labels yield similar results (the difference is probably not statistically significant), suggesting that the distinction between genuinely parallel and backtranslated TED data may not have been necessary. We nevertheless kept the TED-parallel label also for the test data.

4. Multilingual system

Johnson et al. [14] have shown that multilingual translation models can be trained by using training data of various languages and directions and prepending a target language label to each source sentence. One interesting use case of such multilingual models is zero-shot translation, where the

[5] We experimented with a reduced training set consisting of Europarl and NewsCommentary only, but results were not quite as good as with the complete training data.

[6] The best WMT 2013 submissions were the phrase-based statistical systems by the University of Edinburgh team, with BLEU scores of 31.37 in the unconstrained setting and 30.59 in the constrained setting [13].

System	Model type and features		BLEU	NIST	TER
Primary	Bilingual model	With sentence splitting	25.01	6.45	59.48
Contrastive 1	Bilingual model	No sentence splitting	**25.25**	**6.47**	**58.83**
Contrastive 2	Multilingual model	With sentence splitting	22.55	6.10	60.48

Table 6: Submitted systems and official results on the test set.

source language and target language have both been seen by the model, but not in that particular combination. In our case, we are not interested in zero-shot translation, as we do have a sizeable set of Basque-to-English training data. Rather, we wanted to see to what extent multilingual modelling could supplant the creation of synthetic data. To this end, we train a single multilingual model with the following datasets: the parallel Basque–English TED and OpenSubtitles data (as in the baseline model), the parallel English–Spanish TED data in both directions (as used to train the pivot-based backtranslation model), and the Basque–Spanish Open Data Euskadi data (see Table 7). In this setting, we only have English and Spanish as target languages and consequently only use the two target language labels TO_EN and TO_ES. We do not use additional domain labels in this experiment. The model architecture remains the same, but we use a joint trilingual vocabulary consisting of 40k BPE units.

Source	Lines	Source tokens	Target tokens
TED train	5623	97k EU	128k EN
OST	805k	4.8M EU	6.5M EN
TED train	277k	6.3M EN	6.0M ES
TED train	277k	6.0M ES	6.3M EN
ODE	926k	17.5M EU	23.1M ES
TED dev	*1140*	*20k EU*	*27k EN*

Table 7: Data used to train the multilingual model.

Although we used almost the same datasets as in the systems presented above (with the exception of the WMT English–Spanish data), the multilingual model failed to achieve competitive results, with 22.55 BLEU on the validation set. There are several reasons for this lower-than-expected performance. First, the training of the multilingual model was stopped before convergence, after a training time of 72 hours. Nevertheless, the learning curve does not indicate the possibility of substantial improvements if training had continued. Second, the multilingual model has to learn three languages on the source side and two languages on the target side instead of a one-to-one mapping. Its task is thus inherently more complex, and it seems that the three languages in question (Basque, English and Spanish) are typologically too diverse for the model to generalize. Finally, [14] show that good data sampling strategies are crucial when training multilingual models with unbalanced data sizes. In this regard, oversampling the Basque-to-English resources or fine-tuning the model to the target language pair might have helped. Despite its lower performance, we base one of our contrastive submissions on the multilingual model.

5. Submissions

We decided to submit output from two models, the bilingual system trained with all synthetic data and domain labels (last line of Table 5), and the multilingual system described in Section 4.

We have found in a different context [16] that systems trained on single sentences may not be able to translate utterances consisting of several sentences completely. Although there was no particular evidence of such problems occurring in the experiments at hand (since a large portion of the TED training data already contains multi-sentence utterances), we carried out some experiments on this issue. Concretely, we applied a simple sentence splitter to the source text, translated each sentence separately, and merged them back together. In the validation set, 214 (of 1140) lines were split, and sentence splitting improved the BLEU score by 0.26 points. However, qualitative inspection of the results did not show convincing evidence in favor or against sentence splitting. Therefore, we submitted systems with and without sentence splitting.

Also, due to an error in the postprocessing script, the submitted translations were accidentally detokenized with the Basque detokenizer (and some additional rules) rather than the English one. The added rules minimized the adverse effect of this error, such that it only affected two tokens in the test set, resulting in an estimated impact on BLEU score of about 0.01.

Table 6 summarizes the submitted systems with the official results. Sentence splitting turned out to have a slightly negative impact on the translation of the test set, whereas the difference between the bilingual and multilingual system is comparable to the one that was observed with the validation set.

6. Conclusions

The University of Helsinki submissions on Basque–English leverage the existing parallel corpora for other language pairs to create synthetic data of various types. In particular, we have found pivot-based (back-)translation to be a useful approach to increase the amounts of Basque–English training data. In this setting, one side of a parallel corpus is translated to a third language, and this translated output is then aligned with the other side of the original parallel corpus. By using

various synthetic datasets, we were able to increase translation performance from 14.68 BLEU to 25.06 BLEU on the development set.

