Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 189 (2021) 136–141

Procedia Computer Science

www.elsevier.com/locate/procedia

5th International Conference on AI in Computational Linguistics

Impact of Filtering Generated Pseudo Bilingual Texts in Low-Resource Neural Machine Translation Enhancement: The Case of Persian-Spanish

Benyamin Ahmadnia[a,*], Bonnie J. Dorr[b], Raul Aranovich[a]

[a]*Department of Linguistics, University of California at Davis, United States*
[b]*Florida Institute for Human and Machine Cognition (IHMC), Ocala, United States*

Abstract

Although the Neural Machine Translation (NMT) framework has already been shown effective in large training data scenarios, it is less effective for low-resource conditions. To improve NMT performance in a low-resource setting, we extend the high-quality training data by generating a pseudo bilingual dataset and then filtering out low-quality alignments using a quality estimation based on back-translation. We demonstrate that our approach yields significantly higher BLEU scores than those of a set of baselines.

Keywords: computational linguistics; natural language processing; neural machine translation; low-resource languages

1. Introduction

A large-scale bilingual training dataset is an essential resource for training Neural Machine Translation (NMT) systems. Creating high-quality large-scale bilingual training texts requires time, financial support, and professional experts to translate a large amount of texts between two natural languages. As a result, many existing large-scale bilingual corpora are limited to specific languages and domains. In contrast, large monolingual corpora are easier than bilingual corpora to obtain. Although state-of-the-art approaches have generally employed monolingual datasets to generate pseudo bilingual texts [1, 2], such approaches do not fully exploit the training data with an eye toward the quality of the generated texts. Many approaches have thus employed monolingual data for the target language to learn the language model more effectively. However, experiments are primarily on language pairs where the availability of relatively large-scale bilingual datasets supports the viability of this approach.

This paper investigates an approach to expanding the training data effectively by first generating pseudo bilingual texts and then filtering out low-quality alignments. We rely on a quality estimation based on a back-translation ap-

* Corresponding author
E-mail address: ahmadnia@ucdavis.edu

1877-0509
10.1016/j.procs.2021.05.093

proach in order to boost NMT performance through exclusion of low-quality language pairs. Generation of pseudo bilingual texts leverages back-translation, followed by a filtering stage applied to the target monolingual texts for removal of low-quality language pairs. If a target sentence and its back-translation are similar, the synthetic source sentence is deemed appropriate based on its target monolingual sentence and is thus included in the filtered pseudo bilingual dataset. Since low-quality bilingual sentences degrade NMT performance, the quality of the pseudo bilingual dataset is essential. Filtering out low-quality synthetic sentences that would otherwise have been included in our generated bilingual texts yields a high-quality training dataset for low-resource language pairs [3].

We evaluate the impact of our approach on the Persian-Spanish low-resource language pair. We show that although state-of-the-art methods are effective for low-resource language pairs, it is more effective to use a filtered pseudo bilingual dataset as additional training data for low-resource scenarios.

2. Related Work

In the past few years, various approaches have attempted to address the data sparsity problem in machine translation (MT). A common approach is to exploit the availability of monolingual corpora [1, 2, 4]. More recently, data sparsity has been addressed in MT through the exploitation of knowledge from high-resource language pairs, for example, training a single NMT system on a mix of high-resource and low-resource language pairs [6, 7]. Another variant is the application of transfer learning [5], pretraining an NMT system on a high-resource language pair before fine-tuning the system on a target low-resource language pair.

Various approaches employ source-language monolingual texts to enhance translation quality under data-sparse conditions. Domain adaptation in a low-resource setting has been addressed by training a translation model from generated pseudo parallel texts utilizing an in-domain monolingual dataset [8]. In other related work, a pseudo parallel dataset is generated by learning patterns employing source and in-domain monolingual target texts for cross-domain adaptation [9]. In both of these approaches, manual filtering of relatively more accurate translated sentences is used to revise the language model. Our work adopts a similar approach, employing generated pseudo parallel texts derived from translation of a monolingual dataset in the target language, instead of the source language [10]. Our work differs in that we apply automatic filtering to the resulting generated pseudo-parallel texts.

