Translation Studies, Spring 2023, Vol. 21, No. 81, pp. 97–113: **Scientific Research Paper**
DOR: 20.1001.1.17350212.1402.21.1.9.5

Improving English-Persian Neural Machine Translation System through Filtered Back-Translation Method[1]

Pariya Razmdideh[2], Fatemeh Pour-Ali Momen-Abadi[3] & Sajjad Ramezani[4]

Abstract

This study utilizes the neural machine translation (NMT) approach to improve the VRU English-Persian NMT system. In an NMT system, the encoder takes a sequence of source words as inputs and the decoder takes the source vectors through an attention mechanism as input and returns the target words. As English-Persian is a low resource language pair and few researches have been carried out on this language pair, it is important to augment the NMT system with various data. The study explores two methods to enhance the VRU system: back-translation and data filtering. At first, we created NMT models using two corpora, Amirkabir and Persica. To see whether higher ratios of synthetic data leads to decreases or increases in translation performance, we modeled different ratios using back-translation. We found that back-translation significantly improved the VRU NMT system. Second, the filtering method is applied to eliminate noisy data by applying sentence-BLEU, Average Alignment Similarity (AAS), Maximum Alignment Similarity (MAS), combination of AAS and MAS, combination of AAS, MAS, and sent-BLEU. Results show that the combination of AAS, MAS, and sent-BLEU produced the highest growth, with a BLEU score of 30.65. The study concludes that the proposed methods effectively enhance the VRU English-Persian NMT system.

Keywords: AAS, Back-translation, BLEU, Filtering, MAS, Neural machine translation, Tensor2Tensor

[1] This paper was received on 25.01.2023 and approved on 10.04.2023.

[2] Corresponding Author: Assistant Professor of Linguistics, Linguistics Department, Faculty of Foreign Languages, Vali-e-Asr University of Rafsanjan, Iran; email: p.razmdideh@vru.ac.ir

[3] Ph.D. Candidate of Linguistics, Linguistics Department, Faculty of Persian Literature and Foreign Languages, Allameh Tabataba'i University, Tehran, Iran; email: ftemep430@gmail.com

[4] M.A. Graduate Student of Computer Engineering, Computer Engineering Department, Faculty of Engineering, Vali-e-Asr University of Rafsanjan, Iran; email: sajjadramezani90@gmail.com

Introduction

Machine Translation (MT) is a sub-field of computational linguistics that investigates the use of software to translate verbal and non-verbal texts interlingually. MT involves two major approaches: rule-based and data-driven. Within data-driven approaches, three further approaches can be distinguished: Statistical MT (SMT), Example-Based MT (EBMT), and Neural MT (NMT). This research is based on the NMT approach. NMT is a novel approach to MT which is inspired by deep representational learning. NMT attempts to build and train a single large neural network by modeling the entire sentences in a single integrated model. In an NMT module, two sequences of tokens, X= [x_1, x_2, …, xn] and Y= [Y_1, Y_2, …, Y_n] are given in the source and target languages, respectively.

The NMT system aims to model the conditional probability p(y|x) with a single large neural network.

$Y' = argmax\ p(y|x,\vartheta)$

An NMT system is extremely data hungry (Koehn & Knowles, 2017); As a consequence, we will not be in a position to fully employ NMT unless we augment it with a vast amount of monolingual source and/or parallel data. The use of NMT is gaining momentum, due to the numerous substantial works that have been done in this framework. These improvements to NMT have been achieved through methods such as Dual Learning (He et al. 2016) and Round-Tripping (Ahmadnia & Dorr, 2019), Language Model (LM) fusion (including deep fusion and shallow fusion) (Gülçehre et al, 2015), and back-translation (BT) (Sennrich, 2016), which is the focus of this study.

The back-translation technique is applied in either a semi-supervised or an unsupervised MT system (Lample et al. 2018), where both bilingual and monolingual data in the target language are available. This technique begins by training a middle framework using parallel data which is used to translate the target-side monolingual data into the source language. The result is a parallel corpus in which the source-side is synthetic MT output while the target is a genuine text composed by humans. In order to train a system that will translate from the source to the target language, the synthetic parallel corpus is simply added to the real bitext.

In this research, we seek to improve the English-Persian NMT system by increasing the target-side (Persian) monolingual corpus in training and applying the back-translation method. To avoid noises, we applied filtering methods such as measuring the similarity of the vectors and using the translations which are higher than average of three criteria: average AAS, MAS, sent-BLEU, combination of AAS and MAS, and combination of AAS, MAS and sent-BLEU.

