High frequent in-domain word segmentation and forward translation for the WMT21 Biomedical task

Bardia Rafieian and Marta R. Costa-jussà
TALP Research Center, Universitat Politècnica de Catalunya, Barcelona
{bardia.rafieian,marta.ruiz}@upc.edu

Abstract

This paper reports the optimization of using the out-of-domain data in the Biomedical translation task. We firstly optimized our parallel training dataset using the BabelNet in-domain terminology words. Afterward, to increase the training set, we studied the effects of the out-of-domain data on biomedical translation tasks, and we created a mixture of in-domain and out-of-domain training sets and added more in-domain data using forward translation in the English-Spanish task. Finally, with a simple bpe optimization method, we increased the number of in-domain subwords in our mixed training set and trained the Transformer model on the generated data. Results show improvements using our proposed method.

1 Introduction

Domain adaptation is one of the known challenges in Machine Translation since NMT(neural machine translation) models are susceptible to the training data (Koehn and Knowles, 2017). To say, NMT models perform poorly for domain-specific translation when trained on large out-resource data (Chu and Wang, 2018). As a result, due to the limitations of specific domain data, domain adaptation strategies help NMT models by increasing the parallel corpora. There have been several tasks to address domain adaptation which recently, in (Sato et al., 2020) they proposed a vocabulary adaptation to fine-tune the embedding layers of the NMT model by projecting general word embeddings induced from monolingual data in a target domain onto a source-domain embedding space to improve translation score. On the other hand, augmenting bilingual training data with forwarding and backward translation improves the in-domain translation quality (Nayak et al., 2020). Inspired by mentioned ideas, in this work, we implemented our strategy by two essential steps: 1) collecting and augmenting data by forwarding translation and then tuning it using Babelnet to include biomedical sentences 2) Implementing subwords bpe optimization on the train set to study the adaptation of out-of-domain data in the biomedical task. After that, We selected the transformer model (Vaswani et al., 2017) to train our system in different experimental settings. The remainder of the paper is organized as follows. In Sec. 2 we describe data collection and preparation. Sec. 3 explains our bpe optimization strategy to adapt out-of-domain data in the biomedical task. Sec. 4 shows our experimental setups and evaluation results, and finally, we conclude and discuss future works in the Sec. 5.

2 Data production

One of the critical topics in machine translation (MT) is selecting and fitting well-organized domain-relevant data (Wang et al., 2018). This section describes our data preparation approach to tune, clean, and optimize data for our translator model. The details of the dataset are described in the section 4.

2.1 In-domain dataset tuning

The gathered in-domain data is not well-tuned for the biomedical domain, so that we extracted a list of biomedical terms(word level) using the BabelNet API (Navigli and Ponzetto, 2012) by referring to the ”biomedical” tags in the BabelNet: bio-science, technology, medical practice, medical specialty, neurology, and orthopedics. To address it, we gathered a total of 5,800 biomedical terms for both English and Spanish languages. Secondly, we selected the sentences which specifically contain biomedical words. The outcome holds in-domain parallel data which each sentence at least carries a related biomedical term. Algorithm 1 shows our approach to select in-domain sentences.

Proceedings of the Sixth Conference on Machine Translation (WMT), pages 863–867
November 10–11, 2021. ©2021 Association for Computational Linguistics

```
Result: $indomain\ parallel\ dataset$
dataset-tuning;
initialization;
$input(EN\ bio\ words, ES\ bio\ words)$;
$input(standard\ en\ es\ parallel)$
  $init(opt\ en\ es\ parallel)$;
for
  $sentence\ 1\ and\ sentence\ 2\ in\ standard\ en\ es\ parallel:$
  do
    if($any\ token\ sentence\ 1\ in\ (EN\ bio\ words)$)
    and ($any\ token\ sentence\ 2\ in\ (ES\ bio\ words)$) :
    $OptimizedEnEsParallel.append$
    ($sentence\ 1\ and\ sentence\ 2$)
end
$return(OptimizedEnEsParallel)$
```

Algorithm 1: Optimizing the parallel corpus using BabelNet; selecting sentences that contain at least one token in-domain word

2.2 In-domain forward translation

Considering a translation task of L1 → L2, where L1 has more significant monolingual data than L2, a forward translation translates the L1 to L2 and uses the translated L2 to recreate a synthetic parallel corpus. It has been widely reported that forward and back translation brings significant results. (Bogoychev and Sennrich, 2019). We benefited from this fact and produced bilingual data from the English source, which did not have any target or good target parallel translation. However, to ensure the availability of in-domain data, we first passed the previous step on the available monolingual side. Then we translated the source side using our MT model and added bilingual data for retraining. Finally, we merged the in-domain and out-of-domain parallel corpus to achieve a bigger train set.

