Research Article

Benyamin Ahmadnia* and Bonnie J. Dorr

# Augmenting Neural Machine Translation through Round-Trip Training Approach

https://doi.org/10.1515/comp-2019-0019
Received Apr 18, 2019; accepted Sep 12, 2019

**Abstract:** The quality of Neural Machine Translation (NMT), as a data-driven approach, massively depends on quantity, quality and relevance of the training dataset. Such approaches have achieved promising results for bilingually high-resource scenarios but are inadequate for low-resource conditions. Generally, the NMT systems learn from millions of words from bilingual training dataset. However, human labeling process is very costly and time consuming. In this paper, we describe a round-trip training approach to bilingual low-resource NMT that takes advantage of monolingual datasets to address training data bottleneck, thus augmenting translation quality. We conduct detailed experiments on English-Spanish as a high-resource language pair as well as Persian-Spanish as a low-resource language pair. Experimental results show that this competitive approach outperforms the baseline systems and improves translation quality.

**Keywords:** natural language processing, neural machine translation, low-resource language pairs, round-tripping

## 1 Introduction

Neural Machine Translation (NMT) has made considerable progress in recent years. However to achieve acceptable translation output, large sets of aligned parallel sentences are required for the training phase. Thus, as a data-driven paradigm, the quality of NMT output strongly depends on the quality as well as quantity of the provided training data [1]. Unfortunately, in practice, collecting such parallel data in practice is very expensive. As such, parallel bilingually data is limited for many language pairs, e.g. Persian-Spanish [2].

Assuming that large monolingual texts are available, an obvious next step is to leverage these texts to augment the NMT systems' performance. Various approaches have already been developed for this purpose. In some approaches, target monolingual texts are employed to train a Language Model (LM) that is then integrated with MT models trained from parallel texts to enhance translation quality. Although these approaches utilize monolingual text to train a LM, they do not address the shortage of bilingual training datasets [3, 4]. In other approaches, bilingual datasets are automatically generated from monolingual texts by utilizing the Translation Model (TM) trained on aligned bilingual text; the resulting sentence pairs are used to enlarge the initial training dataset for subsequent learning iterations [5, 6]. Although these approaches enlarge the bilingual training dataset, there is no quality control and, thus, the accuracy of the generated bilingual dataset cannot be guaranteed [7].

To tackle the issues described above, we apply a new round-tripping approach that incorporates *dual learning* [8] for automatic learning from unlabeled data, but transcends prior work through effective leveraging of monolingual text. Specifically, the round-tripping approach takes advantage of the bootstrapping methods including self-training and co-training. These methods start their mission from a small set of labelled examples, while also considering one or two weak translation models, and make improvement through the incorporation of unlabeled data into the training dataset.

In the round-tripping approach, the two translation tasks (forward and backward) together make a *closed loop*, i.e., one direction produces informative *feedback* for training the TM for the other direction, and vice versa. The feedback signals—which consist of the language model likelihood of the output model and the reconstruction error of the original sentence—drive the process of iterative updates of the forward and backward TMs.

For the purpose of evaluation, we apply this approach to a bilingually high-resource scenario (English-Spanish) as well as a bilingually low-resource scenario (Persian-Spanish) to leverage monolingual data in a more effective way. By utilizing the round-trip training approach,

***Corresponding Author: Benyamin Ahmadnia:** Department of Computer Science, Tulane University, New Orleans, LA 70118, United States of America; Email: ahmadnia@tulane.edu
**Bonnie J. Dorr:** Institute for Human and Machine Cognition (IHMC), Ocala, FL 34471, United States of America; Email: bdorr@ihmc.us

the monolingual data play a similar role to the bilingual data, effectively reducing the requirement for parallel data. In particular, each model provides guidance to the other throughout the learning process. Our experimental results demonstrate that round-tripping for NMT works well over the baselines. By learning from monolingual data, this approach achieves comparable accuracy to a NMT approach trained from the full bilingual data for the two translation tasks (forward and backward). As discussed in Section 5, our enlarged systems achieve comparable accuracy to a NMT approach trained on a comparable amount of bilingual data.

In general, the round-tripping approach significantly reduces the requirement on the aligned bilingual data, and helps us to translate from scratch even without access to parallel data. This communication game can be played for an arbitrary number of rounds, and the two TMs will get improved through this procedure. In this way, we develop a general learning framework for training NMT models through a round-tripping game.

This paper is organized as follows; Section 2 presents the previous related work. Section 3 describes the language issues. In Section 4, we briefly review the relevant mathematical background of NMT paradigm. Section 5 describes the round-trip training approach. The experimental framework is covered by Section 6. The results analysis is presented in Section 7. Conclusions and future work are discussed in Section 8.

## 2 Related work

It is almost impossible to provide high-quality state-of-the-art NMT systems for rare or low-resource languages because of the dependence on large parallel corpora, which may be available for high-resource languages but require expensive human annotation under low-resource conditions. Due to the small size of training datasets, such systems generally produce inferior translation output. Unfortunately, large quantities of parallel data are not available for a certain number of language pairs. This has triggered a new research challenge in MT related to training an NMT system where availability of parallel texts is not sufficient.

The integration of monolingual data for NMT models was first proposed by Gülçehre et al. [4], who train monolingual LMs independently, and then integrate them during decoding through rescoring of the beam (adding the recurrent hidden state of the LM to the decoder state of the encoder-decoder network). In this approach, an additional controller mechanism controls the magnitude of the LM signal. The controller parameters and output parameters are tuned on further parallel training data, but the LM parameters are fixed during the fine tuning stage.

Jean et al. [9] also report on experiments with reranking of NMT output with a 5-gram LM, but improvements are small. The production of generated parallel texts bears resemblance to data augmentation techniques, where datasets are often augmented with rotated, scaled, or otherwise distorted variants of the limited training set [10].

A similar avenue of research is self-training [11]. The self-training approach as a bootstrapping method typically refers to the scenario where the training dataset is enhanced with training instances with artificially produced output labels (whereas we start with neural network based output, i.e., the translation, and artificially produce an input). We expect that this is more robust towards noise in MT. Also co-training [12] as another bootstrapping method increases the amount of labelled data through effective use of large amounts of unlabelled data. The basic idea of co-training is to check for redundancies in the unlabelled data and then to leverage these to support two or more separate but redundant views in the form of disjoint feature subsets.