Finally, the results are evaluated by using BLEU (Papineni et al, 2002). Few NMT studies have been done on the English-Persian language pair in comparison with other language pairs and none have addressed the use of back-translation and filtering methods to improve the English-Persian NMT system (Ahmadnia et al, 2019; Hoang et al, 2018).

Review of Literature

Although many researches have been carried out on NMT systems, few studies have been carried out on NMT and BT in English-Persian language pair. In this section the most relevant studies are briefly introduced. These researches have worked on different language pairs with Persian. The first researches applied BT on NMT and created pseudo parallel corpora which enhanced the NMTs. the second part of the Literature Review refers to studies which applied BT and Filtered BT to improve NMT.

In applying back-translation in other language pairs, Abdulmumin et al. (2021) carried on their research on English-German language pair. They presented a hybrid approach that improved forward and backward models in back-translation through utilizing the monolingual target data. In this approach, to enhance the backward model utilized the synthetic data through self-learning, and for enhancing the forward model they used standard back-translation; and it showed positive results. Their experimental results show that the advancements accomplished within the performance of the backward models translated into superior forward models, and the results on English-German translation demonstrated that the strategy would be appropriate and generalizable to other language pairs. Likewise, Ahmadnia, and Dorr (2021) proposed an approach to the Farsi-Spanish low-resource NMT system that addressed this inadequacy by leveraging monolingual data from both source and target sides to jointly improve forward and backward models. This approach shares similarities with back-translation due to its iterative process. With the assistance of a monolingual dataset from both source and target languages, their approach aimed to enhance both forward and backward NMT models simultaneously. Moreover, Ahmadnia and Aranovich (2020) compared BT and Re-BT on the case of Spanish-Farsi bilingually Low-Resource NMT. It is stated that NMT requires data augmentation by translating target monolingual data (BT). Re back-translation is a kind of co-training that forward and backward models can be utilized to train one another over both translation directions. In this approach, based on BT a small amount of monolingual data is randomly sampled from the initial monolingual data. BT's quality is important for synthetic data and NMT performance can be improved by re-back-translation over simple back-translation under low-resource condition. Moreover, Sabbagh-Jafari and Ramezani (2022) created translation models based on neural networks which can be improved both in terms of the evaluation criteria (BLEU) and in terms of expanding the range of vocabulary which the model is trained with. Various experiments have been carried out. The first experiment involved several steps, including evaluating the BLEU criterion at the sentence level, measuring the average similarity and maximum similarity between pairs of sentences, combining the average and maximum similarity measures, and incorporating sentences' BLEU scores. It reached the following results: 1- There was a 0.57 increase in Amirkabir's BLEU. 2- The translation model that was trained with the refined pair of corpora by mapping the word vectors with the neural network and combined with the MAS criterion is finally ranked first. 3- The random selection or the use of all the pseudo-parallel sentences did not have a good effect on the quality of the models and reduced the BlEU value. Also, Fadaee and Monz (2018) explored diverse facets of the BT method to obtain an enhanced understanding of its performance. The findings of this research indicate that predicting words in the target language which benefits from back-translation can be challenging. These difficult words are the ones with high prediction loss during training when the translation model converges. They applied back-translation on an NMT system by targeting difficult words on English→German WMT17. Their findings thus show that words with high prescription losses in the target language benefited most from the additional back-translated data. They also discovered the prediction loss to identify weaknesses of translation model and to provide the additional synthetic data targeting these shortcomings for improved translation quality.

Neural Machine Translation

NMT is inspired by deep neural networks based on encoding and decoding of source and target sentences. In this approach, the proper data is obtained through corpora. Unlike conventional translation systems, each component of the neural translation model is taught simultaneously to maximize translation performance. Two main factors in NMT are encoder and decoder. The encoder is used to display the input sentence which is a sequence of words. To study the sequence of words we refer to their hidden matrix and we process them with RNN. This process creates a situation in which each encodes its left context (all the previous words). To obtain its right context a RNN is carried out from right to left or to be precise from the beginning to the end of sentence. This process creates a Bidirectional RNN. An input sequence x_1^n is accepted and a h_1^n, sequence of contextualized representations, is generated by the encoder. Decoder is also an RNN. It receives a display of input context and the predicted previous hidden status and output word. There is a Context state between encoder and decoder; it has a context vector, c, which is the function of h1n. Then the decoder accepts the input, c, and creates an arbitrary length sequence of hidden states h^{m1}, then its output is $y^m{}_1$ (Cho et al. 2014).

