The Fujitsu DMATH submissions for WMT21 News Translation and Biomedical Translation Tasks

Ander Martínez
Fujitsu, Ltd.
4-1-1, Kamikodanaka, Nakahara-ku,
Kawasaki 211-8588, Japan
ander@fujitsu.com

Abstract

This paper describes the Fujitsu DMATH systems used for WMT 2021 News Translation and Biomedical Translation tasks. We focused on low-resource pairs, using a simple system. We conducted experiments on English-Hausa, Xhosa-Zulu and English-Basque, and submitted the results for Xhosa→Zulu in the News Translation Task, and English→Basque in the Biomedical Translation Task, abstract and terminology translation subtasks. Our system combines BPE dropout, sub-subword features and back-translation with a Transformer (base) model, achieving good results on the evaluation sets.

1 Introduction

WMT has been exploring the state of the art in MT for many years, and, particularly in recent editions, the participants have shown impressive results. However, often times, these results require very heavy or complex systems, trained on dozens of GPUs. Participants compete for a margin that places them above the rest, combining multiple methods from the latest research.

In recent years, different variants of the Transformer (Vaswani et al., 2017) architecture have been popular for NMT, so can be seen when inspecting the submissions to previous editions of WMT. In our systems, we use the Transformer `base` configuration, the smaller one. Our implementation is based on Sockeye 2 (Hieber et al., 2020; Domhan et al., 2020).

We combine several techniques or strategies for low-resource pairs. These techniques are described in Section 2.

We conducted a few experiments on language pairs Xhosa-Zulu and English-Hausa, from the News Transltion task, and on English-Basque, from the Biomedical Translation task. The results of our experiments are shown in Section 3.

2 Techniques

This section describes the strategies used for our NMT models. The first two, *bpe dropout* and *sub-subword features*, were used in all the subtasks, while the last one was only used for the biomedical translation subtasks.

2.1 BPE dropout

BPE dropout (Provilkov et al., 2020) was introduced as an alternative to Kudo (2018). Provilkov et al. found that the main drawback to the subword regularization method is its complexity, since it requires training a unigram language model and uses uses EM and Viterbi algorithms to sample segmentations.

BPE dropout works on BPE vocabulary models (Sennrich et al., 2016b), that is, the vocabularies are built in the same way as vanilla BPE. While the unigram language model subword regularization method uses a statistical model and dynamic programming to be able to sample different segmentations from the same sequence, BPE dropout uses random noise to discard certain merges, randomly generating a different sequence of subwords each time. This is so because BPE does not store the frequencies of each subword, only the order of the merges. Merges are discarded with a probability p, which is usually 0.1. Provilkov et al. concluded through several experiments that BPE dropout achieves better results.

Our systems use BPE dropout during training, with a dropout proability p of 0.1.

2.2 Sub-subword features

The main idea of the Sub-subword feature method (Martinez et al., 2021) is to build the embedding matrices from the n-gram features of the subwords in the vocabulary. The features used to produce the embeddings are selected by an algrithm before training, and the neural network that produces the embeddings is trained with the rest of the model.

Proceedings of the Sixth Conference on Machine Translation (WMT), pages 162–166
November 10–11, 2021. ©2021 Association for Computational Linguistics

	Sentences	Words in source	Words in target	Word ratio
Xh-Zu	94,323	1,356,127	1,325,168	1.02
En-Ha	752,287	11,044,101	11,713,109	1.06
En-Eu	2,627,745	23,225,786	17,472,145	1.33

Table 1: Statistics of the datasets used for BT experiments. The rows of the table are ordered from smallest to largest, the source language being that of the pair. The ratios are the number of words of the largest language compared to the other.

The method has a regularizing effect, particularly effective under low-resource settings. The sub-subword feature method can be used with BPE and BPE dropout, to achieve better results.

2.3 Back-Translation

Back-translation (BT) (Sennrich et al., 2016a) can be used with monolingual data of the target language, to improve low-resource language pair performance. BT is a type of distant supervision, in which a model of the opposite direction to the one that one wants to build is used to synthesize more parallel data. The method requires training an opposite model, and the synthesized data is noisy. Still BT has been used extensively with good results reported (Poncelas et al., 2018; Edunov et al., 2018).

The effectiveness of the BT method depends largely on the quality of the monolingual corpora used. Monolingual corpora compiled automatically using web crawlers in combination with automatic language detection are prone to be noisy. Particularly for low-resource languages for which language detection has lower accuracy.

For example, we noted that the Hausa Extended Common Crawl corpus published for WMT21 contained a large number of Japanese song lyrics written in Latin alphabet.