Improving NMT with source-side monolingual data, following similar work on phrase-based Statistical Machine Translation (SMT) [13], remains possible future work. Domain adaptation of neural networks via continued training has been shown to be effective for neural LMs by Ter-Sarkisov et al. [14].

Sennrich et al. [15] noted that encoder-decoder NMT architectures already have the capacity to learn the same information as in the LM, and they explore strategies to train on monolingual data without changing the neural network architecture. By pairing monolingual training data with an automatic back-translation, they are able to treat the data as additional parallel training data, with substantial improvements.

Artetxe et al. [16] proposed a novel method to train an NMT system in an unsupervised manner, relying on monolingual corpora. Their model builds upon the recent work on unsupervised embedding mappings, and consists of a slightly modified attentional encoder-decoder model (which selectively focuses on sub-parts of the sentence during translation) that is trained on monolingual corpora alone using a combination of denoising and back-translation.

Relatedly, Lample et al. [17] proposed a model that takes sentences from monolingual corpora in two different languages and maps them into the same latent space. By learning to reconstruct in both languages from this shared feature space, the model effectively learns to translate without using any labeled data. Along the same lines, Yang et al. [18] introduced an extension that used two independent encoders that shared partial weights that extract high-level representations of the input sentences. In addition, two different Generative Adversarial Networks (GANs), namely the local GAN and global GAN, were proposed to enhance the cross-language translation.

One of the common solutions to the lack of parallel data is the use of a pivot (bridge) language technique [19]. This technique induces a systematic approach to MT when a proper bilingual corpus is lacking or the existing ones are weak. Ahmadnia et al. [20] indicated that the pivot language technique outperforms direct SMT processes currently in use between Persian and Spanish languages. They investigated both sentence pivoting and phrase pivoting, demonstrating that phrase-level pivoting outperforms sentence-level pivoting for Persian-Spanish SMT. They also suggested a method called *combination model* in which the standard direct translation model and the best triangulation pivoting model are blended in order to reach a high-quality translation.

Round-tripping approach has already been utilized in phrase-based SMT by Ahmadnia et al. [21]. In this work, forward and backward models produce informative feedback to iteratively update the TMs during the training of the system until convergence.

# 3 Language issues

Low-resource languages, also known as resource poor, 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. 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.

## 3.1 Persian language issues

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 morphologically rich, with many characteristics shared only by Arabic. It makes no use of articles (*a*, *an*, *the*) and no distinction between capital and lowercase 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 speeches such as nouns, subjects, adverbs and verbs in different locations in the sentence, and sometimes 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.

## 3.2 Spanish language issues

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. There are a few significant differences. 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 Spanish language.

## 3.3 Farsi-Spanish divergences

A number of *divergences* [22, 23] between low-resource (e.g., Farsi) and high-resource (e.g., Spanish) languages pose many challenges in the translation from one to the other, or vice versa. In Farsi, 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 Farsi, the sentences follow a "Subject", "Object", "Verb" (SOV) order, and in Spanish, the sentences follow the "Subject", "Verb", "Object" (SVO) order [24]. Such distinctions are exceedingly prevalent and thus pose many challenges for machine translation.

## 4 Neural machine translation

NMT consists of an encoder and a decoder. Following Bahdanau et al. [1], we adopt an *attention-based* encoder-decoder model, i.e., one that selectively focuses on subparts of the sentence during translation. Consider a source sentence $x = x_1, x_2, ..., x_J$, the encoder transforms it to an internal representation $h = h_1, h_2, ..., h_J$ and then a decoder decodes $h$ to the target sentences $y = y_1, y_2, ..., y_I$. The problem of translation from the source language to the target is solved by finding the best target language sentence $\hat{y}$ that maximizes the conditional probability:

$$\hat{y} = \underset{y}{\operatorname{argmax}}\, P(y|x) \tag{1}$$

The NMT models the conditional probability of the target sentence as:

$$P(y|x) = \prod_{i=1}^{I} P(y_i|y_{<i}, x) \tag{2}$$

Both the encoder and decoder components are Recurrent Neural Networks (RNNs). The encoder converts the source words into a sequence of vectors, and the decoder generates target words one-by-one based on the conditional probability shown in the Equation (2). More specifically, the encoder takes a sequence of source words as inputs and returns forward hidden vectors $\overrightarrow{h_j}(1 \leq j \leq J)$ of the forward-RNN:

$$\overrightarrow{h_j} = f(\overrightarrow{h_{j-1}}, x) \tag{3}$$

Similarly, we obtain backward hidden vectors $\overleftarrow{h_j}(1 \leq j \leq J)$ of the backward-RNN, in the reverse order.

$$\overleftarrow{h_j} = f(\overleftarrow{h_{j-1}}, x) \tag{4}$$

The forward and backward vectors are concatenated to make source vectors $h_j(1 \leq j \leq J)$ based on Equation (5):

$$h_j = \left[\overrightarrow{h_j}; \overleftarrow{h_j}\right] \tag{5}$$

The decoder takes source vectors as input and returns target words. It starts with the initial hidden vector $h_J$ (concatenated source vector at the end), and generates target words in a recurrent manner using its hidden state and an output context. The conditional output probability of a target language word $y_i$ is defined as follows:

$$P(y_i|y_{<i}, x) = \operatorname{softmax}(f(d_i, y_{i-1}, c_i)) \tag{6}$$

where $f$ is a non-linear function and $d_i$ is a the hidden state of the decoder at step $i$:

$$d_i = g(d_{i-1}, y_{i-1}, c_i) \tag{7}$$

where $g$ is a non-linear function taking its previous state vector with the previous output word as inputs to update its state vector. $c_i$ is a context vector to retrieve source inputs in the form of a weighted sum of the source vectors $h_j$, first taking as input the hidden state $d_i$ at the top layer of a stacking Long Short-Term Memory (LSTM) [25]. The goal is to derive a context vector $c_i$ that captures relevant source information to help predict the current target word $y_i$. While these models differ in how the context vector $c_i$ is derived, they share the same subsequent steps. $c_i$ is calculated as follows:

$$c_i = \sum_{j=1}^{J} \alpha_{t,j} h_j \tag{8}$$

where $h_j$ is the annotation of source word $x_j$ and $\alpha_{t,j}$ is a weight for the $j^{th}$ source vector at time step $t$ to generate $y_i$:

$$\alpha_{t,j} = \frac{\exp(\operatorname{score}(d_i, h_j))}{\sum_{j'=1}^{J} \exp(\operatorname{score}(d_i, h_{j'}))} \tag{9}$$

The above score function may be defined in variety of ways as discussed by Luong et al. [26]. We use *dot* attention for this score function calculated as follows:

$$\operatorname{score}(d_i, h_j) = d_i^T h_j \tag{10}$$

This scalar product score basically means the decoder puts more weights (attention) to source vectors close to its state vector $d_i$.

