American Journal of Artificial Intelligence
2020; 4(2): 42-49
http://www.sciencepublishinggroup.com/j/ajai
doi: 10.11648/j.ajai.20200402.11
ISSN: 2639-9717 (Print); ISSN: 2639-9733 (Online)

Spanish-Turkish Low-Resource Machine Translation: Unsupervised Learning vs Round-Tripping

Tianyi Xu[1], Ozge Ilkim Ozbek[2], Shannon Marks[2], Sri Korrapati[2], Benyamin Ahmadnia[1, 3, *]

[1]Department of Computer Science, Tulane University of Louisiana, New Orleans, United States
[2]Department of Linguistics, Tulane University of Louisiana, New Orleans, United States
[3]Department of Linguistics, University of California, Davis, United States

Email address:
txu9@tulane.edu (T. Xu), oozbek@tulane.edu (O. I. Ozbek), smarks3@tulane.edu (S. Marks), skorrapati@tulane.edu (S. Korrapati), ahmadnia@tulane.edu (B. Ahmadnia)

[*]Corresponding author

To cite this article:
Tianyi Xu, Ozge Ilkim Ozbek, Shannon Marks, Sri Korrapati, Benyamin Ahmadnia. Spanish-Turkish Low-Resource Machine Translation: Unsupervised Learning vs Round-Tripping. *American Journal of Artificial Intelligence.* Special Issue: *Machine Translation for Low-Resource Languages*. Vol. 4, No. 2, 2020, pp. 42-49. doi: 10.11648/j.ajai.20200402.11

Received: May 28, 2020; **Accepted**: June 18, 2020; **Published**: July 23, 2020

Abstract: The quality of data-driven Machine Translation (MT) strongly depends on the quantity as well as the quality of the training dataset. However, collecting a large set of training parallel texts is not easy in practice. Although various approaches have already been proposed to overcome this issue, the lack of large parallel corpora still poses a major practical problem for many language pairs. Since monolingual data plays an important role in boosting fluency for Neural MT (NMT) models, this paper investigates and compares the performance of two learning-based translation approaches for Spanish-Turkish translation as a low-resource setting in case we only have access to large sets of monolingual data in each language; 1) Unsupervised Learning approach, and 2) Round-Tripping approach. Either approach completely removes the need for bilingual data or enables us to train the NMT system relying on monolingual data only. We utilize an Attention-based NMT (Attentional NMT) model, which leverages a careful initialization of the parameters, the denoising effect of language models, and the automatic generation of bilingual data. Our experimental results demonstrate that the Unsupervised Learning approach outperforms the Round-Tripping approach in Spanish-Turkish translation and vice versa. These results confirm that the Unsupervised Learning approach is still a reliable learning-based translation technique for Spanish-Turkish low-resource NMT.

Keywords: Computational Linguistics, Natural Language Processing, Neural Machine Translation, Low-Resource Languages, Unsupervised Learning, Round-Tripping

1. Introduction

Learning-based translation with monolingual data is an undesirable task due to multiple possible outcomes in the mapping of source language and target language sentences [1]. Machine Translation (MT) proves to be a front runner of recent successful advances in tackling challenges within the field of Natural Language Processing (NLP), but the reliance on large, high-quality sets of bilingual data for current learning algorithms still poses a major issue. While many ambiguities arise from the mapping of source and target sentences with the use of monolingual data, learning algorithms leveraged with such monolingual data for both languages increase translation accuracy. The currently employed model for MT research is an attention-based encoder-decoder model [2]. The core of this approach is translation via a learning-based model trained on neural networks. This model has gained much attention in the recent state-of-the-art research in which the attention-based framework improves upon the encoder-decoder Neural Machine Translation (NMT) model by allowing for variable-length input-output pairs for source sentences and their target translations. The use of NMT can be considered less preferable in low-resource translation tasks as a result of the obvious drawback of NMT, a strong reliance on a high-volume quality parallel corpus. Low-resource language-pair translation is a problem for all MT that must be addressed.