The process of NMT approaches is: 1) Learning phrase representation 2) Seq2Seq Learning with NN 3) Effective Approaches to Attention-Based 4) Convolutional Seq2Seq Learning 5) Transformer 6) An Attention-Based User Behavior, and 7) Universal Transformer.

NMT which is implemented as an encoder-decoder network with recurrent neural networks has gained state-of-the art performance for most of language pairs by using monolingual data. Although there are other methods like multilingual to improve NMT's performance the focus of this research is monolingual parallel data. To boost fluency of NMT, it is better to train target-side monolingual data. These methods are: 1) Integrating a separately trained RNN-LM into the NMT decoder (Gülçehre et al. 2015). 2(Using a single dummy token (Sennrich et. al, 2016) 3) Copying the target sentence over to the source side (Currey et. al, 2019) 4) back-translation (Sennrich et. al, 2017) which is implemented in various ways as different iterative back-translation and filtered back-translation which is the main focus of this research.

Back-Translation

NMT requires a lot of data and it is based on the statistics of large parallel corpora. Back-translation (BT) can be carried out to improve NMT systems (Edunov, et al. 2018). The back-translation operates in two ways: 1) Semi-supervised system: The supervised learning models need a dataset with some labels and observations as their inputs. They then learn from the labeled datasets through a learning algorithm to predict the new unseen observations that are given to the model. The model is able to provide targets for any new data after sufficient training. Finally, the output can be compared with the correct intended output to find the errors and modify itself. 2) Unsupervised MT system: For unsupervised learning model, a dataset with some observations and without labeling is needed. A function can be inferred to describe a hidden structure of unlabeled data not to predict the right output but to find the data and to draw the inferences from datasets to describe the hidden structures from unlabeled data. Back-translation is the reverse translation direction. The middle framework is trained by the parallel data which is used to translate the target-side monolingual data into the source language. At the end, a parallel corpus is in which the source-side is synthetic MT output is produced. Parallel corpus and the real bitext together train a system that will translate from the source to the target language. Back translation is effective for both high and low-resource languages. To create synthetic source sentences, it uses beam (successful in finding high probability outputs) or just greedy search. Back translation's training process to obtain pseudo training data for translation from the source (L_S) to the target (L_T) language has three steps: 1) A Back translation model is trained $L_{T\rightarrow}$ L_S 2) L_T monolingual data is back translated into L_S to generate pseudo L_S (source-side) monolingual data; in this step, alignment is needed. 3) A forward-translation model (LS → LT) is trained.

Filtered Back-Translation

In filtering technique, a naive filtering model is applied based on sentence similarity to filter out noisy pseudo parallel data (Fadaee & Monz, 2018). By using the trained NMTs a monolingual source sentence S is translated to target sentence T. To gain the sentence embedding of S and T, the MUSE (Multilingual Universal Sentence) Encoder is applied. Based on the obtained embedding of S and T, the cosine similarity is calculated. A pair is treated noisy if the cosine score is below a certain threshold (it is based on the cosine score on the entire monolingual data). This filtering can be used to sample sentence pairs from the true parallel data. The entire parallel data are sort out in decreasing order of similarity scores; then, the pairs that have the text in the same language in the source-side and target-side using Langid library are eliminated. At the end, the top n sentence pairs from the above data where n is the number of samples required from the true parallel data is returned. Sentence-level Similarity Metrics, AAS, MAS, sent-BLEU, Combination of

AAS and MAS and Combination of AAS, MAS and sent-BLEU for Filtering are used in this study.

Tensor2Tensor

Tensor2Tensor (T2T) is a library of deep learning models and datasets which are designed to make deep learning research faster and accessible. Usability and performance are stressed out in T2T. Through its use of TensorFlow and various T2T-specific abstractions, models can be trained on Central processing unit (CPU), Graphics processing unit (GPU), (single or multiple), and TPU (locally and in the cloud), usually with no or minimal device specific code or configuration. T2T's five key components are: 1) Datasets 2) Device configuration 3) Hyper parameters 4) Model 5) Estimator and Experiment.