In this paper, we denote all the parameters to be optimized in the neural network as $\Theta$ and denote $C$ as the dataset that contains source-target sentence pairs for the training phase. Hence, the learning objective is to seek the optimal parameters $\Theta^*$:

$$\Theta^* = \underset{\Theta}{\operatorname{argmax}} \sum_{(x,y)\in C} \sum_{(i=1)}^{I} \log P(y_i|y_{<i}, x; \Theta) \tag{11}$$

## 5 Method description

Round-tripping involves two related translation tasks: the outbound-trip (source-to-target direction) and the inbound-trip (target-to-source direction). The defining traits of these forward and backward tasks are that they form a closed loop and both produce informative feedback that enables simultaneous training of the TMs. In fact, round-tripping leverages monolingual data in the most effective and influential way possible, for both the source and the target languages. This approach enables the monolingual data to play a role that is similar to the bilingual data, and this helps in the gradual reduction of the requirement on bilingual text during the training phase [7].

Generally, the round-tripping procedure for NMT is described as follows:

- The first translation system understands the language $X$ and it sends a message in this language to the other translation system. The second translation system understands language $Y$. After checking the message, it sends a notification to the first translation system.
- After receiving the message by the first translation system from the second one, it checks the message and then sends a notification to the second translation system as well.
- After receiving this feedback, both translation systems know about the performance of the two TMs, and as a result of this feedback, they make the required changes.

The structure of round-tripping contains bootstrapping methods including self-training as well as co-training. In general, these methods aim to augment translation quality. They leverage a small set of labelled data coupled with a set of weak TMs (emergent from training on initial small bilingual corpora) to make improvements through incorporation of unlabelled data into the training dataset. The round-trip training approach resembles self-training because the outbound-trip produces translations for monolingual source sentences which are then used to retrain itself. The round-tripping also resembles co-training because the inbound-trip gives signals by helping to select high-quality translations from the *n*-best list which are then used to retrain the outbound-trip.

According to the round-tripping idea, in order to identify high-quality translations among many (potentially noisy) translations on the target side of the generated bilingual sentence pairs, two important points are essential:

- A candidate translation must be a well-formed sentence in the target language. (It should be an understandable as well as a clear even if it is not a correct translation of its corresponding source sentence).
- In addition to being a well-formed sentence on the target-side, the candidate translation should be high quality (accurate) for its corresponding source sentence as well.

We assume availability of: (1) monolingual datasets ($C_X$ and $C_Y$) for the source and target languages; and (2) two weak TMs that bidirectionally translate sentences from source and target languages. The goal of the round-tripping approach is to augment the accuracy of the two TMs by employing the two monolingual datasets instead of a bilingual text.

We start by translating a sample sentence in one of the monolingual datasets, as the outbound-trip (forward) translation to the target language. This step generates more bilingual sentence pairs between the source and target languages. We then translate the resulting sentence pairs backward through the inbound-trip translation to the original language. This step finds high-quality sentences throughout the entirety of the generated sentence pairs. Evaluating the results of this round-tripping approach will provide an indication of the quality of the two TMs, and will enable their enhancement, accordingly. This process is iterated for several rounds until both TMs converge.

We define $K_X$ as the number of sentences, in $C_X$ and $K_Y$ as the number of sentences in $C_Y$. We take $P(.|S;\Theta_{XY})$ and $P(.|S;\Theta_{YX})$ to be two neural TMs in which $\Theta_{XY}$ and $\Theta_{YX}$ are supposed as their parameters. We also assume the existence of two LMs for languages $X$ and $Y$ trained in advance either by using other resources, or using the monolingual data ($C_X$ and $C_Y$). Each LM takes a sentence as input and produces a real number, based on target-language fluency (LM correctness) together with translation accuracy (TM correctness). This score represents the confidence of the translation quality of the sentence in its own language.

We start with a sentence in $C_X$ and denote $S_{sample}$ as a translation output sample. This step has a score as follows:

$$R_1 = LM_Y(S_{sample}) \tag{12}$$

The $R_1$ score indicates the well-formedness of the output sentence in language $Y$.

Given the translation output $S_{sample}$, we employ the log probability value of $s$ recovered from the $S_{sample}$ as the score of the construction:

$$R_2 = \log P(S|S_{sample};\Theta_{YX}) \tag{13}$$

Note that $R_2$ is the sum of logs of the individual output word scores which are selected by softmax for $s$, and this sum is the figure-of-merit selected for the beam search used in our round-trip training algorithm (described below).

We adopt the LM score and construction score as the total reward score:

$$R_{total} = \alpha R_1 + (1-\alpha)R_2 \tag{14}$$

where $\alpha$ is an input hyper-parameter.

The total reward score is considered a function of $S$, $S_{sample}$, $\Theta_{XY}$ and $\Theta_{YX}$. To maximize this score, we optimize the parameters in the TMs utilizing Stochastic Gradient Descent (SGD) [27]. According to the forward TM, we sample the $S_{sample}$ and then compute the gradient of the expected score ($E[R_{total}]$) where $E$ is taken from $S_{sample}$:

$$\nabla_{\Theta_{XY}} E[R_{total}] = E[R_{total} \nabla_{\Theta_{XY}} \log P(S_{sample}|S; \Theta_{XY})] \quad (15)$$

$$\nabla_{\Theta_{YX}} E[R_{total}] = E[(1-\alpha) \nabla_{\Theta_{YX}} \log P(S|S_{sample}; \Theta_{YX})] \quad (16)$$

Algorithm 1 shows the round-trip training procedure.