Even in Statistical Machine Translation (SMT), low-resource language pairs pose problems in tasks such as rare-word translation. Investigating techniques that apply to low-resource languages in both NMT and SMT is important; however, as NMT benefits more from this approach only NMT is explored in this paper.

Current MT research employs and improves unsupervised learning [1] and round-tripping [3-5], both of which incorporate monolingual data. There are several common principles underlying their success to be identified, though these approaches differ in technical aspects. These approaches generally use an inferred bilingual dictionary to initialize their system, which then relies on the data for training by leveraging strong language models. The final common component of these system algorithms is that they create a supervised problem from the unsupervised one through the generated pseudo-bilingual sentence pairs to constrain the latent representations produced. These representations are to be shared across both the source and target languages. Investigating the effectiveness of the mentioned learning-based approaches, unsupervised learning, and round-tripping, on overall translation quality in the low-resource conditions over the attention-based Neural MT (Attentional NMT) model can be done by employing the low-resource language pair Spanish-Turkish. The linguistic differences between Spanish and Turkish motivated the use of this language pair as the case study for this paper. Spanish and Turkish being from different language families and having significant differences in their linguistic parameters and features may pose a challenge for MT tasks.

Turkish, an agglutinating language, makes use of affixes to convey information such as a person, number, and tense. Since each word carries meaning through the use of bound morphemes, word order in Turkish is not as strict as in Spanish. This structural variety leads to alternative forms of sentences describing a single semantic event and information but with subtle distinctions that need to be carefully analyzed in context. Questions in Turkish are formed by adding a specific free morpheme, or one of its allomorphs, at the end of the sentence or after the constituent in the sentence that is being questioned. Turkish sentences are not required to have subjects since the information on the person and number is already contained in the verb with morphological markers. Questions in Spanish follow the exact same word order as declarative sentences with no additional morphemes. The information about whether the sentence is declarative or a question is conveyed with intonation in spoken language and two question marks (one at the end of the sentence and an inverse one at the beginning of the sentence) in written language.

In Spanish, nouns are usually preceded by articles that agree with the gender and number aspect, such as *el/la/los/las* and *un/una/unos/unas*. Turkish lacks articles, such as "the" and "a", so there is no equivalent for *el*/*la*/*los*/*las* and *un*/*una*/*unos*/*unas* in Spanish. This lack of determiner immediately adds the requirement of gender identification in the Turkish-Spanish translation direction. Another important feature is that the third singular person pronoun is not gendered in Turkish, and they share one pronominal representation *o*; therefore, knowing the gender of the subject when translating the sentence to another language is usually not possible. On the other hand, Spanish is a synthetic language in which verb forms agree with their subject in features such as gender, most notably. The canonical word order in Turkish is Subject-Object-Verb, and adjectives always immediately precede nouns. Spanish uses the Subject-Verb-Object order and the subject is not required to be present in the sentence since the verb already conveys that information.

NMT applications are used in this translation task by employing Unsupervised Learning and Round-Tripping to handle some of the issues in this alignment and translation task, though experimentally investigating the language pair is not the salient contribution of this research, which can be attributed to the investigation of two learning approaches in low-resource NMT tasks. Unsupervised Learning is a type of Machine Learning (ML) that looks for previously undetected patterns in a dataset with no pre-existing labels and minimum human supervision. Unsupervised MT can be accomplished by leveraging the initialization of the translation models, language modeling, and iterative back-translation. However, improving translation quality between Spanish and Turkish languages is still challenging. Round-Tripping is an approach that involves the two tasks of inbound-trip and outbound-trip. Recently, a round-tripping approach incorporated with dual learning [6] solved problems for automatic learning. The Round-Tripping method considers translation models and makes improvements by the unlabeled data in the training dataset. Our results demonstrate that the Unsupervised Learning approach outperforms the Round-Tripping approach in Spanish-Turkish translation and vice versa that confirms that the unsupervised approach is still a reliable learning-based translation technique for Spanish-Turkish low-resource NMT.

This paper is organized as follows; Section 2 investigates the previous related work. Section 3 reviews the related mathematical background. Section 4 describes the methodologies. The experimental framework is presented in Section 5. The results analysis and evaluation are covered in Section 6. Conclusions are provided in Section 7.