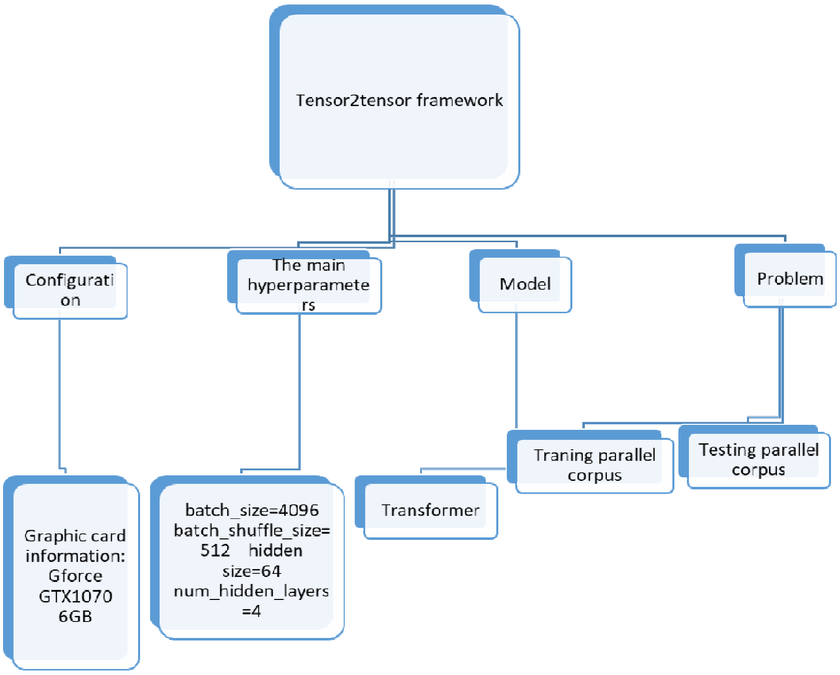

Tensor2tensor framework
Configurati on
The main hyperparamete rs
Model
Problem
Traning parallel corpus
Testing parallel corpus
Graphic card information: Gforce GTX1070 6GB
batch_size=4096 batch_shuffle_size= 512 hidden size=64 num_hidden_layers =4
Transformer

Figure 1. Tensor2tensor framework

Corpora

Since NMT requires vast amount of data, authentic corpora and/or synthetic corpora play an important role in our research so we utilized two kinds of corpora.

Amirkabir Corpus

The Amirkabir Bilingual Farsi-English Corpus (AFEC) holds a prominent position as a valuable resource in the field of machine translation. This bilingual parallel English-Farsi corpus, known for its considerable size, high quality, and broad coverage of news issues, has been widely regarded as a reliable benchmark (Jabbari et al. 2012). The availability of the Amirkabir Corpus has facilitated the training of English-Persian neural machine translation systems, enabling researchers to explore the capabilities and limitations of these models. Amirkabir bilingual corpus has 575592 pairs of parallel sentences produced by Amirkabir University.

Persica Corpus

This corpus includes news texts collected from the ISNA news agency website in the early 1990s. Some of these texts have been published in the format of Persica (Eghbalzadeh et al. 2012). The initial texts included separate news on different subjects and had subject tags, which in order to be used in systems such as speech recognition and machine translation, all of them came in the form of one file, and normalization of the text, removal of additional tags, and writing corrections specific to the Persian language was carried out automatically. Also, the whole text is included in different lines based on the dot character, as a final mark. Corpus texts are saved in utf8 encoding format. The total volume of the texts after the aforementioned pre-processing is about 4.8 GB and the number of tokens is about 550 million tokens.

Method

The NMT system we are using is tensor-to-tensor which is GPU based (Nvidia 1070Ti) and consists of three main parts: encoder that encodes source side and provides embeddings words' vectors go to training frame, decoder that decodes the vectors which is responsible for the production of output, and body that is the main function and relates all the parts together which is a connective between encoder and decoder. The learning rate is 0.1 for the first 10 epochs and the batch size is 1096. The English-Persian NMT system was trained using Amirkabir bilingual Persian-English corpus. To increase target-side monolingual corpus, Persica corpus was used. We proposed two interesting scenarios in this research to improve the VRU NMT system:

First, we trained the NMT system from English to Persian. Due to the lack of data in Persian corpora, the data from Persica corpus would be picked on different ratios of Amirkabir bilingual Persian-English corpus. The selected target side monolingual data were trained in the English-Persian NMT system to make the synthetic corpus named VRU Synthetic target corpus. To make VRU Synthetic target corpus, we classified the sentences according to the number of words. The classification intervals were: 5 and 10, 10 and 20, 20 and 30,.....,200 and 300. Secondly, according to the frequency of sentences in each category, we randomly extracted sentences from those categories in proportion to the frequency of each category, so that 7 million sentences were obtained.

Second, translation models were created from VRU corpus which contained seven million sentences. Amirkabir Model has *575591*(~600) sentences. The system trained using this corpus with 30.50 BLEU score. Firstly, according to this model, we take the same number of sentences from our seven million pseudo-parallel corpus and create a translation model for 600K data (+Synthetic 1:1), which resulted in 28.66 BLEU score. To see whether higher ratios of synthetic data leads to decreases or increases in translation performance, we followed the pattern and add 600K data at each step that created + synthetic 1:2,1:3,1:4,1:5,1:6,1:7,1:8. Each ratio has different BLEU score. Overall, BT enhanced our NMT and raised the BLEU score.