Based on Equations (15) and (16), we are able to adopt any sampling approach to estimate the expected gradient. Considering that random sampling brings very large variance and sometimes unreasonable results in MT, we use beam-search [28] to achieve acceptable translations for SGD computation, i.e., we greedily generate *n*-best sample translations and use the averaged value on the beam-search results to approximate the true SGD.[1]

To start the round-trip training approach, the systems are initialized using warm-start TMs trained from initial small bilingual data. The goal is to see whether the round-tripping improves the baseline accuracy. We retrain the baseline systems by enlarging the initial small bilingual corpus: we add the optimized generated bilingual sentences to the initial parallel text. The new *enlarged* translation system contains both the initial and optimized generated bilingual text. For each translation task, we train the round-trip training approach.

**Algorithm 1:** Round-trip training for NMT

```
Input: Monolingual dataset in source and target
    languages ($C_X$ and $C_Y$), initial translation models in
    outbound and inbound trips ($\Theta_{XY}$ and $\Theta_{YX}$),
    language models in source and target languages ($LM_X$
    and $LM_Y$), trade-off parameter between 0 and 1 ($\alpha$),
    beam search size ($N$), learning rates ($\gamma_{1,t}$ and $\gamma_{2,t}$).
1: repeat:
2: $t = t + 1$.
3: Sample sentences $S_X$ and $S_Y$ from $C_X$ and $C_Y$
   respectively.
4: // Update model starting from language X.
   Set $S = S_X$.
5: // Generate top-N translations using $\Theta_{XY}$.
   Generate sentences $S_{sample,1}, \ldots, S_{sample,N}$.
6: for $n = 1, \ldots, N$ do
7: // Set LM score for $n^{th}$ sampled sentence.
   $R_{1,n} = LM_Y(S_{sample,n})$.
8: // Set TM score for $n^{th}$ sampled sentence.
   $R_{2,n} = logP(S|S_{sample,N}; \Theta_{YX})$.
9: // Set total score of $n^{th}$ sampled sentence.
   $R_n = \alpha R_{1,n} + (1-\alpha) R_{2,n}$.
10: end for
11: // SDG computing for $\Theta_{XY}$.
    $\nabla_{\Theta_{XY}} \hat{E}[R_{total}] = \frac{1}{N} \sum_{n=1}^{N}$
    $\left[R_n \nabla_{\Theta_{XY}} \log P(S_{sample,n}|S; \Theta_{XY})\right]$.
12: // SDG computing for $\Theta_{YX}$.
    $\nabla_{\Theta_{YX}} \hat{E}[R_{total}] = \frac{1}{N} \sum_{n=1}^{N}$
    $\left[(1-\alpha) \nabla_{\Theta_{YX}} \log P(S|S_{sample,n}; \Theta_{YX})\right]$.
13: // Model update.
    $\Theta_{XY} \leftarrow \Theta_{XY} + \gamma_{1,t} \nabla_{\Theta_{XY}} \hat{E}[R_{total}]$.
14: // Model update.
    $\Theta_{YX} \leftarrow \Theta_{YX} + \gamma_{2,t} \nabla_{\Theta_{YX}} \hat{E}[R_{total}]$.
15: // Update model starting from language $Y$.
    Set $S = S_Y$.
16: Go through lines 5 – 14 symmetrically.
17: until convergence.
```

## 6 Experimental framework

We apply the round-tripping approach to English-Spanish (En-Es) and Persian-Spanish (Pe-Es) translation tasks, and evaluate the results. For the high-resource scenario (En-Es) we utilize the English-Spanish bilingual corpora from WMT'18[2] [29] which contains 10M sentence pairs extracting from *Europarl*, *News-Commentary*, *UN* and *Common Crawl* collections. We also concatenate *newstest2012* and *newstest2013* as the validation set, and use *newstest2014* as the testing set. For the low-resource scenario (Pe-Es) we use the Persian-Spanish small bilingual corpora from *GNOME* corpus[3] [29] which contains about 500K parallel sentence pairs. We also use parallel sentences extracted from *Tanzil* collection[4] and *KDE4* [29] as the validation and testing datasets, respectively. For all experiments, we utilize 1M parallel sentences from the *OpenSubtitles2018* corpus[5] [29], as the monolingual data. For all Persian experiments, the Persian set contains explicit diacritic marks to

[1] We used beam sizes 500 and 1000.

[2] http://www.statmt.org/wmt18/translation-task.html

[3] http://opus.nlpl.eu/GNOME-v1.php

[4] http://opus.nlpl.eu/Tanzil-v1.php

[5] http://opus.nlpl.eu/OpenSubtitles-v2018.php

indicated vowels. (This simplification is discussed further in Section 7).

We implemented the DyNet-based model architecture [30] on top of *Mantis* [31] which is an implementation of the attentional sequence-to-sequence (Seq-to-Seq) NMT. For each language, we constructed a vocabulary with the most common 50K words in the parallel corpora, and OOV words were replaced with a special token <*UNK*>. For monolingual corpora, sentences containing at least one OOV word were removed. Additionally, sentences with more than 80 words were removed from the training set.[6] The encoders and decoders make use of Long Short-Term Memory (LSTM) with 500 embedding dimensions, 500 hidden dimensions. For training, we used the SGD algorithm as the optimizer. The batch size was set as 64 with 20 batches pre-fetched and sorted by sentence lengths.

We compare the system based on the optimized round-trip training (round-tripping) through two translation systems; the first one is the standard NMT system (baseline), and the second one is the system that generates pseudo bilingual sentence pairs from monolingual corpora to assist the training step (self-training). For the pseudo-NMT we used the trained NMT model to generate pseudo bilingual sentence pairs from monolingual text, removed sentences with more than 80 words (as above), merged the generated data with the original parallel training data, and then trained the model for testing. Each of the translation systems was trained on a single GPU until their performances stopped improving on the validation set. This approach required an LM for each language. We trained an RNN-based LM [32] for each language using its corresponding monolingual corpus. The LM was then fixed and the log-likelihood of a received message was utilized for scoring the TM.