In general, data filtering is achieved in multilingual domain adaptation through application of Phrase-Based Statistical MT (PBSMT) systems: sentences are extracted from large corpora to optimize the language model as well as the translation model [11, 12]. Such approaches are most closely related to our work in that they build a quality estimator to obtain high-quality parallel sentence pairs. This approach has been shown to achieve better translation performance and a reduction in time-complexity with a small high-quality corpus. This prior method filters data by calculating similarity between source and target sentences; our approach differs in that it calculates similarity between monolingual and synthetic target sentences.

In other related work [13], dynamic data selection is undertaken during NMT training. To sort and filter the training dataset, language models are employed, from the source and target sides of in-domain and out-of-domain data, to calculate cross-entropy scores [3]. In our own work, back-translation is used to filter data, as an approximation to meaning preservation.

Dual Learning [14] as well as Round-Tripping [15, 16] simultaneously train two models through a reinforcement learning process. These approaches use monolingual data for both source and target languages and generate informative feedback signals to train the translation models. While these approaches are shown to alleviate the issue of noisy data by increasing coverage, our work aims to remove the noisy data. In addition, the aforementioned approaches assume a high-recourse language pair to cold-start the reinforcement learning process, while in our work, we employ low-resource language pairs for which high-quality seed NMT models are difficult to obtain.

3. Language Issues

Low-resource languages are those that have fewer technologies and datasets relative to some measure of their international importance. The biggest issue with low-resource languages is the extreme difficulty of obtaining sufficient resources for effective machine translation. Natural Language Processing (NLP) methods that have been created for

analysis of low-resource languages are likely to encounter similar issues to those faced by documentary and descriptive linguists whose primary endeavor is the study of minority languages. Lessons learned from such studies are highly informative to NLP researchers who seek to overcome analogous challenges in the computational processing of these types of languages [15, 19].

MT has proven successful for a number of language pairs. However, each language comes with its own challenges, and Persian is no exception. Persian suffers significantly from the shortage of digitally available parallel and monolingual texts. It is a morphologically rich language, with many characteristics shared only by Arabic. It differs from many high-resource languages in that it makes no use of articles (*a*, *an*, *the*) and does not distinguish between capital and lower-case letters. Symbols and abbreviations are rarely used. As a consequence of being written in the Arabic script, Persian uses a set of diacritic marks to indicate vowels, which are generally omitted except in infant writing or in texts for those who are learning the language. Sentence structure is also different from that of English. Persian places parts of speech (e.g., nouns, subjects, adverbs and verbs) in different locations in the sentence, and sometime even omits them altogether. Some Persian words have many different accepted spellings, and it is not uncommon for translators to invent new words. This can result in Out-Of-Vocabulary (OOV) words [15, 19].

Spanish utilizes the Latin alphabet, with a few special letters; vowels with an acute accent *(á, ú, é, ó, í)*, *u* with an umlaut (*ü*), and an *n* with a tilde (*ñ*). Due to a number of reforms, the Spanish spelling system is almost perfectly phonemic and, therefore, easier to learn than the majority of languages. Spanish is pronounced phonetically, but includes the trilled *r* which is somewhat complex to reproduce. In the Spanish IPA, the letters *b* and *v* correspond to the same symbol *b* and the distinction only exists in regional dialects. The letter *h* is silent except in conjunction with *c*, *ch*, which changes the sound into *tf*. The Spanish language punctuation is very close to, but not the same as, English. For example, in Spanish, exclamation and interrogative sentences are preceded by inverted question and exclamation marks. Also, in a Spanish conversation, a change in speakers is indicated by a dash, while in English, each speaker's remark is placed in separate paragraphs. Formal and informal translations address several different characteristics. Inflection, declination and grammatical gender are important features of the Spanish language [15, 19].