2. Related Work

Due to the unavailability of a sufficient amount of bilingual data for training low-resource NMT systems, several methods have been proposed to improve the quality of translation. One of the most popular methods is using a third language which has more bilingual data available with the two languages and can serve as a bridge between them [7-11]. In this method, the source language is first translated to the pivot (bridge) language as an intermediate step, and then to the target language. However, other more advanced approaches have been shown to get better results in terms of enhancing the baseline model, like a teacher-student

framework [12, 13]. Other attempts include using the monolingual data in both source and target languages in combination with the parallel data in order to train the NMT system. One method to do this is to create synthetic parallel data by back translating the monolingual data [14]. However, a better method to train the NMT system based on monolingual data is the Round-Tripping approach [3-5], where two translation systems are trained to translate in opposite directions, allowing them to teach each other through the learning process. This method, however, initially requires at least a small amount of parallel data, unlike the unsupervised approach, which does not use any bilingual data.

Much of the recent research in MT has focused on the NMT approach, in which a hidden-layer neural network is used for the translation task. NMT offers many advantages over its Statistical MT (SMT) counterpart, such as minimal domain knowledge requirement, ready-to-implement beam search decoding, avoidance of storing large phrase tables by use of a recurrent neural network (RNN) [15]. Initially, NMT was applied in formal domain translation, but since then NMT application has expanded to low-resource settings, spoken language domains, and other translation tasks [15]. Due to these clear advantages, much work has been done to develop a sound NMT framework. The current state-of-the-art architecture for NMT uses the attentional encoder-decoder based framework. The encoder-decoder setup allows for the neural network translation where the attention-based aspect allows for variable input-output pair lengths, allowing translations to be more flexible in the lengths of words [15]. Though the attentional model has been compared to the alignment algorithm in SMT, there is no certainty in the attentional models' actual calculations [16].

Because NMT focus was preceded by SMT focus in MT literature, much effort has been put into comparing these two methods. SMT has been shown to outperform NMT at out-of-domain translation, low-resource settings, and in the translation of sentences of 60 words or longer, but NMT was able to out-perform SMT in high-resource settings and low-frequency word translations [16]. Although much of the work done in regard to the performance ability of NMT suggests that it is a stronger approach than Phrase-Based SMT (PBSMT), some of the research still seeks to revisit this claim.

The fact that NMT performance can be worse than PBSMT resulting from a lack of adaptation to low-resource settings calls for a reassessment of the validity of the previously absolute argument that NMT outperforms PBSMT [17]. NMT approaches have been outperformed by the PBSMT approach specifically in low-resource settings because of the nature of NMT being greatly dependent on the corpus size and quality. The NMT performance results, being worse in out-of-domain or low-resource environments, have been attested to its heavy reliance on the training data, resulting in a bad level of out-of-date error and unintended outcomes of beam search [16]. This conclusion then suggests that data augmentation of the parallel corpus may help improve the performance of NMT systems in these situations. Another reasonable response is to train the NMT system on other data in addition to the parallel corpus.

Low-resource NMT is very sensitive to hyperparameters such as vocabulary size, and word dropout, but NMT systems can be competitively trained without relying on auxiliary resources through the sole use of a parallel corpus [17]. This has practical relevance for languages where large amounts of monolingual data involving related languages are not available. A related study was focused on only using parallel data; however, the results are also relevant for work on using auxiliary data to improve low-resource MT [17]. Although their research began with reassessing the validity of the claims being made about NMT, their results did demonstrate that NMT is a suitable choice in low data settings and can outperform PBSMT with far less parallel training data than previously claimed.

Relevant to low-resource tasks, there has been an investigation into how suitable translation can be achieved when the only source and target monolingual corpora are available. The vast majority of language pairs have very little parallel data, so there is a need for an expansion in MT application to incorporate or leverage monolingual data. Initializing the MT system with an inferred bilingual dictionary, then leveraging strong Language Models (LMs) by training the sequence-to-sequence system as a denoising auto-encoder, turning the unsupervised problem into a supervised one by automatic generation of sentence pairs via back-translation, and finally constraining the latent representations produced by the encoder to be shared across the two languages successfully provide remarkable empirical results, especially considering the fully unsupervised setting [18].