We employ Bilingual Evaluation Understudy (BLEU) [33] (using *multi-bleu.perl* script from Moses) as the evaluation metric. BLEU is calculated for individual translated segments by comparing them with a data set of reference translations. The scores of each segment, ranging between 0 and 100, are averaged over the entire evaluation dataset to yield an estimate of the overall translation quality (higher is better). We additionally use: (1) *Accuracy*[7], also ranging between 0 and 100, which indicates the number of correct translations among the total number of translations (higher is better); and (2) *Perplexity*[8], which represents how certain we are that a predicted translation is correct (lower is better). The more certain we are, the less information we gain from the translation predictions, thus lower perplexity is better. It is fair to say that perplexity is the de facto standard for evaluating LMs. However, small perplexity cannot guarantee that the accuracy of predicted translations is actually high [34].

We note that accuracy and perplexity measure different aspects of translation quality. Accuracy refers to the degree of correctness of the highest-probability hypothesis. Perplexity measures the probability of the correct hypothesis or, more generally, how probable the observed data are. Accuracy is of great interest in our studies, but there are three challenges to computing accuracy: (1) supervised data are required; (2) a measure for degree of correctness is required; (3) optimization is difficult for accuracy because accuracy it is generally not a continuous function of the parameters (i.e., an epsilon change in the parameters may not change which hypothesis has the highest-probability). As such, in addition to BLEU, we have reported both Accuracy and Perplexity.

## 7 Results and analysis

The baseline systems for English-Spanish and Persian-Spanish are trained separately, while the round-tripping approach conducts joint training. We summarize the overall performances in Tables 1 and 2 (En-Es), and Tables 3 and 4 (Pe-Es):

As seen in Tables 1 to 4, the round-tripping systems outperform the others in all translation directions (English-to-Spanish and vice-versa as well as Persian-to-Spanish and vice-versa). The results demonstrate the effectiveness of the round-trip training approach. The base-

**Table 1:** Performance of NMT systems for English-to-Spanish.

| NMT systems | BLEU | Accuracy | Perplexity |
|---|---|---|---|
| **baseline** | 34.66 | 17.01 | 198.07 |
| **self-training** | 35.13 | 21.21 | 164.41 |
| **round-tripping** | 37.06 | 24.13 | 153.36 |

**Table 2:** Performance of NMT systems for Spanish-to-English.

| NMT systems | BLEU | Accuracy | Perplexity |
|---|---|---|---|
| **baseline** | 35.71 | 19.19 | 199.15 |
| **self–training** | 35.98 | 20.66 | 176.13 |
| **round–tripping** | 38.78 | 21.31 | 148.98 |

[6] The average sentence length is 47; an upper bound of 80 ensured exclusion of non-sentential and other spurious material.

[7] How correct is the highest-probability hypothesis?

[8] How probable is the correct hypothesis?

**Table 3:** Performance of NMT systems for Persian-to-Spanish.

| NMT systems | BLEU | Accuracy | Perplexity |
|---|---|---|---|
| **baseline** | 30.12 | 20.98 | 198.66 |
| **self–training** | 31.91 | 24.87 | 179.56 |
| **round–tripping** | 35.66 | 26.15 | 174.67 |

**Table 4:** Performance of NMT systems for Spanish-to-Persian.

| NMT systems | BLEU | Accuracy | Perplexity |
|---|---|---|---|
| **baseline** | 28.02 | 22.21 | 189.78 |
| **self–training** | 30.21 | 27.93 | 168.85 |
| **round–tripping** | 33.97 | 30.63 | 151.64 |

**Table 5:** Performance of the baseline and enlarged NMT systems for English-to-Spanish translation task.

| NMT systems | BLEU | Accuracy | Perplexity |
|---|---|---|---|
| **baseline** | 34.66 | 17.01 | **198.07** |
| **enlarged** | 38.77 | 24.59 | 148.18 |

**Table 6:** Performance of the baseline and enlarged NMT systems for Spanish-to-English translation task.

| NMT systems | BLEU | Accuracy | Perplexity |
|---|---|---|---|
| **baseline** | 35.71 | 19.19 | 199.15 |
| **enlarged** | 39.89 | 22.55 | 139.43 |

**Table 7:** Performance of the baseline and enlarged NMT systems for Persian-to-Spanish translation task.

| NMT systems | BLEU | Accuracy | Perplexity |
|---|---|---|---|
| **baseline** | 30.12 | 20.98 | 198.66 |
| **enlarged** | 34.85 | 22.02 | 160.11 |

**Table 8:** Performance of the baseline and enlarged NMT systems for Spanish-to-Persian translation task.

| NMT systems | BLEU | Accuracy | Perplexity |
|---|---|---|---|
| **baseline** | 28.02 | 22.21 | 189.78 |
| **enlarged** | 33.43 | 23.45 | 155.55 |

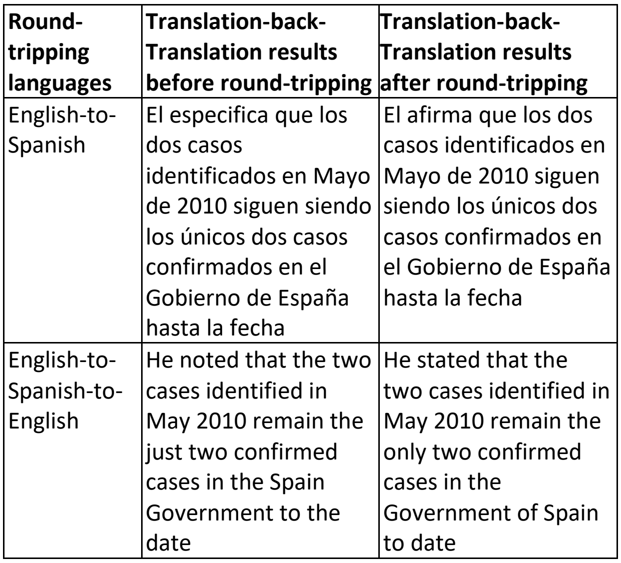
line systems outperform the self-training ones in all cases because of the noise in the generated bilingual sentences used by self-training. Upon further examination, this result might have been expected given that the aim of round-trip training is to optimize the generated bilingual sentences by selecting the high-quality sentences to get better performance over self-training systems. When the bilingual corpus size is small, the round-tripping makes a larger improvement. This outcome is an indication that the round-trip training approach makes effective use of monolingual data for these low-resource/high-resource pairings. (Further generalization of this point is discussed in Section 8.)