Third, the most qualified sentences filtered to add to the real bilingual corpus. These filtered sentences were gained through some criteria which are sent-BLEU, AAS, MAS, Combination of AAS and MAS, Combination of AAS, MAS, and sent-BLEU. We filtered the pseudo-parallel corpus and extracted the high quality sentences and added them to the Amirkabir parallel corpus. Then, we created a new translation model and evaluated its quality according to the BLEU score.

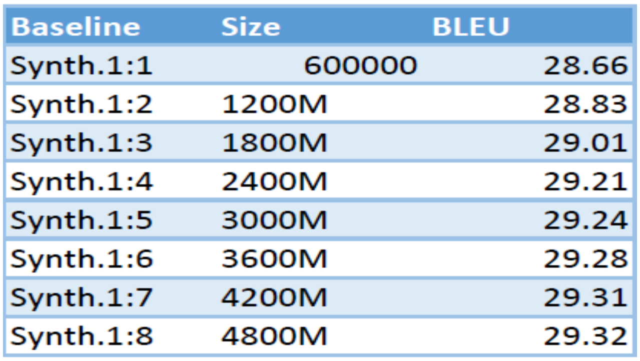
Data Description and Analysis

The ultimate purpose of this study is to improve the English-Persian NMT system which is trained in the NLP laboratory of the Vali-e-Asr University of Rafsanjan. At first, we trained the NMT system with Amirkabir data. Then we applied Back-Translation. In back-translation, we used different portions of our original data set at each step. Then, we measured BLEU for each synthetic ratio e.g. 1:1, 1:2, …1:8.

To evaluate BLEU in different ratios, we considered Amirkabir Model which has *575591*(~600) sentences (*30.50* BLEU score) as a baseline. First, according to this model, we took the same number of sentences from our seven million pseudo-parallel corpus and created a translation model for 600K data (+Synthetic 1:1), which resulted in *28.66* BLEU score. To see whether higher ratios of synthetic data leads to decreases or increases in translation performance, we followed the pattern and added 600K data at each step that created + synthetic 1:2,1:3,1:4,1:5,1:6,1:7,1:8. Second, we doubled the number of sentences and created a translation model for 1200M data (+Synthetic 1:2) and resulted in *28.83* BLEU score which is a slight rise in comparison with +Synthetic 1:1. Third, a translation model consists of 1800M data (+Synthetic 1:3) was created and resulted in *29.01* BLEU score which was considered an improvement compared to +synthetic 1:1 and 1:2. Fourth, a translation model consists of 2400M data (+Synthetic 1:4) was created and resulted in *29.21* BLEU score which was considered an improvement compared to +synthetic 1:1, 1:2 and 1:3. Fifth, a translation model with 3000M data (+Synthetic 1:5) was created and resulted in *29.24* BLEU score which was also considered an improvement compared to +synthetic 1:1, 1:2 and 1:3. Sixth, a translation model including 3600M data (+Synthetic 1:6) was created and resulted in *29.28* BLEU score. Next, a translation model consisted of 4200M data (+Synthetic 1:7) was created and resulted in *29.31* BLEU score which was considered an improvement compared to +synthetic 1:6. At the end, a translation model containing 4800M data (+Synthetic 1:8) was created and resulted in *29.32* BLEU score which was considered an improvement compared to +synthetic 1:5. We did not go any further because sometimes augmenting NMT with various data has a negative effect and eight steps are enough to answer our question. Table (1) depicts that BLEU score increases in different ratios. To introduce new context and improve the fluency of the translation target side monolingual data was added to the training data. Based on Lambert et al. (2011) research, adding synthetic source and real target data improves the phrase-based machine translation (PBMT) and NMT.

As table (1) shows, we could improve our NMT system by using Back-Translation.

Table 1: Back-translation BLEU of systems with different ratios of real: synthetic data



Baseline	Size	BLEU
Synth.1:1	600000	28.66
Synth.1:2	1200M	28.83
Synth.1:3	1800M	29.01
Synth.1:4	2400M	29.21
Synth.1:5	3000M	29.24
Synth.1:6	3600M	29.28
Synth.1:7	4200M	29.31
Synth.1:8	4800M	29.32

Filtered BT

Filtered Back-translation proposed by Jaiswal et al (2020) improves NMT system. In this technique, a filtering model is applied based on sentence similarity to filter out noisy pseudo parallel data. There are three filtering metrics sent-BLEU, AAS, and MAS (Song and Roth, 2015).