Our systems used 2 million backtranslated sentences to improve performance.

2.4 Multilingual model

Johnson et al. (2017) introduced multilingual models to NMT. Multilingual models are capable of translating more than one pair. For this, they used a simple approach that consists of using a special symbol inserted in the source sentence, indicating the target language. The architecture of the model can be the same as that of non-multilingual models. In their experiments, they showed that, although the performance of pairs with more resources worsens when sharing a model with other pairs, the performance of pairs with fewer resources improves. Multilingual models allow translation between pairs with zero resources. This is known as zero-shot translation.

Much research has been done on Multilingual Neural Machine Translation (MNMT). Dabre et al. (2020) published a comprehensive survey that summarizes different ideas and techniques for MNMT.

For the English-Basque Biomedical task, we tried using multilingual models too. In particular, for the terminology translation subtask, we included the English-Spanish terminology from MeSpEN (Villegas et al., 2018). The terminology was included as training data, using the method described in this section. A more sophisticated vocabulary integration method could have given better results (Post and Vilar, 2018; Bergmanis and Pinnis, 2021).

3 Experiments

We conducted experiments on Xhosa → Zulu, Zulu → Xhosa, English → Hausa, Hausa → English and English → Basque. Notice that the WMT21 Biomedical Translation Task for English-Basque was only in the English → Basque direction, and not Basque → English.

Table 1 shows the statistics for three language pairs. The rows are ordered from smallest to largest. The Xhosa-Zulu and English-Hausa data were published in the WMT21 news translation task. Both are classified as low-resource in the task description, but Xhosa and Zulu are two closely-related languages, and English and Hausa, two distant languages. The English-Basque data were published for the biomedical task of WMT21. The English-Basque dataset cannot be considered low-resource, with 2.6M parallel sentences, but it represents two distant languages. The Basque language has a complex morphology that makes its generation difficult.

Word ratios can hint about the similarity or dissimilarity of the languages. Xhosa and Zulu are related languages, and that is why they show a ra-

	Xh→Zu	**Zu→Xh**	**En→Ha**	**Ha→En**	**En→Eu**
Baseline	6.5 (.416)	6.3 (.421)	12.0 (.412)	13.0 (.403)	16.5 (.456)
+SSWF	9.3 (.470)	8.5 (.468)	12.5 (.420)	14.5 (.429)	**17.3** (.471)
+BT	9.2 (.471)	8.6 (.467)	17.5 (.480)	**16.7** (.460)	16.4 (.462)
+BT+SSWF	**9.7** (.478)	**8.8** (.470)	**18.0** (.482)	15.5 (.461)	16.4 (.463)

Table 2: BT results for various language pairs. (+/-) BT indicates the use or non-use of BT data. The results follow the format "BLEU (CHRF2)". Best BLEU results are shown in bold and the best CHRF2 are underlined.

tio close to one. English and Hausa are distant languages, but their morphological characteristics result in sequences of similar length.

Table 3 shows the hyperparameters used to train the models. The Transformer hyperparameters are those of the *base* model. We use a relatively large vocabulary size of 32k subwords. Although Sennrich and Zhang (2019) showed that smaller vocabularies give better results on low-resource datasets, larger vocabularies work well when using sub-subword features (Martinez et al., 2021).

We used 4,000 warmup steps schedule as described in Vaswani et al. (2017) with an initial learning rate of 2.0 and evaluated the development cost every 2,000 updates. The model was reloaded from the best checkpoint when the development cost did not improve, and training stopped after 3 consecutive stallings.

Hyperparameter	**Value**
Vocabulary size	32,000 subwords
BPE dropout p	0.1
Batch size	4,096 (×2 GPUs)
Warmup steps	4,000
Learning rate	2.0
Encoder layers	6
Decoder layers	6
Attention heads	8
Transformer size	512
Hidden layer size	2,048
Dropout	0.1
Label smoothing ϵ	0.1
FTE layers †	3
FTE size †	3,072

Table 3: Hyperparameters used in our models. † FTE (feature-to-embedding) network size for sub-subword feature (+SSWF) models.

For the *News Translation Task* participants need agree to contribute to the manual evaluation about eight hours of work, per system submission. In consideration of this workload, we decided to submit only the Xhosa → Zulu system to the News Translation Task.

Table 2 shows the results for the languages in Table 1. The BT data were translated using the sub-subword feature (+SSWF) model. The BT data contain 2 million pairs of sentences. The English-Basque model shown in this table does not use the multilingual approach described in Subsection 2.4.