Improving the pseudo-parallel corpus method used for low-resource NMT, where a parallel corpus is created by translating the monolingual data in both languages with an already existing MT system, by filtering the pseudo-parallel corpus using back-translation for evaluation has proven effective [19]. This method is found to increase the translation quality in MT for low-resource languages, in which the translation quality heavily depends on the quantity and quality of the limited amount of training data.

Other research focuses on improving NMT in low-resource settings by improving the parallel corpus itself. This line of research attempts to leverage the training data for the creation of new training data using rare words. One instance of this follows a similar method in computer vision research, but a key difference is that the augmentation does not preserve semantic content [20]. Using Translation Data Augmentation as a novel approach that modifies existing sentences in a parallel corpus, rare words are substituted for similar words such that more examples of the rare word were included in the training data. This is similar to computer vision research in which an image undergoes transformations to produce multiple distinct versions of the same image, allowing for the system to better train. Unlike in vision research, the labels are weakly preserved where the target

and source sentences of the augmentation are of different meanings to the original, while also keeping the translation accurate. Whereas in computer vision and image alteration does not change the image's semantic label, in MT a word alteration should change the semantic label.

Round-Tripping can be applied to both source and target languages. It can enable the data to play a role that is similar to the parallel bilingual data [3-5]. This innovative approach helps in the gradual reduction of the requirement on parallel bilingual data during the training process. Round-Tripping is shown to help solve the training data scarcity problem by making effective use of monolingual data, and this approach has been shown to improve NMT environments as well [5]. The Round-Tripping approach produces informative feedback to update translation models until convergence [3].

Unsupervised Learning is an ML technique drawing inferences from data sets consisting of input data without labeled responses. Applying this method in MT, models are trained without using any labeled data. Previous work uses encoder-decoder structure in unsupervised learning MT systems [21]. The outputs of encoders for two languages can be constrained and modified into the same latent space [33]. In addition, encoders can be improved by denoising auto-encoders and with adversarial training methods. After that, iterative back-translation is applied to parallel data to help cross-lingual training. Unsupervised MT works via suitable initialization of the translation models, language modeling, and iterative back-translation [18]. These three aspects underlie the success of much Unsupervised MT research. PBSMT systems often outperform NMT systems in the fully unsupervised setting and by combining these systems they can greatly outperform previous approaches from the literature [18]. The general focus in MT research on improving NMT gives hope to the potential of unsupervised MT in further studies.

Several deficiencies of existing unsupervised SMT approaches were identified and addressed by exploiting sub-word information, developing a theoretically well-founded unsupervised tuning method [22]. Large improvements over the previous state-of-the-art in unsupervised MT are seen with these upgrades, but still, the increasing popularity of NMT calls for the extension of the unsupervised approach and improvements from previous SMT research to NMT research.

In addition, Unsupervised Learning is extended to low-resource languages. One instance is to propose a model that takes sentences from monolingual corpora in two different languages [1]. Without using any labeled data, this unsupervised model can be reconstructed in both languages from this shared feature space and can translate more effectively.

3. Mathematical Background

In an NMT system, the main components are the encoder and the decoder, which are both Recurrent Neural Networks (RNN) [2]. The encoder takes a source sentence and transforms it into an internal representation, which is then taken by the decoder and transformed into one or more target sentences.