Tables 5 to 8 show the performance of the baseline alongside of the *enlarged* translation systems, in all directions tested in the original NMT experiments, where the enlarged systems utilize the training dataset of both the baseline and the round-tripping systems. As seen here, the performance of each enlarged NMT system is better than its corresponding baseline in all translation directions. The improvements indicate that the round-tripping systems are promising for tackling training data scarcity and also help for enhancing translation quality.

We note that Tables 7 and 8 show that the results of enlarged systems are somewhat lower than the corresponding results for round-tripping systems in Tables 3 and 4. This is to be expected, as the enlarged systems are essentially conventional NMT systems (not the round-trip algorithm), trained on more data (a larger bilingual corpus). Thus, these experiments support the effectiveness of the round-trip algorithm, not just for improving translation output, but for selecting high-quality sentence pairs to increase the size of the original small bilingual corpus (baseline) for the source and target languages, which is important in restricted resource scenarios.

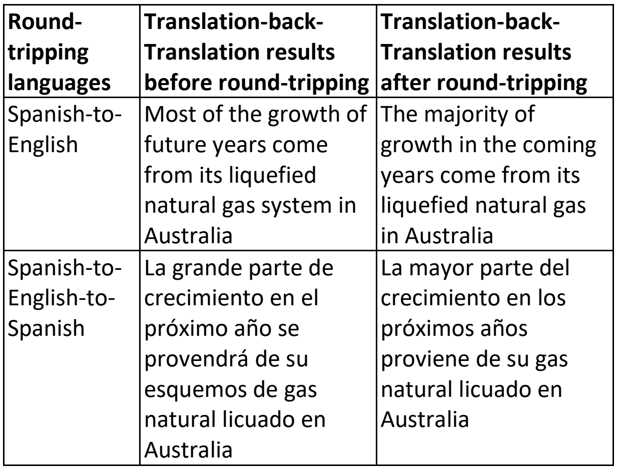
We further note that the translation systems are initialized with TMs trained from small bilingual data corpora. In the experiments, to transition from the initial model trained from bilingual data to the model training purely from monolingual data, we adopt the following strategy: at the starting point, for each mini batch, we train on half the sentences from monolingual data and half the sentences from bilingual data (sampled from the dataset used to train the initial model). The goal is to maximize the weighted sum of the reward score based on monolingual data as well as the likelihood on bilingual data. As training proceeds, we gradually increase the percentage of monolingual sentences in the mini batch, until no bilingual data were used at all. Additionally, although self-training outperforms baseline, its improvements are not significant. We expect that the quality of pseudo bilingual sentences generated from the monolingual data is not high, which limits the performance gain of self-training. One might need to select and filter the generated pseudo bilingual sentence pairs to get better performance for self-training systems.

Figures 1 and 2 provide examples in English, Spanish, and Persian, to compare the self-reconstruction output of models before and after round-tripping. It is clear that



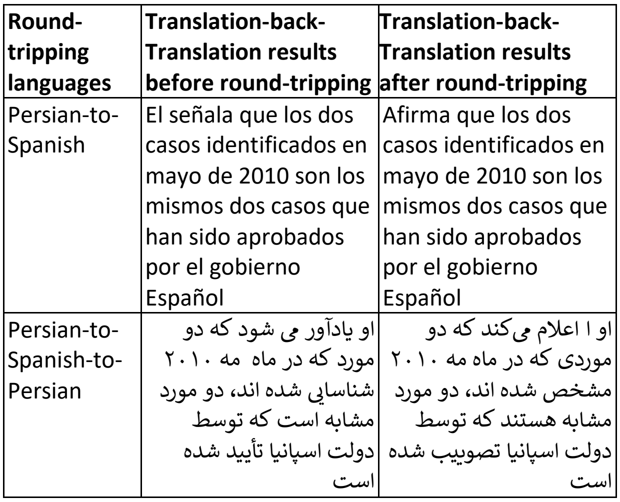
| Round-tripping languages | Translation-back-Translation results before round-tripping | Translation-back-Translation results after round-tripping |
|---|---|---|
| English-to-Spanish | El especifica que los dos casos identificados en Mayo de 2010 siguen siendo los únicos dos casos confirmados en el Gobierno de España hasta la fecha | El afirma que los dos casos identificados en Mayo de 2010 siguen siendo los únicos dos casos confirmados en el Gobierno de España hasta la fecha |
| English-to-Spanish-to-English | He noted that the two cases identified in May 2010 remain the just two confirmed cases in the Spain Government to the date | He stated that the two cases identified in May 2010 remain the only two confirmed cases in the Government of Spain to date |

**Figure 1:** Translation-back-translation performance through round-trip training approach in case of English-Spanish, for source English sentence: *He specifies that the two cases identified in May 2010 remain the only two cases confirmed in the Government of Spain to date*.