Our contrastive multilingual model performed less well, although it saw almost the same data as the bilingual model and its auxiliary models used to create the synthetic data. It remains to be seen if better balancing of the training data, possibly including some fine-tuning, as well as the inclusion of domain labels and additional Spanish–English training data could make this model more competitive. Also, both approaches could be combined by training a multilingual model with added synthetic data.

7. References

[1] M. Cettolo, C. Girardi, and M. Federico, "Wit^3: Web inventory of transcribed and translated talks," in *Proceedings of the 16th Conference of the European Association for Machine Translation (EAMT)*, Trento, Italy, May 2012, pp. 261–268.

[2] J. Tiedemann, "Parallel data, tools and interfaces in OPUS," in *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, N. Calzolari, K. Choukri, T. Declerck, M. U. Doğan, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, and S. Piperidis, Eds. Istanbul, Turkey: European Language Resources Association (ELRA), May 2012.

[3] P. Lison and J. Tiedemann, "OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles," in *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, N. Calzolari, K. Choukri, T. Declerck, S. Goggi, M. Grobelnik, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, and S. Piperidis, Eds. Paris, France: European Language Resources Association (ELRA), May 2016.

[4] P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst, "Moses: Open source toolkit for statistical machine translation," in *ACL'07: Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics, Demonstration Session*. Association for Computational Linguistics, 2007, pp. 177–180.

[5] 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*, 2017, pp. 5998–6008.

[6] M. Junczys-Dowmunt, R. Grundkiewicz, T. Dwojak, H. Hoang, K. Heafield, T. Neckermann, F. Seide, U. Germann, A. Fikri Aji, N. Bogoychev, A. F. T. Martins, and A. Birch, "Marian: Fast neural machine translation in C++," in *Proceedings of ACL 2018, System Demonstrations*. Melbourne, Australia: Association for Computational Linguistics, July 2018, pp. 116–121.

[7] R. Sennrich, B. Haddow, and A. Birch, "Neural machine translation of rare words with subword units," 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. 1715–1725.

[8] ——, "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.

[9] J. Tiedemann, "Character-based pivot translation for under-resourced languages and domains," in *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, 2012, pp. 141–151.

[10] Y. Cheng, Y. Liu, Q. Yang, M. Sun, and W. Xu, "Neural machine translation with pivot languages," *CoRR*, vol. abs/1611.04928, 2016.

[11] S. Ren, W. Chen, S. Liu, M. Li, M. Zhou, and S. Ma, "Triangular architecture for rare language translation," in *Proceedings of ACL 2018*, Melbourne, Australia, 2018.

[12] O. Bojar, C. Buck, C. Callison-Burch, C. Federmann, B. Haddow, P. Koehn, C. Monz, M. Post, R. Soricut, and L. Specia, "Findings of the 2013 Workshop on Statistical Machine Translation," in *Proceedings of the Eighth Workshop on Statistical Machine Translation*. Sofia, Bulgaria: Association for Computational Linguistics, August 2013, pp. 1–44.

[13] N. Durrani, B. Haddow, K. Heafield, and P. Koehn, "Edinburgh's machine translation systems for European language pairs," in *Proceedings of the Eighth Workshop on Statistical Machine Translation*. Sofia, Bulgaria: Association for Computational Linguistics, August 2013, pp. 114–121.

[14] M. Johnson, M. Schuster, Q. Le, M. Krikun, Y. Wu, Z. Chen, N. Thorat, F. Viégas, 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.

[15] S. Tars and M. Fishel, "Multi-domain neural machine translation," in *Proceedings of the 21st Annual Conference of the European Association for Machine Translation (EAMT'2018)*, J. A. Pérez-Ortiz, F. Sánchez-Martínez, M. Esplà-Gomis, M. Popović, C. Rico, A. Martins, J. V. den Bogaert, and M. L. Forcada, Eds., 2018.

[16] A. Raganato, Y. Scherrer, T. Nieminen, A. Hurskainen, and J. Tiedemann, "The University of Helsinki submissions to the WMT18 news task," in *Proceedings of the Third Conference on Machine Translation (WMT18)*. Association for Computational Linguistics, 2018.

8. Appendix

All models presented in this paper were trained using the parameter settings described in `https://github.com/marian-nmt/marian-examples/tree/master/transformer`, which correspond roughly to the base setup of [5].

The relevant parameters are as follows:

```
marian --type transformer
--max-length 200 --mini-batch-fit
-w 10000 --maxi-batch 1000
--early-stopping 10 --valid-freq 5000
--valid-metrics cross-entropy
perplexity translation
--valid-mini-batch 64 --beam-size 6
--normalize 0.6 --enc-depth 6
--dec-depth 6 --transformer-heads 8
--transformer-postprocess-emb d
--transformer-postprocess dan
--transformer-dropout 0.1
--label-smoothing 0.1
--learn-rate 0.0003 --lr-warmup 16000
--lr-decay-inv-sqrt 16000
--optimizer-params 0.9 0.98 1e-09
--clip-norm 5 --tied-embeddings-all
--sync-sgd --exponential-smoothing
--seed 1111
```