The internal representation consists of a sequence of vectors. The forward-RNN returns the following forward hidden vectors:

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

and the backward-RNN returns the following backward hidden vectors:

$$\overleftarrow{h_j} = f(\overleftarrow{h_{j-1}}, x) \tag{2}$$

Source vectors are obtained by the concatenation of the forward and backward vectors as the following:

$$h_j = [\overrightarrow{h_j}; \overleftarrow{h_j}] \tag{3}$$

The decoder uses its hidden state and an output context to output target words in a recurrent manner. The decoder takes the input source vectors (starting with the concatenated vector h_j) and outputs target words one-by-one by checking the conditional probability of each potential output. The conditional probability is formulated as follows, where x is the source sentence, y is the target sentence, and h is the internal representation:

$$P(y|x) = \prod_{i=1}^{I} P(y_i|y_{<i}, x) \tag{4}$$

The conditional probability of the whole target word is formulated as the following, where y_i is the target word, f is a nonlinear function, and d_i is the decoder's hidden state at the i^{th} step:

$$P(y_i|y_{<i}, x) = softmax\,[f(d_i, y_{i-1}, c_i)] \tag{5}$$

The hidden state $d_i$is modeled as following, where g is an RNN function and $c_i$is a context vector:

$$d_i = g(d_{i-1}, y_{i-1}, c_i) \tag{6}$$

The function g updates its state vector by taking the previous state vector and the output word as input. The context vector takes the weighted sum of the source vectors in order to get source inputs, taking the d_i at the top layer of a Long Short-Term Memory (LSTM) [23] stack as input. The context vector serves to represent the relevant information from the source in order to facilitate the prediction of the target word y_i.

One possible model to derive a context vector is the following, which we are going to use in this work:

$$c_i = \sum_{j=1}^{J} \alpha_{i,j}, h_j \tag{7}$$

here, $\alpha_{i,j}$ represents a weight of the source vector j at time step i and is derived by the following score function (discussed further in [24]):

$$\alpha_{i,j} = \frac{exp(score(d_i, h_j))}{\sum_{j'=1}^{J} exp(score(d_i, h_{j'}))} \tag{8}$$

The system is trained using N training sentences in order the maximize the log-likelihood probability described as following, where x^n and y^n represent source-target pairs:

$$L_\theta = arg\ \max_\theta \left(\frac{1}{N}\sum_{n=1}^{N} logP_\theta(x^n)\right) \quad (9)$$

4. Methodology

Our methodology relies on both the Unsupervised Learning approach as well as the Round-Tripping approach.

The Unsupervised Learning approach follows a standard encoder-decoder architecture through an attention mechanism [2]. NMT systems are employed to predict the translations in a parallel corpus. When there is the case that researchers only have access to monolingual corpora, it becomes necessary to rely on the architecture modifications to a training system such as a dual structure [6], a shared encoder, and a fixed embedding in the encoder, in order to train the entire system in an unsupervised learning approach. This is done using two strategies: 1) denoising and 2) back-translation.

1) Denoising: This system can be directly trained to reconstruct its own input exploiting the shared encoder and the dual structure of MT. This is to confirm that the entire system can be optimized to receive an input sentence in a given language, encode it using the shared encoder, and reconstruct the input sentence using the decoder of that language. Since we utilize pre-trained cross-lingual embeddings in the shared encoder, the encoder is expected to learn to compose the embeddings of both languages (source and target) in a language-independent fashion. Similarly, each decoder should learn to decompose this representation into its corresponding language. At inference time, we replace the decoder with that of the target language, so it generates the translation of the input text from the language-independent representation given by the encoder. This ideal behavior is severely compromised by the fact that the resulting training procedure is essentially a trivial copying task. Because of this, the optimal solution for this task would not need to capture any real knowledge of the languages involved, because there would be many degenerated solutions that blindly copy all the elements in the input sequence. If this were the case, the system would at best make very literal word-by-word substitutions when used to translate from one language to another at inference time.

2) Back-translation: In spite of the denoising strategy, the given training procedure is still a copying task with some synthetic alterations that involve a single language at each time, without considering our final goal of translating between two languages. In order to train our system in a true translation setting without violating the constraint of using nothing but monolingual corpora, we propose to adapt the back-translation approach proposed by Sennrich et al. [14] to our scenario. More concretely, given an input sentence in one language, we use the system in inference mode with greedy decoding to translate it to the other language [14]. This way, we obtain a generated bilingual sentence pair and train the system to predict the original sentence from this synthetic translation.