Sent-BLEU

We used the BLEU metric to evaluate translated texts in MT. In BLEU, the weight combination of the common n-grams between the translated sentence (candidate) and the reference or references of that sentence is calculated. The first method of calculating BLEU is to calculate the accuracy of the n-gram. And then, based on the number of words in the candidate sentence and the reference sentence, a brevity penalty for accuracy is considered. When the accuracy is high, the number of candidate and reference words should be the same. Finally, BLEU is obtained. The result of evaluation through BLEU at the sentence level is a number between 0 and 1. When two sentences are completely concordant the result is 1 otherwise the result is a number between 0 and 1. BLEU is language-independent at the sentence level and its calculation is inexpensive and it is a method of evaluating sentences based on human evaluation. Our proposed filtering method was to use Persica's Persian sentences as source sentences and considered Persica's synthetic Persian sentences as candidate sentences. Therefore, we used the NLTK tool and calculated the BLEU score of the pair of candidate and reference sentences. Finally, there is a BLEU score for each pair of sentences, and we normalized these values. Then, we considered their average as the threshold and extracted the pseudo-parallel sentences according to the index of the sentences whose BLEU score is better than or equal to the threshold.

The experiment is related to the translation model whose training data is composed of 575591 pair of sentences from Amirkabir corpus and 1900000 pair of filtered sentences with sent-BLEU score. This model was tested by Amirkabir model. Finally, we reached to 30.60 BLEU score. This score showed improvement in NMT system.

Table 2: Sent-BLEU-Amirkabir training corpus
BLEU score improved from 30.50 to 30.60.

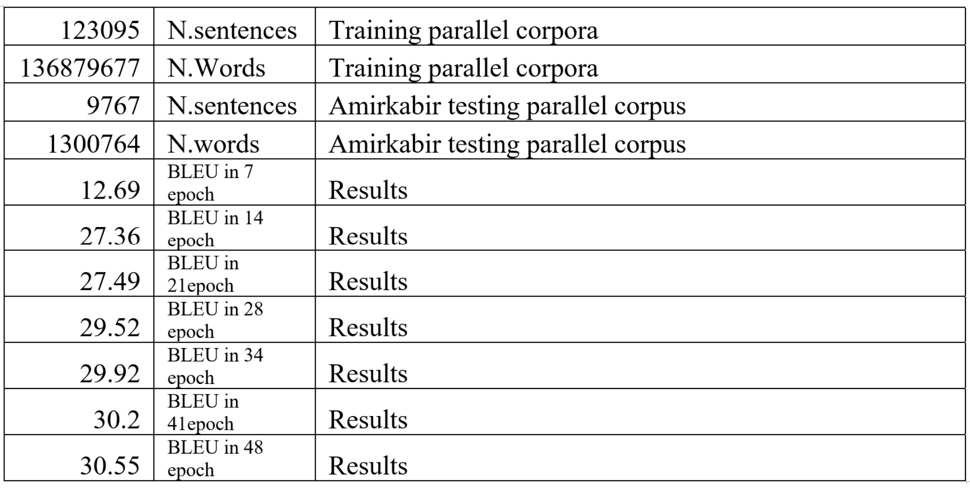
123095	N.sentences	Training parallel corpora
136879677	N.Words	Training parallel corpora
9767	N.sentences	Amirkabir testing parallel corpus
1300764	N.words	Amirkabir testing parallel corpus
27.05	BLEU in 7 epoch	Results
29.02	BLEU in 14 epoch	Results
29.15	BLEU in 21epoch	Results
29.97	BLEU in 28 epoch	Results
29.92	BLEU in 34 epoch	Results
30.33	BLEU in 41epoch	Results
30.60	BLEU in 48 epoch	Results

MAS

The cosine similarity between the most similar word from the monolingual target sentence and each word from the synthetic target sentence is the MAS score. This metric calculates the similarity of two corresponding sentences, one of which is from the Persica (*y)* corpus and the other from the VRU *(y')* corpus. Our method is that for two corresponding sentences, it selects the first word of the sentence *(y)* and then calculates the cosine angle between all the words of the sentence *(y')* and then calculates their MAS. This process goes on for all of the sentences in *y* and finally the MAS is obtained for the sentence. This value is calculated for all the corresponding sentences and finally we would have one MAS value for each pair of sentences. Then we considered their average as the threshold limit and extracted the sentences of the pseudo-parallel corpus according to the index of the sentences, whose MAS value is better than or equal to the threshold limit.