3 Subword BPE optimization

Byte Pair Encoding, or BPE, is a subword segmentation algorithm that encodes rare and unknown words as sequences of subword units by merging the most frequent consecutive byte pair into a new subword (Sennrich et al., 2015). Since we enriched the train set with out-of-domain data, We propose "bpe-terms in-domain optimization" to handle open vocabulary problems and enhancing the morphology when out-of-domain data is available. Consequently, increasing the frequency of in-domain words in the subword bpe training raises the chance of having in-domain words in the vocabulary. As a result, out-of-domain data will not affect the quality of the model on translating the in-domain words, while they let the model learn on an enormous corpus. We performed this strategy by first learning the subwords on 10x duplicated in-domain parallel sentences with a size of eight million mixed with smaller out-of-domain corpora (no duplication) and then applying the trained subword model on the standard-sized corpus. After that, we expect to have the biomedical in-domain words directly translated to the target language without breaking them into subwords.

4 Experiments

Experiments illustrated in this section study the effects of the out-of-domain data on in-domain(biomedical) translation task as well as the possibility of adapting it by performing a tuned subword-bpe segmentation algorithm 3 to improve the translation quality. We split this section into four parts which start with data collection and prepossessing. Then, we describe the training system and, finally, the evaluation scores of the competition.

4.1 Data collection

We rely on the WMT21 official webpage to collect the (en/es) parallel in-domain data. Out of the provided resources, in particular, for the in-domain train set, we selected UFAL, Pubmed, Medline, IBECS (Villegas et al., 2018) and UNcorpus (Ziemski et al., 2016) along with the OPUS collection (Tiedemann, 2012). Next, we cleaned the data by removing empty lines, duplicates, and very short and long sentences. Also, to perform our experiments on out-of-domain data, we collected the parallel sentences provided from the same WMT21 official website.

4.2 Data preprocessing

To prepare our data for training, we followed the standard pipelines by performing normalization, tokenization, and removing words that contain non-alphabetic characters using Moses (Koehn et al., 2007). Then, we removed concise and long sentences by keeping the thresh-hold between 2 and 30 words for each sentence and implemented the strategy described in section 2 to select in-domain sentences. As a report, we collected 6,855,049 in-domain and added 1,965,824 out-of-domain parallel data (English/Spanish). We also translated 1,558,834 in-domain UFAL monolingual English data to Spanish and added it to our bilingual corpus for retraining the en/es model.

4.3 Training on optimized segmented data

Our method focuses on data preparation and investigates how the out-of-domain data affects the BLEU score. We imply that tuning the vocabulary of subwords would improve the accuracy of the in-domain translation(biomedical) even though some of the data is out of the domain.

The two crucial factors applied in our experiments are preprocessing the parallel corpus with BabelNet, and tuning the learning step of subwords to adapt out-of-domain data. following the strategy, four experiments have been done with two different trainsets, in-domain and mixture of in-domain and out-of-domain data:

1. In the first experiment, we used word-level data in both the source and target sides to evaluate the impact of out-of-domain usage in an in-domain task.

2. In the second experiment, we applied subword-bpe level on both source and target side with shared embeddings; however, the data were preprocessed by using Babelnet (described in section 2) to adjust the in-domain sentences in the train set for all the experiments.

3. We used the same strategy as the second experiment but with applying BPE-dropout (Provilkov et al., 2019) on both the source and target side of the data.

4. The last experiment was carried out by using tuned in-domain subword level data on both source and target sides as explained in the section 3.

In all experiments, we trained baselines on word-level and subword-bpe level to measure the proposed methods.

We selected a vocabulary size of 50k tokens and trained the data by the Transformer model with its default parameters using Open-nmt (Klein et al., 2017) neural machine translation framework.

4.4 Evaluation and results

The evaluation has been done on WMT18 and WMT19 test sets based on the BLEU score. We compared the trained models with word-level, standard subword bpe level, bpe drop out and tuned subword bpe level of the parallel corpus in the trainset to follow our experiments. We also studied the results with three types of trainsets:

- in-domain
- fair mixture of in-domain and out-of-domain sentences
- an unfair mixture of in-domain and out-of-domain with more in-domain sentences

We started and continued each training until it accomplished the best BLEU score on the validation set. We realized that using bpe dropout in the trainset gives worse results than the standard bpe level in terms of the BLEU score. Also, as expected, the worst results belong to word level and hybrid wordlevel+subword level trainset. On the other hand, using out-of-domain data in an in-domain task caused a dramatic drop in the BELU score. In this regard, there was a slight improvement in BLEU score by increasing the frequency of biomedical words in the mixture of in-domain and out-of-domain trainset in both fair and unfair distribution of each domain sentence. For WMT21 competition, we selected the models which achieved the highest scores in the wmt18 and wmt19 en2es and es2en test sets.