Contrary to back-translation, which is considered standard, and uses an independent model to back-translate the entire corpus at one time, we exploit the dual structure of the proposed architecture to back-translate each mini-batch using the model that is being trained itself. This way, as training progresses and the model improves, it will produce better synthetic sentence pairs through back-translation, which will serve to further improve the model in the following iterations. During training, we alternate these different training objectives from mini-batch to mini-batch. This way, given two languages source and target, each iteration would perform one mini-batch of denoising for source, another one for target, one mini-batch of back-translation from source to target, and another one from target to source.

Round-Tripping involves two related translation tasks: 1) the outbound-trip (source-to-target), and 2) the inbound-trip (target-to-source). 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 translation models. This approach enables the monolingual data to play a role that is similar to the bilingual data. This helps in the gradual reduction of the requirement on bilingual text during the training phase. In Round-Tripping, the first translation system understands the source language and it sends a message in this language to the other translation system. The second translation system understands the target language. After checking the message, it sends a notification to the first translation system. After receiving the message from the second translation system, the first one 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 translation models, and as a result of this feedback, they make the required changes.

According to the Round-Tripping approach, 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: 1) a candidate translation must be a well-formed sentence in the target language, and 2) the candidate translation should be high-quality for its corresponding source sentence as well. Assume the availability of (1) monolingual datasets for the source and target languages; and (2) two weak translation models that bi-directionally translate sentences from source and target languages.

Since the Round-Tripping approach aims at augmenting the accuracy of the two translation models by employing the two monolingual datasets instead of a bilingual text, a sample sentence is first translated in one of the monolingual data sets, as the outbound-trip (forward) translation to the target language. This step generates more bilingual sentence pairs between the source and target languages. Then the resulting sentence pairs are translated 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 a round-tripping approach will provide

an indication of the quality of the two translation models, and will enable their enhancement, accordingly.

5. Experimental Framework

The experiments were conducted employing the Spanish-Turkish dataset collected from GNOME and Ubuntu bilingual corpora [25] (1.1K sentences and 8.95K words). The systems are evaluated using tokenized BLEU scores as computed by the *multi-bleu.perl* script. As for the training data (~700 sentences), the proposed systems have been tested under two different settings:

1) Unsupervised-Learning; that is one of the scenarios under consideration in this work, where the system has only access to monolingual corpora. For that purpose, we utilized the Europarl corpus.
2) Round-Tripping; that is the other scenario under consideration in our work, where the system has access to the monolingual corpus (Europarl) [26] as well as the bilingual corpus (GNOME + Ubuntu).
3) Baseline; where the system has only access to the bilingual corpora (direct translation). For that purpose, we employed the collection of GNOME and Ubuntu as our bilingual training dataset.

For the corpus preprocessing, we performed tokenization and truecasing using standard Moses tools. Byte Pair Encoding (BPE) was applied [27]. While BPE is known to be an effective way to overcome the rare word problem in standard NMT, it is less clear how it would perform in our more challenging unsupervised learning scenario, as it might be difficult to learn the translation relations between sub-word units. For that reason, experiments at the word level in the unsupervised learning scenario were also run, limiting the vocabulary to the most frequent 10K tokens and replacing the rest with a special token <*UNK*>. We accelerate training by discarding all sentences with more than 30 elements (either BPE units or actual tokens). We utilized the monolingual corpora to independently train the embeddings for each language using *word2vec* [28]. The training step of the system is done with the cross-entropy loss function and a batch size of 25 sentences (for the unsupervised learning system, denoising was used by itself as well as alongside back-translation, in order to better analyze the contribution of the latter). The DyNet-based model architecture [29] was implemented on top of *Mantis* [30] which is an implementation of the attentional NMT.

In the experiments based on the Round-Tripping approach, in case of using monolingual corpora, sentences containing at least one Out-Of-Vocabulary (OOV) word were removed. The encoders and decoders make use of LSTM with 500 embedding dimensions and 500 hidden dimensions. Stochastic Gradient Descent (SGD) was employed as the optimizer with a learning rate of 0.1. Using the Mantis implementation, training each system took about one week on a single GPU. Also, at training time, *greedy* decoding was used for back-translation, while actual inference at test time was done using beam-search with a beam size of 10 [14, 31].