The experiment is related to the translation model whose training data is composed of 575591 pair of sentences from Amirkabir corpus and 2100000 pair of filtered sentences with MAS score. This model was tested by Amirkabir model. Finally, we reached to 30.55 BLEU score. This score showed a slight improvement in NMT system.

Table 3: MAS Amirkabir Training corpus



123095	N.sentences	Training parallel corpora
136879677	N.Words	Training parallel corpora
9767	N.sentences	Amirkabir testing parallel corpus
1300764	N.words	Amirkabir testing parallel corpus
12.69	BLEU in 7 epoch	Results
27.36	BLEU in 14 epoch	Results
27.49	BLEU in 21epoch	Results
29.52	BLEU in 28 epoch	Results
29.92	BLEU in 34 epoch	Results
30.2	BLEU in 41epoch	Results
30.55	BLEU in 48 epoch	Results

Based on the obtained BLEU scores, our NMT system improved from 30.50 to 30.55.

AAS

The average cosine similarity between vectors of the whole words in monolingual and synthetic target sentences is the AAS score (Imankulova et al. 2018). This metric calculates the average similarity of two corresponding sentences, one of which is the Persian form of Persica y and the other is of the synthetic form of Persica y'. Our method is that for two corresponding sentences, it chooses the first word of the sentence y, and then calculates the cosine angle between all the words of the sentence y' and then calculates their sum. This process goes on for all of the sentences in y. Finally, the average of the desired values is obtained. This value is calculated for all corresponding sentences and finally we will have an AAS value for each pair of sentences which are normalized. Then we consider their average as the threshold limit and extract the pseudo-parallel sentences according to the index of the sentences, whose AAS value is greater than or equal to the threshold limit.

The experiment is related to the translation model whose training data is composed of 575591 pair of sentences from Amirkabir corpus and 1200000 pair of filtered sentences with AAS score. This model was tested by Amirkabir model. Finally, we reached to 30.17 BLEU score. This score did not show improvement in NMT system.

Table 4: AAS Amirkabir training corpus

123095	N.sentences	Training parallel corpora
136879677	N.Words	Training parallel corpora
9767	N.sentences	Amirkabir testing parallel corpus
1300764	N.words	Amirkabir testing parallel corpus
25.93	BLEU in 7 epoch	Results
28.23	BLEU in 14 epoch	Results
29.14	BLEU in 21epoch	Results
29.65	BLEU in 28 epoch	Results
30.14	BLEU in 34 epoch	Results
30.06	BLEU in 41epoch	Results
30.17	BLEU in 48 epoch	Results

According to the table, our NMT system improved 0.31 unit in BLEU score. But it did not improve our NMT system based on the baseline.

Combination of AAS and MAS

Using the MAS and AAS values obtained in the previous two methods, we presented the combined method of using both metrics and finally the simultaneous effect of the two MAS and AAS values on filtering high quality sentences. From summing the normalized data of both metrics for each sentence and using the mean threshold, we extracted sentences whose combined value is better than or equal to the threshold.

The experiment is related to the translation model whose training data is composed of 575591 pair of sentences from Amirkabir corpus and 1750000 pair of filtered sentences with the combination of AAS and MAS. This model was tested by Amirkabir model. Finally, we reached to 30.40 BLEU score. This score did not show improvement in NMT system.

Table 5: Combination of AAS and MAS Amirkabir Training corpus

123095	N.sentences	Training parallel corpora
136879677	N.Words	Training parallel corpora
9767	N.sentences	Amirkabir testing parallel corpus
1300764	N.words	Amirkabir testing parallel corpus
25.57	BLEU in 7 epoch	Results
28.23	BLEU in 14 epoch	Results
25.52	BLEU in 21epoch	Results
28.95	BLEU in 28 epoch	Results
29.46	BLEU in 34 epoch	Results
29.99	BLEU in 41epoch	Results
30.44	BLEU in 48 epoch	Results

Combination of AAS and MAS did not improve our NMT system. Baseline's BLEU score is 30.50 but the obtained BLEU in the combination of AAS and MAS is 30.44.

Combination of AAS, MAS and sent-BLEU

We provided high quality filtered sentences by using the obtained value of AAS, MAS, and BLEU criteria and finally the simultaneous effect of three values. From summing the normalized data of all three criteria for each sentence and using the average threshold, we extracted the sentences whose combined value is greater than or equal to the threshold.

The experiment is related to the translation model whose training data is composed of 575591 pair of sentences from Amirkabir corpus and 2260000 pair of filtered sentences with sent-BLEU score. This model was tested by Amirkabir model. Finally, we reached to 30.65 BLEU score. This score showed improvement in NMT system.