A number of *divergences* [17, 18] between low-resource (e.g., Persian) and high-resource (e.g., Spanish) languages pose many challenges in the translation from one to the other, or vice versa. In Persian, the modifier precedes the word it modifies, and in Spanish the modifier follows the head word (although it may precede the head word under certain conditions). In Persian, the sentences follow a "Subject", "Object", "Verb" (SOV) order, and in Spanish, the sentences follow the "Subject", "Verb", "Object" (SVO) order [19]. Such distinctions are exceedingly prevalent and thus pose many challenges for machine translation [15, 19].

4. Methodology

This section describes the application of a method for filtering a bilingual dataset generated via back-translation of a monolingual text under low-resource conditions. The resulting dataset is then employed as an additional training corpus. We then bootstrap an attentional NMT model by iterating the filtering process until convergence.

This method contains the following steps: 1) Translating monolingual target sentences using a model trained on bilingual corpus in target-source direction to produce synthetic source sentences. In this step, we obtain an "Unfiltered" pseudo bilingual texts as an addition to the unfiltered dataset; 2) Back-translating the synthetic source sentences using a model trained on bilingual corpus in source-target direction to obtain a synthetic target sentence; 3) Calculating sentence-level similarity metric scores using the monolingual target sentences as reference and the synthetic target sentences as candidates; 4) Sorting the monolingual target sentences and their corresponding synthetic source sentences, in descending order of sentence-level similarity metric scores, and filtering out sentences with low scores. The threshold is determined by the quality translated sentences in the development set; 5) Employing the filtered synthetic source sentences as the source-side and the monolingual target sentences as the target-side of the pseudo bilingual dataset. The result is referred to as "Filtered" pseudo bilingual texts, which are used as additional training data [3].

After the filtering step, bootstrapping proceeds as follows: 1) Bootstrap$_a$ employs a pseudo bilingual corpus created using the "Parallel" model as additional data to train our NMT system; 2) Bootstrap$_b$ selects the best model on the development set from "Bootstrap$_a$" and trains its target-to-source model.

We employ target sentences from the generated bilingual texts that have been filtered out in the previous iteration to train the best model. If there is no improvement over the previous iteration, bootstrapping is terminated. The

filtered pseudo bilingual dataset and translation model are returned as output, and this process is repeated. Even if the monolingual target sentences remain the same, the synthetic source sentences are refreshed at each iteration. In other words, the translation quality of both the "Unfiltered" and "Filtered" pseudo bilingual datasets is improved through the bootstrapping process until the termination criterion is met.

5. Experiments and Results

In our case study, we utilized the *Tanzil* corpus that contains about 67K sentence pairs. We splitted these into 60K sentence pairs for the training set, 5K sentence pairs for the validation set, and about 2K sentence pairs for the test set. To perform Persian-to-Spanish-to-Persian translation, for the Persian-to-Spanish experiment, we sampled an additional 5,389 Spanish monolingual sentences from the *GNOME* corpus. We also sampled 8,239 Persian monolingual sentences from the same dataset for Spanish-to-Persian translation.

We compared our system to the following baseline systems: 1) Parallel system trained on a parallel corpus in both directions, that is used to generate a bilingual corpus (Bootstrap$_{a,b}$ *Parallel* in the case of bootstrapping); 2) Unfiltered system trained on a concatenated parallel corpus with all generated bilingual corpora without any filtering (Bootstrap$_{a,b}$ *Unfiltered* in the case of bootstrapping).

Following [21] for the implementation, we employed a *Transformer* [20] on top of PyTorch, which uses a 6-layer Long Short-Term Memory (LSTM) encoder-decoder and the hidden layer of 1024. The training uses a mini-batch of 256 and the Stochastic Gradient Descent (SGD) with an initial learning rate of 0.1. We set the size of word embeddings layer to 512. We also set dropout to 0.1. We used a maximum sentence length of 50 words. We also set a beam size of 8, and the model continues for 20 epochs (in both training and test steps) on a single GPU. For evaluation, we adopt the Bilingual Evaluation Understudy (BLEU) [22] metric.