BLEU [32] is employed 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). Additionally, *ACCURACY* is used, also ranging between 0 and 100, which indicates the number of correct translations among the total number of translations (higher is better) [33].

6. Results Evaluation

Tables 1 and 2 show the BLEU scores as well as the ACCURACY scores obtained by all the tested variants in Spanish-Turkish translation and vice versa.

Table 1. *Spanish-Turkish translation results in applying Unsupervised-Learning systems including "Denoising", "Back-Translation", and "BPE" compared to state-of-the-art including "Baseline" and "Round-Tripping".*

Translation Systems	BLEU	ACCURACY
Baseline	14.51	12.04
Unsupervised / Denoising	13.33	11.45
Unsupervised / back translation	21.97	18.77
Unsupervised / BPE	21.69	18.30
Round-Tripping	19.86	16.54

Table 2. *Turkish-Spanish translation results in applying Unsupervised-Learning systems including "Denoising", "Back-Translation", and "BPE" compared to state-of-the-art including "Baseline" and "Round-Tripping".*

Translation Systems	BLEU	ACCURACY
Baseline	15.67	12.92
Unsupervised / Denoising	13.92	11.88
Unsupervised / back translation	22.19	19.04
Unsupervised / BPE	21.84	18.51
Round-Tripping	20.17	16.89

As seen in the tables, the unsupervised systems (including back-translation and BPE) demonstrate good results, especially considering that they were trained on monolingual corpora. These results confirm the notion that the unsupervised systems are capable of achieving much more than literal translations since monolingual corpora are stronger than the baseline system of word-by-word substitutions. The results demonstrate that an unsupervised system can learn to account for the internal structure of languages as well as learn how to employ context information of the given language.

The tables provided above also express that back-translation is essential for an unsupervised NMT system to work properly. The results from the denoising technique alone are below the baseline, while the results from the back-translation technique demonstrate noteworthy improvements. Test accuracies from our experiments also confirm this claim. They show that the unsupervised system containing denoising alone obtains a per-word accuracy of 11.45 for Spanish-to-Turkish (11.88 vice versa), whereas the one with back-translation achieves a better accuracy of 18.77 (19.04 vice versa).

The unsupervised training procedure would not work employing back-translation alone (without denoising). This

nuance is because the initial translations would be meaningless sentences produced by a random NMT model, encouraging the system to follow the structure of completely ignoring the input sentence and simply learning a language model of the target language.

Furthermore, the results show that BPE is slightly beneficial. BPE is a translation system that does not handle OOV's in any way because it is a word-level system, so it does fail to translate unknown words, making the beneficial quality seen in our results a surprising one. Although BPE generally does not translate unknown words, these findings may suggest the ability of this system to translate some unknown words as well as display some new errors. BPE is of little help when translating infrequent named entities while a baseline NMT model would easily learn to copy the named entities using BPE.

From the results, both denoising and back-translation can be deduced to play an important role during training. Denoising enforces the system to capture broad word-level equivalences, while back-translation enforces it to learn more subtle relations in an increasingly natural setting.

The results of the Round-Tripping systems demonstrate that this model can exploit a small bilingual dataset. As seen in the tables, these results are better than the baseline and comparable with unsupervised/BPE NMT models. In fact, the round-tripping systems demonstrate better results than the comparable NMT systems trained in the full bilingual corpus in almost every case. The hypothesis is that this improvement is because the domain of both the monolingual and the bilingual corpora utilized match that of the test set.

The relatively poor results of the comparable NMT models confirm that the additional constraints in these systems (which were introduced to enable unsupervised learning), may also be a factor limiting its potential performance. Due to this possibility, it may be that the system could be further improved during training in many ways. For example, one improvement would be to employ fixed cross-lingual embeddings in the encoder. This may result in improvement because doing so enforces the encoder to utilize a common word representation for both languages, which is necessary for the early stages of training. However, this application may also limit everything that is learned throughout the process. Because of this caveat, progressively updating the weights of the encoder embeddings as training progresses may be necessary. Another possibility would be decoupling the shared encoder into two independent encoders during training, or progressively reducing the noise level.