Table 1 describes our (en2es) results on a mixture of 2.7 million in-domain + 1.7 million out-of-domain parallel sentences (described the data in the section 2). As well, Table 2 shows the results on 2.7 million in-domain parallel sentences and also a mixture of 8 million in-domain + 1.7 million out-of-domain parallel data (all of that data). Similarly, we show the (es2en) results in the tables 3 and 4

5 Conclusion and future works

This work presented a method to adapt out-of-domain data in an in-domain(biomedical) task to improve the BLEU score. We tuned the parallel data by BabelNet, then found and increased the frequency of biomedical words in subword-learning to raise the weight of in-domain words in the vocabulary. Our results obtained in a different mixture of datasets show that our method improves the BLEU score compared with the standard subword-bpe approach. In the future, we plan to extend our approach to more low-resource languages and domains. Moreover, we plan to increase out-of-domain data and configure the frequency of in-domain words based on the domain type.

Acknowledgements

This work was supported by the project ADAVOICE, PID2019-107579RB-I00 / AEI / 10.13039/501100011033.

Dataset: 2.7m indomain+ 1.7m out-of-domain		
EXP type	**wmt18**	**wmt19**
Word level indomain+out-of-domain	35.0	36.6
Word level Indomain+ subword level out-of-domain	34.5	36.1
Subword level indomain+ subword level out-of-domain (baseline)	35.6	42.4
10x freq subword indomain+subword out-of-domain (our approach)	**39.8**	**42.7**
bpe dropout indomain + bpe dropout out-of-domain	38.5	41.9

Table 1: en2es BLEU score results on hybrid dataset using different word segmentation approaches, word level, hybrid, standard bpe, bpe dropout and tuned subword bpe

	Dataset: 2.7m indomain		**Dataset: 8m in + 1.7m out**	
EXP type	**wmt18**	**wmt19**	**wmt18**	**wmt19**
subword bpe in domain (baseline)	39.8	42.1	**40.1**	42.8
10x freq subwords indomain (our approach)	**39.9**	**42.2**	39.2	**43.0**
bpe dropout	39.7	39.2	37.1	41.7

Table 2: en2es BLEU score results on solid indomain and eight million hybrid datasets using different word segmentation approaches, word level, hybrid, standard bpe, bpe dropout and tuned subword bpe

Dataset: 2.7m indomain+ 1.7m out-of-domain		
EXP type	**wmt18**	**wmt19**
Word level indomain+out-of-domain	NA	NA
Word level Indomain+ subword level out-of-domain	NA	NA
Subword level indomain+ subword level out-of-domain (baseline)	38.1	43.23
10x freq subword indomain+subword out-of-domain (our approach)	**39.6**	**43.3**

Table 3: es2en BLEU score results on hybrid dataset using different word segmentation approaches, word level, hybrid, standard bpe, bpe dropout and tuned subword bpe

	Dataset: 2.7m indomain		**Dataset: 8m in + 1.7m out**	
EXP type	**wmt18**	**wmt19**	**wmt18**	**wmt19**
subword bpe in domain (baseline)	**42.1**	**44.0**	**43.0**	44.1
10x freq subwords indomain (our approach)	41.9	43.6	42.3	44.1

Table 4: es2en BLEU score results on hybrid indomain+out-of-domain dataset and unfair distribution.

References

Nikolay Bogoychev and Rico Sennrich. 2019. Domain, translationese and noise in synthetic data for neural machine translation. *CoRR*, abs/1911.03362.

Chenhui Chu and Rui Wang. 2018. A survey of domain adaptation for neural machine translation. *CoRR*, abs/1806.00258.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander Rush. 2017. OpenNMT: Open-source toolkit for neural machine translation. In *Proceedings of ACL 2017, System Demonstrations*, pages 67–72, Vancouver, Canada. Association for Computational Linguistics.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic. Association for Computational Linguistics.

Philipp Koehn and Rebecca Knowles. 2017. Six challenges for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 28–39, Vancouver. Association for Computational Linguistics.

Roberto Navigli and Simone Paolo Ponzetto. 2012. Babelnet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193:217–250.

Prashant Nayak, Rejwanul Haque, and Andy Way. 2020. The ADAPT's submissions to the WMT20 biomedical translation task. In *Proceedings of the Fifth Conference on Machine Translation*, pages 841–848, Online. Association for Computational Linguistics.

Ivan Provilkov, Dmitrii Emelianenko, and Elena Voita. 2019. Bpe-dropout: Simple and effective subword regularization. *CoRR*, abs/1910.13267.

Shoetsu Sato, Jin Sakuma, Naoki Yoshinaga, Masashi Toyoda, and Masaru Kitsuregawa. 2020. Vocabulary adaptation for distant domain adaptation in neural machine translation. *CoRR*, abs/2004.14821.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015. Neural machine translation of rare words with subword units. *CoRR*, abs/1508.07909.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in opus. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. European Language Resources Association (ELRA).

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *CoRR*, abs/1706.03762.

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

Wei Wang, Taro Watanabe, Macduff Hughes, Tetsuji Nakagawa, and Ciprian Chelba. 2018. Denoising neural machine translation training with trusted data and online data selection. *CoRR*, abs/1809.00068.

Michał Ziemski, Marcin Junczys-Dowmunt, and Bruno Pouliquen. 2016. The United Nations parallel corpus v1.0. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3530–3534, Portorož, Slovenia. European Language Resources Association (ELRA).