We used our 60K sentence pairs to train the first *Parallel* models in both directions. We then utilized these generated models to create a pseudo bilingual corpus by translating 5,389 Spanish monolingual sentences. A concatenation of parallel and pseudo bilingual sentences is then used to train the *Unfiltered* model. The application of this model yields results indicating that the use of all pseudo bilingual texts (as an additional dataset) reduces the BLEU scores. These results suggest that unfiltered data contain many incorrect sentence pairs which lead to reduced NMT accuracy. (See the top section of Table 1.)

We constructed *Parallel* baseline NMT systems (in both directions) employing an available bilingual corpus to obtain our filtered pseudo-parallel corpus.

Table 1. Translation results.

Threshold	**Development**	**Test**
Parallel Persian-Spanish	17.53	16.13
Parallel Spanish-Persian	20.47	18.71
Unfiltered	19.86	18.05
Filtered	23.69	22.23
Bootstrap$_a$ Parallel Persian-Spanish	16.05	15.32
Bootstrap$_a$ Parallel Spanish-Persian	19.23	17.69
Bootstrap$_a$ Unfiltered	25.03	23.25
Filtered	25.29	24.37
Bootstrap$_b$ Parallel Persian-Spanish	19.35	18.78
Bootstrap$_b$ Parallel Spanish-Persian	23.48	22.20
Bootstrap$_b$ Unfiltered	25.96	24.20
Filtered	27.32	26.02

The *Bootstrap$_a$* results (middle section of Table 1) and *Bootstrap$_b$* (bottom section of Table 1) show that the filtered models outperform the baselines in both directions. Overall, the results indicate that, by being selective about which sentence pairs to use from the pseudo bilingual texts, translation performance is improved in our low-resource experiments.

6. Conclusions

The model trained utilizing filtered pseudo bilingual texts as additional dataset yielded better translation performance than the baselines for our case study. We further improve translation performance over the bootstrapping approach. Our results suggest that translation accuracy heavily depends on data size as well as data quality. While prior work [2] showed that using a pseudo bilingual corpus as additional data yields in large performance improvements, we show even greater improvements from the creation of a better pseudo bilingual corpus. The size of the usable pseudo bilingual corpus for low-resource language pairs is very small, which indicates that filtering out very noisy data results in higher accuracy of the NMT system.

Acknowledgements

The authors would like to acknowledge the financial support received from the Linguistics Department at UC Davis (USA).

References

[1] Zhang, Jiajun, and Zong, Chengqing. (2016) "Exploiting Source-side Monolingual Data in Neural Machine Translation." *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 1535–1545.

[2] Sennrich, Rico, and Haddow, Barry, and Birch, Alexandra. (2016) "Improving Neural Machine Translation Models with Monolingual Data." *Proceedings of the 54th Annual Meeting of Association for Computational Linguistics*, 86–96.

[3] Imankulova, Aizhan, and Sato, Takayuki, and Komachi, Mamoru. (2019) "Filtered Pseudo-Parallel Corpus Improves Low-Resource Neural Machine Translation." *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, **9** (2).

[4] Gülçehre, Çaglar, and Firat, Orhan, and Xu, Kelvin, and Cho, Kyunghyun, and Barrault, Loïc, and Lin, Huei-Chi, and Bougares, Fethi, and Schwenk, Holger, and Bengio, Yoshua. (2015) "On Using Monolingual Corpora in Neural Machine Translation." *ArXiv*, abs/1503.03535.

[5] Zoph, Barret, and Yuret, Deniz, and May, Jonathan, and Knight, Kevin. (2016) "Transfer Learning for Low-Resource Neural Machine Translation." *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 1568–1575.

[6] Firat, Orhan, and Sankaran, Baskaran, and Al-Onaizan, Yaser, and Yarman Vural, Fatos T., and Cho, Kyunghyun. (2016) "Effective Approaches to Attention-based Neural Machine Translation." *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 268–277.