7. Conclusions

This paper investigated the impact of a reliable approach to train an Attentional NMT system in a completely unsupervised manner. This project was created based on existing work on unsupervised cross-lingual embeddings and incorporated them in a modified attentional encoder-decoder model. Employing a shared encoder with fixed cross-lingual embeddings, the system was able to be trained from monolingual corpora alone, combining denoising, and back-translation. The experimental results show the effectiveness of applying the unsupervised scenario, obtaining considerable improvements in the BLEU scores as well as the ACCURACY scores over a baseline system and a system based on a round-trip training approach. The results also confirm the quality of our unsupervised system; it is able to model complex cross-lingual relations and produce high-quality translations. Furthermore, the research shows that combining the unsupervised method with a small bilingual dataset can bring further improvements, showing its potential interest beyond the strictly unsupervised scenario.

Acknowledgements

We would like to express our sincere gratitude to Dr. Michael W. Mislove (Tulane University of Louisiana, USA) for all his unconditional support.

References

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

[2] 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.

[3] 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.

[4] Ahmadnia B., Dorr B. J., Augmenting Neural Machine Translation through Round-Trip Training Approach, Open Computer Science, 2019, 9 (1): 268-278.

[5] 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.

[6] He D., Xia Y., Qin T., Wang L., Yu N., Liu T., Ma W., Dual learning for machine translation, Proceedings of the 30th Conference on Neural Information Processing Systems, 2016.

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

[8] Ahmadnia B., Serrano J., Direct translation vs. pivot language translation for Persian-Spanish low-resourced statistical machine translation system, Proceedings of the 18th International Conference on Artificial Intelligence and Computer Science, 2016.

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

[10] 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.

[11] Ahmadnia B., Serrano, Haffari G., Balouchzahi NM., Direct-bridge combination scenario for Persian-Spanish low-resource statistical machine translation, Proceedings of Artificial Intelligence and Natural Language, 2018, 67-78.

[12] Firat O., Sankaran B., Al-Onaizan Y., Yarman Vural F. T., Cho K., Effective approaches to attention-based neural machine translation, Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2016, 268–277.

[13] Chen, Y., Liu Y., Cheng Y., Li V., A teacher-student framework for zero- resource neural machine translation, Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, 2017, 1925-1935.

[14] 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.

[15] 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.

[16] Koehn P., Knowles R., Six Challenges for Machine Translation, Proceedings of the First Workshop on Neural Machine Translation, 2017.

[17] Sennrich R., Zhang B., Revisiting Low-Resource Neural Machine Translation: A Case Study, Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019, 211-221.

[18] Lample G., Ott M., Conneau A., Denoyer L., Ranzato M., Phrase-based & neural unsupervised machine translation, Proceedings of EMNLP, 2018, 5039-5049.

[19] Imankulova, A., Sato, T., Komachi, M., Improving Low-Resource Neural Machine Translation with Filtered Pseudo-parallel Corpus, Proceedings of the 4th Workshop on Asian Translation, 2017, 70-78.

[20] Fadaee M., Bisazza A., Monz C., Data augmentation for low-resource neural machine translation, Association for Computational Linguistics, 2019.

[21] Artetxe, M., Labaka, G., Agirre, E., & Cho, K. 2017. Unsupervised neural machine translation. arXiv preprint arXiv: 1710.11041.

[22] Artetxe M., Labaka G., Agirre E., An effective approach to unsupervised machine translation, Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019, 194-203.

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

[24] Luong T., Pham H., Manning C. D., Effective approaches to attention-based neural machine translation, Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2015, 1412–1421.

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

[26] Koehn P., Europarl: A parallel corpus for statistical machine translation, Proceedings of the 10th Machine Translation Summit, 2005, 79-86.

[27] 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, 2016b, 1715-1725.

[28] Mikolov T., Sutskever I., Chen K., Corrado G., Dean J., Distributed representations of words and phrases and their compositionality. In Advances in Neural Information Processing Systems 26, 2013, 3111-3119.

[29] 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.

[30] 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.

[31] 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.

[32] 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.

[33] 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.