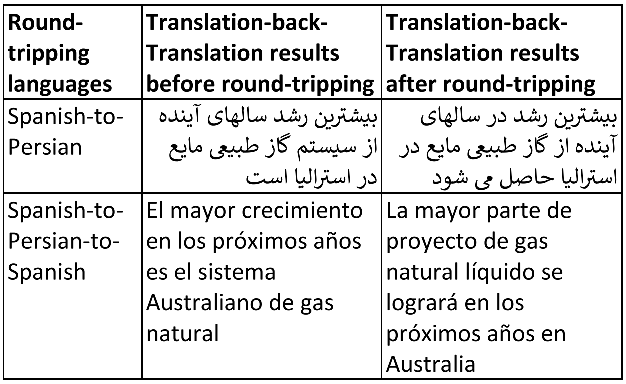
| Round-tripping languages | Translation-back-Translation results before round-tripping | Translation-back-Translation results after round-tripping |
|---|---|---|
| Spanish-to-English | Most of the growth of future years come from its liquefied natural gas system in Australia | The majority of growth in the coming years come from its liquefied natural gas in Australia |
| Spanish-to-English-to-Spanish | La grande parte de crecimiento en el próximo año se provendrá de su esquemos de gas natural licuado en Australia | La mayor parte del crecimiento en los próximos años proviene de su gas natural licuado en Australia |

**Figure 2:** Translation-back-translation performance through round-trip training approach in case of English-Spanish, for source Spanish sentence: *La mayoría del crecimiento en los próximos años provendrá de su esquemas de gas natural licuado en Australia*.



| Round-tripping languages | Translation-back-Translation results before round-tripping | Translation-back-Translation results after round-tripping |
|---|---|---|
| Persian-to-Spanish | El señala que los dos casos identificados en mayo de 2010 son los mismos dos casos que han sido aprobados por el gobierno Español | Afirma que los dos casos identificados en mayo de 2010 son los mismos dos casos que han sido aprobados por el gobierno Español |
| Persian-to-Spanish-to-Persian | او یادآور می شود که دو مورد که در ماه مه ۲۰۱۰ شناسايي شده اند، دو مورد مشابه است که توسط دولت اسپانیا تأیید شده است | او ا اعلام می‌کند که دو موردی که در ماه مه ۲۰۱۰ مشخص شده اند، دو مورد مشابه هستند که توسط دولت اسپانیا تصوییب شده است |

**Figure 3:** Translation-back-translation performance through round-trip training approach in case of Persian-Spanish, for source Persian sentence:
او مشخص می کند که دو موردی که در ماه مه ۲۰۱۰ شناسايي شده اند، همان دو موردی هستند که تا کنون در دولت اسپانیا تأیید شده اند



| Round-tripping languages | Translation-back-Translation results before round-tripping | Translation-back-Translation results after round-tripping |
|---|---|---|
| Spanish-to-Persian | بیشترین رشد سالهای آینده از سیستم گاز طبیعی مایع در استرالیا است | بیشترین رشد در سالهای آینده از گاز طبیعی مایع در استرالیا حاصل می شود |
| Spanish-to-Persian-to-Spanish | El mayor crecimiento en los próximos años es el sistema Australiano de gas natural | La mayor parte de proyecto de gas natural líquido se logrará en los próximos años en Australia |

**Figure 4:** Translation-back-translation performance through round-trip training approach in case of Persian-Spanish, for source Spanish sentence: *La mayoría del crecimiento en los próximos años provendrá de su esquemas de gas natural licuado en Australia*.

after round-tripping, the reconstruction is enhanced for all directions, i.e., English-Spanish and English-Spanish-English (Figure 1), Spanish-English and Spanish-English-Spanish (Figure 2), Persian-Spanish and Persian-Spanish-Persian (Figure 3), and Spanish-Persian and Spanish-Persian-Spanish (Figure 4).

For example, in Figure 1, *especifica* is more adequately conveyed as *afirma* in the English-to-Spanish round-trip, en route to the more natural sounding *stated* (as opposed to *noted*) for the English-to-Spanish-to-English round-trip. Analogously, in Figure 2, *majority of growth in the coming years* is a more fluent translation than *most of the growth of future years* in the Spanish-to-English round-trip, en route to the more natural sounding phrase *La mayor parte del crecimiento* (as opposed to *La grande parte de crecimiento*) for the Spanish-to-English-to-Spanish round-trip.

Similarly, in Figure 3, *señala* is more adequately conveyed as *afirma* in the Persian-to-Spanish round-trip, en route to the more natural sounding اعلام می کند (as opposed to یادآور می شود) for the Persian-to-Spanish-to-Persian round-trip. Analogously, in Figure 4, بیشترین رشد در سال های آینده is a more fluent translation than بیشترین رشد سال های آینده in the Spanish-to-Persian round-trip, en route to the more natural sounding phrase *La mayor parte de proyecto* (as opposed to *El mayor crecimiento*) for the Spanish-to-Persian-to-Spanish round-trip.

## 8 Conclusions and future work

In this paper, we applied a round-tripping approach based on retraining scenario to tackle training data scarcity in NMT systems. An exciting finding of this work is that it is possible to learn translations directly from monolingual data of the two languages. We employed either high-resource and low-resource language pairs examples and verified the hypothesis that, regardless of the amount of training resources, this approach outperforms the baseline. The results demonstrate that round-trip training is promising and better utilizes the monolingual data.

Although our experiments make use of only one language pair from each of the low-resource and high-resource conditions (Persian-Spanish and English-Spanish, respectively), the experimental results demonstrate a promising first step toward generalizing these results. We have demonstrated first that our approach works well under both the low-resource and high-resource conditions for these particular pairings. We view generalizing to additional low-resource pairings as an important future opportunity for investigation about the approach effectiveness on a larger scale for a broader set of language pairs.

Many Artificial Intelligence (AI) tasks are naturally in dual form. Examples are: (1) speech recognition paired with text-to-speech; (2) image captioning paired with image generation; and (3) question answering paired with question generation. Thus, a possible future direction would be to design and test the round-tripping approach for more tasks beyond MT. We note that round-tripping is not restricted to two tasks only. Indeed, the key idea is to form a closed loop so feedback signals are extracted by comparing the original input data with the final output data. Therefore, if more than two associated tasks form a closed loop, this approach can applied in each task for improvement of the overall model, even in the face of unlabeled data.

Another future direction is the handling of unvowelized Persian. Because Arabic script is used in our work with Persian, a set of diacritic marks is available to indicate vowels. These are generally omitted for more experienced readers, which makes Persian text more challenging. For our purposes, we factored out this challenge—using explicit diacritic marks—to focus on demonstrating the utility of the round-tripping approach for this initial study. Having a proof of concept leaves us in a better position for future work that addresses the additional (orthogonal) challenge of pairings that involve a language with unvowelized script.