[7] Johnson, Melvin, and Schuster, Mike, and Le, Quoc V., and Krikun, Maxim, and Wu, Yonghui, and Chen, Zhifeng, and Thorat, Nikhil, and Viégas, Fernanda, and Wattenberg, Martin, and Corrado, Greg, and Hughes, Macduff, and Dean, Jeffrey. (2017) "Google's Multilingual Neural Machine Translation System: Enabling Zero-Shot Translation", *Transactions of the Association for Computational Linguistics*, 339–351.

[8] Bertoldi, Nicola, and Federico, Marcello. (2009) "Domain Adaptation for Statistical Machine Translation with Monolingual Resources." *Proceedings of the Fourth Workshop on Statistical Machine Translation*, 182–189.

[9] Hsieh, An-Chang, and Huang, Hen-Hsen, and Chen, Hsin-Hsi. (2013) "Uses of Monolingual In-Domain Corpora for Cross-Domain Adaptation with Hybrid MT Approaches." *Proceedings of the Second Workshop on Hybrid Approaches to Translation*, 117–122.

[10] Adjeisah, Michael, and Liu, Guohua, and Omwenga Nyabuga, Douglas, and Nuetey Nortey, Jinling Song. (2021) "Pseudotext Injection and Advance Filtering of Low-Resource Corpus for Neural Machine Translation." *Computational Intelligence and Neuroscience*.

[11] Moore, Robert C., and Lewis, William. (2010) "Intelligent Selection of Language Model Training Data." *Proceedings of the ACL 2010 Conference*, 220–224.

[12] Axelrod, Amittai, and He, Xiaodong, and Gao, Jianfeng. (2011) "Domain Adaptation via Pseudo In-Domain Data Selection." *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, 355–362.

[13] van der Wees, Marlies, and Bisazza, Arianna, and Monz, Christof. (2017) " Dynamic Data Selection for Neural Machine Translation." *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 1400–1410.

[14] He, Di, and Xia, Yingce, and Qin, Tao, and Wang, Liwei, and Yu, Nenghai, and Liu, Tie-Yan, and Ma, Wei-Ying. (2016) "Dual learning for machine translation." *Proceedings of the 30th Conference on Neural Information Processing Systems*.

[15] Ahmadnia, Benyamin, and Dorr, Bonnie J. (2019) "Augmenting Neural Machine Translation through Round-Trip Training Approach." *Open Computer Science*, **9** (1): 268–278.

[16] Ahmadnia, Benyamin, and Dorr, Bonnie. (2019) "Bilingual Low-Resource Neural Machine Translation with Round-Tripping: The Case of Persian-Spanish." *Proceedings of Recent Advances in Natural Language Processing*, 18–24.

[17] Dorr, Bonnie J. (1994) "Machine Translation Divergences: A Formal Description and Proposed Solution." *Computational Linguistics*, **20** (4):597–633.

[18] Dorr, Bonnie J., and Pearl, Lisa, and Hwa, Rebecca, and Habash, Nizar. (2002) "DUSTer: A Method for Unraveling Cross-Language Divergences for Statistical Word-Level Alignment." *Proceedings of the 5th conference of the Association for Machine Translation in the Americas*.

[19] Ahmadnia, Benyamin, and Serrano, Javier, and Haffari, Gholamreza. (2017) "Persian–Spanish Low-Resource Statistical Machine Translation Through English as Pivot Language." *Proceedings of Recent Advances in Natural Language Processing*, 24–30.

[20] Vaswani, Ashish, and Shazeer, Noam, and Parmar, Niki, and Uszkoreit, Jakob, and Jones, Llion, and Gomez, Aidan N, and Kaiser, Łukasz, and Polosukhin, Illia. (2017) "Attention is all you need." *Proceedings of the 31st Conference on Advances in Neural Information Processing Systems*, 5998–6008.

[21] Ahmadnia, Benyamin. (2020) "Linked Data Effectiveness in Neural Machine Translation." *Proceedings of the 2020 4th International Symposium on Computer Science and Intelligent Control*, 231–234.

[22] Papineni, Kishore, and Roukos, Salim, and Ward, Todd, and Zhu, Wei-Jing. (2001) "BLEU: A Method for Automatic Evaluation of Machine Translation." *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, 311–318.