Table 6: Combination of AAS, MAS and BLEU Amirkabir Training corpus

123095	N.sentences	Training parallel corpora
136879677	N.Words	Training parallel corpora
9767	N.sentences	Amirkabir testing parallel corpus
1300764	N.words	Amirkabir testing parallel corpus
0.50	BLEU in 7 epoch	Results
28.52	BLEU in 14 epoch	Results
29.73	BLEU in 21epoch	Results
29.71	BLEU in 28 epoch	Results
29.98	BLEU in 34 epoch	Results
30.35	BLEU in 41epoch	Results
30.65	BLEU in 48 epoch	Results

Combination of AAS, MAS and BLEU showed a growth in BLEU score. In table 6, the BLEU scores in 41 and 48 epochs were 29.99 and 30.40 respectively. In this table, the BLEU score in 41 is 30.35 and in 48 epochs was 30.65.

Conclusion

In this study, we conducted an investigation into the performance of English-Persian NMT translation models, focusing on the utilization of back-translation and the filtering of noisy data. Our findings demonstrate that employing back-translation in different ratios has led to superior results compared to the NMT VRU system. Furthermore, we implemented various filtering techniques, including sent-BLEU, AAS, MAS, and combinations thereof, to eliminate noisy data and enhance the NMT system's performance.

Specifically, our experiments yielded the following outcomes:

1. Filtering 1,900,000 sentences using sent-BLEU resulted in a notable improvement, as evidenced by a BLEU score of 30.60.
2. Filtering 1,200,000 sentences using AAS did not lead to a significant improvement in the NMT system, as indicated by a BLEU score of 30.17.
3. Filtering 2,100,000 sentences using MAS resulted in a slight improvement, with a BLEU score of 30.55.
4. Filtering 1,750,000 sentences using the combination of AAS and MAS did not show notable improvement, with a BLEU score of 30.40.
5. Filtering 2,260,000 sentences using the combination of sent-BLEU, AAS, and MAS showcased a considerable enhancement in the NMT system, as reflected by a BLEU score of 30.65.

Overall, the filtering of noisy data has had a significant and positive impact on the performance of our NMT system. These findings highlight the importance of data quality and demonstrate the potential for further optimizing NMT models through effective filtering techniques.

Acknowledgements

The work was supported by a grant from Vali-e-Asr Universityof Rafsanjan under Pata research contract.

Works Cited:

Abdulmumin, I., Galadanci, B.S., Isa, A., Kakudi, H.A., & Sinan, II. (2021). A Hybrid Approach for Improved Low Resource Neural Machine Translation using Monolingual Data. *Engineering Letters*, 29 (4), 1478–1493.

Ahmadnia, B., & Aranovich, R. (2020). An Effective Optimization Method for Neural Machine Translation: The Case of English-Persian Bilingually Low-Resource Scenario. *Proceedings of the 7th Workshop on Asian Translation (WAT 2020).*

Ahmadnia, B., & Dorr, B. J. (2019). Augmenting Neural Machine Translation through Round-Trip Training Approach. Open Computer Science, 9, 268–278.

Ahmadnia, B., Dorr, B. J., & Aranovich, R. (2021). Impact of Filtering Generated Pseudo Bilingual Texts in Low-Resource Neural Machine Translation Enhancement: The Case of Persian-Spanish. *Procedia Computer Science*, 189, 136–141.

Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning phrase representations using RNN encoder-decoder for statistical machine translation, *Conference on Empirical Methods in Natural Language Processing.*

Currey, A., & Heafield, K. (2019). Zero-Resource Neural Machine Translation with Monolingual Pivot Data. Proceedings of the 3rd Workshop on Neural Generation and Translation (WNGT 2019), *Association for Computational Linguistic,* 99–107.

Edunov, S., Ott, M., Auli, M., & Grangier, D. (2018). Understanding Back-Translation at Scale. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 489–500.

Eghbalzadeh, H, B., Khadivi, Sh., & Khodabakhsh, A. (2012). Persica: A Persian Corpus for Multipurpose Text Mining and Natural Language Processing. In *Sixth International Symposium Telecommunication (IST)*, IEEE, Tehran.

Fadaee, M., & Monz, C. (2018). Back-Translation Sampling by Targeting Difficult Words in Neural Machine Translation. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics,* 436–446.

Gülçehre, C., Firat, O., Xu, K., Cho, K., Barrault, L., Lin H., Bougares, F., Schwenk, Hand Bengio, Y. (2015). On Using Monolingual Corpora in Neural Machine Translation.1503.03535v2.