**Acknowledgement:** We would like to express our sincere gratitude to Dr. Michael W. Mislove (Tulane University, USA), Dr. K. Brent Venable (University of West Florida, USA), Dr. Carola Wenk (Tulane University, USA), and Dr. Anastasia Kurdia (Tulane University, USA) for all their support. We also would like to acknowledge the Florida Institute for Human and Machine Cognition (USA).

## References

[1] Bahdanau D., Cho K., Bengio Y., Neural machine translation by jointly learning to align and translate, Proceedings of the International Conference on Learning Representations, 2015

[2] Ahmadnia B., Serrano J., Employing pivot language technique through statistical and neural machine translation frameworks: The case of under-resourced Persian-Spanish language pair, International Journal on Natural Language Computing, 2017, 6(5), 37–47

[3] Brants T., Popat A. C., Xu P., Och F. J., Dean J., Large language models in machine translation, Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, 2007, 858–867

[4] Gülçehre Ç., et al., On using monolingual corpora in neural machine translation, ArXiv, Vol. abs/1503.03535, 2015

[5] Ueffing N., Haffari G., Sarkar A., On using monolingual corpora in statistical machine translation, Journal of Machine Translation, 2008

[6] Sennrich R., Haddow B., Birch A., Improving neural machine translation models with monolingual data, Proceedings of the 54th Annual Meeting of Association for Computational Linguistics, 2016

[7] Ahmadnia B., Haffari G., Serrano J., Statistical machine translation for bilingually low-resource scenarios: A round-tripping approach, Proceedings of the 3rd IEEE International Conference on Machine Learning and Natural Language Processing, 2018, 261–265

[8] He D., et al., Dual learning for machine translation, Proceedings of the 30th Conference on Neural Information Processing Systems, 2016

[9] Jean S., Cho K., Memisevic R., Bengio Y., On using very large target vocabulary for neural machine translation, ArXiv, Vol. 412.2007, 2015

[10] Rowley H. A., Baluja S., Kanade T., Neural network-based face detection, IEEE Transactions on Pattern Analysis and Machine Intelligence, 1998, 20(1), 23–38

[11] McClosky D., Charniak E., Johnson M., Effective self-training for parsing, Proceedings of the Main Conference on Human Language Technology Conference of the North American Chapter of

the Association of Computational Linguistics, 2006, 152–159

[12] Blum A., Mitchell T., Combining labeled and unlabeled data with co-training, Proceedings of the 11th Annual Conference on Computational Learning Theory, 1998, 92–100

[13] Schwenk H., Investigations on large-scale lightly-supervised training for statistical machine translation, Proceedings of IWSLT, 2008, 182–189

[14] Ter-Sarkisov A., Schwenk H., Barrault L., Bougares F., Incremental adaptation strategies for neural network language models, Proceedings of the 3rd Workshop on Continuous Vector Space Models and their Compositionality, 2015, 48–56

[15] Sennrich R., Haddow B., Birch A., Neural machine translation of rare words with subword units, Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, 2016, 1715–1725

[16] Artetxe M., Labaka G., Agirre E., Cho K., Unsupervised neural machine translation, ArXiv, Vol. abs/1710.11041, 2017

[17] Lample G., Conneau A., Denoyer L., Ranzato M., Unsupervised machine translation using monolingual corpora only, Proceedings of the International Conference on Learning Representations, 2018

[18] Yang Z., Chen W., Wang F., Xu B., Unsupervised neural machine translation with weight sharing, Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, 2018, 46–55

[19] Wu H., Wang H., Pivot language approach for phrase-based statistical machine translation, Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, 2007, 856–863

[20] Ahmadnia B., Serrano J., Haffari G., Balouchzahi N., Direct-bridge combination scenario for Persian-Spanish low-resource statistical machine translation, Proceedings of the 7th International Conference of Artificial Intelligence and Natural Language, 2018, 67–78

[21] Ahmadnia B., Haffari G., Serrano J., Round-trip training approach for bilingually low-resource statistical machine translation systems, International Journal of Artificial Intelligence, 2019, 17(1), 167–185

[22] Dorr B. J., Machine translation divergences: A formal description and proposed solution, Computational Linguistics, 1994, 20(4), 597–633

[23] Dorr B. J., Pearl L., Hwa R., Habash N., 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, 2002

[24] Ahmadnia B., Serrano J., Haffari G., Persian-Spanish low-resource statistical machine translation through English as pivot language, Proceedings of the 9th International Conference of Recent Advances in Natural Language Processing, 2017, 24–30

[25] Hochreiter S., Schmidhuber J., Long short-term memory, Neural Computation, 1997, 9(8), 1735–1780

[26] Luong T., Sutskever I., Le Q. V., Vinyals O., Zaremba W., Addressing the rare word problem in neural machine translation, Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing, 2015, 11–19

[27] Sutton R., Mcallester D., Singh S., Mansour Y., Policy gradient methods for reinforcement learning with function approximation, Proceedings of Advances in Neural Information Processing Systems, 2000, 1057–1063

[28] Sutskever I., Vinyals O., le Q. V., Sequence to sequence learning with neural networks, Proceedings of Advances in Neural Information Processing Systems, 2014, 3104–3112

[29] Tiedemann J., Parallel data, tools and interfaces in OPUS, Proceedings of the 8th International Conference on Language Resources and Evaluation, 2012

[30] Mi H., Wang Z., Ittycheriah A., Supervised attentions for neural machine translation, Proceedings of the International Conference on Empirical Methods in Natural Language Processing, 2016, 2283–2288

[31] Cohn T., Huang C. D. V., Vymolova E., Yao K., Dyer C., Haffari G., Incorporating structural alignment biases into an attentional neural translation model, Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics Human Language Technologies, 2016, 876–885

[32] Mikolov T., Karafiat M., Burget L., Cernocky J., Khudanpur S., Recurrent neural network based language model, Proceedings of INTERSPEECH, 2010, 1045–1048

[33] Papineni K., Roukos S., Ward T., Zhu W.J., BLEU: A method for automatic evaluation of machine translation, Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, 2001, 311–318

[34] Ahmadnia B., Kordjamshidi P., Haffari G., Neural machine translation advised by statistical machine translation: The case of Farsi-Spanish bilingually low-resource scenario, Proceedings of the 17th IEEE International Conference on Machine Learning and Applications, 2018, 1209–1213