The results show that the sub-subword features (+SSWF) improve the results of the corresponding −SSWF models under low-resource settings. In the case of Hausa → English, the +SSWF system did not achieve better BLEU scores than the corresponding −SSWF system, but achieved better CHRF2.

Despite its noisy nature, we decided to use the *Extended Common Crawl* Hausa corpus. The results show that the data, although noisy, was effective in improving the performance.

The English → Basque biomedical abstract translation did not improve when using back-translation data. It is possible that the cause for this was the domain mismatch of the monolingual data, that was not exclusively from scientific papers' abstracts.

All models were trained on two *NVIDIA Tesla P100* GPUs. The Xhosa-Zulu models are trained in about 2.5 hours, and the English-Hausa models are trained in about 10 hours.

Table 4 shows the result of combining the English-Basque training data with the MeSpEN English-Spanish terminology (Villegas et al., 2018). The MeSpEN terminology dictionary that we used contained 125,519 term pairs after cleaning.

Model	**BLEU**	**chrF2**
En→Eu	16.47	.456
En→Eu +SSWF	17.34	**.471**
En→ {Eu,Es} +SSWF	**17.44**	.470

Table 4: NMT result of combining the English-Basque training data with the MeSpEN English-Spanish terminology.

The scores displayed were obtained by evaluating the trained models on a test set sampled from the provided data for abstract translation. The data used to build the development and test sets were removed from the training data. The results show the BLEU and CHRF2 scores for abstract translation, but we did not prepare any evaluation set for terminology translation, as we wanted to include WMT20 terminology in the training data.

The same models were used for abstract translation and terminology translation. Manual examination of the produced transations hinted better performance for the the model trained with English-Spanish terminology.

In consideration of the results, we decided to submit two systems to the abstract translation and terminology translation subtasks. One of the systems incorporated the MeSpEN terminology, and the other one did not. Both systems did not use backtranslated data.

4 Conclusions

We built and submitted three lightweight systems that used sub-subword features to build the embeddings. We evaluated the approach with different configurations and the results showed the adequacy of the approach.

The relatively small models could possibly use larger hyperparameters and other techniques to achieve better results, but we think the current results can show the strenght of the techniques that were applied.

References

Toms Bergmanis and Mārcis Pinnis. 2021. Facilitating terminology translation with target lemma annotations. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3105–3111, Online. Association for Computational Linguistics.

Raj Dabre, Chenhui Chu, and Anoop Kunchukuttan. 2020. A survey of multilingual neural machine translation. *ACM Comput. Surv.*, 53(5).

Tobias Domhan, Michael Denkowski, David Vilar, Xing Niu, Felix Hieber, and Kenneth Heafield. 2020. The sockeye 2 neural machine translation toolkit at AMTA 2020. In *Proceedings of the 14th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pages 110–115, Virtual. Association for Machine Translation in the Americas.

Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. Understanding back-translation at scale. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 489–500, Brussels, Belgium. Association for Computational Linguistics.

Felix Hieber, Tobias Domhan, Michael Denkowski, and David Vilar. 2020. Sockeye 2: A toolkit for neural machine translation. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 457–458, Lisboa, Portugal. European Association for Machine Translation.

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

Taku Kudo. 2018. Subword regularization: Improving neural network translation models with multiple subword candidates. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 66–75, Melbourne, Australia. Association for Computational Linguistics.

Ander Martinez, Katsuhito Sudoh, and Yuji Matsumoto. 2021. Sub-subword n-gram features for subword-level neural machine translation. *Journal of Natural Language Processing*, 28(1):82–103.

Alberto Poncelas, Dimitar Shterionov, Andy Way, Gideon Maillette de Buy Wenniger, and Peyman Passban. 2018. Investigating backtranslation in neural machine translation.

Matt Post and David Vilar. 2018. Fast lexically constrained decoding with dynamic beam allocation for neural machine translation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1314–1324, New Orleans, Louisiana. Association for Computational Linguistics.

Ivan Provilkov, Dmitrii Emelianenko, and Elena Voita. 2020. BPE-dropout: Simple and effective subword regularization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1882–1892, Online. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016a. 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.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016b. 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)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Rico Sennrich and Biao Zhang. 2019. Revisiting low-resource neural machine translation: A case study. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 211–221, Florence, Italy. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Marta Villegas, Ander Intxaurrondo, Aitor Gonzalez-Agirre, Montserrat Marimon, and Martin Krallinger. 2018. The mespen resource for english-spanish medical machine translation and terminologies: Census of parallel corpora, glossaries and term translations. *Language Resources and Evaluation*.