EMPLOYING PIVOT LANGUAGE TECHNIQUE THROUGH STATISTICAL AND NEURAL MACHINE TRANSLATION FRAMEWORKS: THE CASE OF UNDER-RESOURCED PERSIAN-SPANISH LANGUAGE PAIR

Benyamin Ahmadnia and Javier Serrano

Autonomous University of Barcelona, Cerdanyola del Valles, Spain

ABSTRACT

The quality of Neural Machine Translation (NMT) systems like Statistical Machine Translation (SMT) systems, heavily depends on the size of training data set, while for some pairs of languages, high-quality parallel data are poor resources. In order to respond to this low-resourced training data bottleneck reality, we employ the pivoting approach in both neural MT and statistical MT frameworks. During our experiments on the Persian-Spanish, taken as an under-resourced translation task, we discovered that, the aforementioned method, in both frameworks, significantly improves the translation quality in comparison to the standard direct translation approach.

KEYWORDS

Statistical Machine Translation, Neural Machine Translation, Pivot Language Technique

1. INTRODUCTION

The purpose of the statistical machine translation is to translate a source language sequences into a target language ones by assessing the plausibility of the source and the target sequences in relation to existing bodies of translation between the two languages. A huge shortcoming in *SMT* is the lack of consistent parallel data for many language pairs and corpora of this type [2]. In order to overcome this shortcoming, researchers have developed different ways to connect source and target languages with only a small parallel corpus, that is used to generate a systematic *SMT* when a proper bilingual corpus is lacking or the existing ones are weak [5, 10, 13, 28, 29]. This is an important issue when there are languages with inefficient *NLP* (Natural Language Processing) resources that are not able to provide an *SMT* system. Nevertheless, there are sufficient resources between them and some other languages.

Afterwards, the goal of neural machine translation is to build a single neural network that can be jointly tuned to maximize the translation quality [26]. The *NMT* has built state-of-the-art for many pairs of languages only by using parallel training data set, and has shown competitive results in recent researches [3, 20, 26]. In comparison with conventional *SMT* [22], competitive translation quality has been obtained on well-resourced pairs of languages such as *English-French* or *German-English*.

In spite of these achievements, there are also some shortcomings. The *NMT* systems indicate poorer performance in comparison to a standard tree-to-string *SMT* system for under-resourced pairs of languages, because the neural network is a data-driven approach [31]. The *NMT* is non-trivial because it directly maximizes the probability of the target sentences given the source

DOI: 10.5121/ijnlc.2017.6503

sentences without modeling latent structures. In order to bridge the source-target translation model through the source-pivot and the pivot-target translation models, we need to use a joint training for the *NMT*. We have to make the translation path from the source language to the target one, with the bridge language translations. We investigate a kind of connection terms, which uses a small source-target parallel corpus to guide the translation path with the bridge language translations, so that we can connect these two directional models.

The remainder of this article is organized as follows; We introduce the structures of both translation frameworks in Section 2, and the concepts of the pivoting method in Section 3. In Section 4 we describe and analyze the experiments. Section 5 describes the related works, and Section 6 gives a conclusion of the article.

2. TRANSLATION FRAMEWORKS

In this section we will introduce the architectures of the statistical machine translation systems, and the neural machine translation systems, which are used to deal with our experiments and the translation process.

2.1. Statistical MT Framework

The statistical machine translation paradigm has, as its most important elements, the idea; that probabilities of the source and the target sentences can find the best translations. Frequently used paradigms of *SMT* on the log-linear model are the *phrase-based*, the *hierarchical phrase-based*, and the *ngram-based*. In our experiments we use the phrase-based *SMT* system with the maximum entropy framework [4]:

$$\hat{t}_1^I = \arg\max_{t_1^I} P(s \mid t) \qquad (1)$$

The phrase-based *SMT* model is an example of the noisy-channel approach, where we can present the translation hypothesis (*t*) as the target sentence (given (*s*) as a source sentence), maximizing a log-linear combination of feature functions:

$$\hat{t}_1^I = \arg\max_{t_1^I} \left\{ \sum_{m=1}^{M} \lambda_m h_m(s_1^J, t_1^J) \right\} \qquad (2)$$

This equation called the log-linear model, where λ_m corresponds to the weighting coefficients of the log-linear combination, and the feature functions $h_m(s,t)$ to a logarithmic scaling of the probabilities of each model. The translation process involves segmenting the source sentences into source phrases, translating each source phrase into a target phrase, and reordering these target phrases to yield the target sentence.

2.2. Neural MT framework

Neural machine translation aims at designing a comprehensible trainable model. In this model, all components are tuned based on a training corpora to raise the translation accuracy and performance. Building and training a single, large neural network that reads a sentence and outputs a correct translation are the chief purposes of *NMT*. Any neural network which maps a source sentence to a target one is considered as an *NMT* system, where all sentences are assumed to terminate with a special *"end-of-sentence" (<eos>)* token. More concretely, an *NMT* system uses a neural network to parameterize the following conditional distributions for $1 \leq j \leq m$:

$$P(t_j \mid t_{<j}, s_{\leq n}) \qquad (3)$$

By doing so, it becomes possible to compute and therefore maximize the log probability of the target sentence given the source sentence:

$$\log P(t \mid s) = \sum_{j=1}^{m} \log P(t_j \mid t_{<j}, s_{\leq n}) \qquad (4)$$

There are many ways to parameterize these conditional distributions. For example, Kalchbrenner and Blunsom (2013) used a combination of a convolutional neural network and a Recurrent Neural Network (*RNN*) [20], Sutskever et al. (2014) used a deep Long/Short-Term Memory (*LSTM*) model [26], Cho et al. (2014) used an architecture similar to the *LSTM* [8], and Bahdanau et al. (2015) used a more elaborate neural network architecture that uses an attentional mechanism over the input sequence [3, 15].

3. Pivot Language Technique

Translation systems in terms of both *SMT* and *NMT*, have made great strides in translation quality. State-of-the-art have shown that, high-quality translation output is dependent on the availability of massive amounts of parallel texts in the source and the target languages. However, there are a large number of languages that are considered low-density, either because the population speaking those languages is not very large, or even if millions of people speak those languages, insufficient amounts of parallel texts are available in those languages.

This technique is an idea to generate a systematic machine translation when a proper bilingual corpus is lacking or the existing ones are weak. This article shows that, how such corpora can be used to achieve high translation quality through the pivot language technique, and we investigate the performance of this strategy through our considered translation frameworks.

3.1. Pivoting Strategy for SMT

According to [29], pivot-based strategies that employed for *SMT* systems can be classified into these categories:

1. The *"transfer method"* also known as cascade or sentence translation pivot strategy, which translates the text in the source language to the pivot, using a source-pivot translation model, and then to the target language using a pivot-target translation model.

2. The *"multiplication method"* also identified as triangulation or phrase translation pivot strategy, which merges the corresponding translation probabilities of the translation models for the source-pivot and the pivot-target languages, generates a new source-target translation model.

3. The *"synthetic corpus method"* which tries to create a synthetic source-target corpus by translating the pivot part in the source-pivot corpus, into the target language with a pivot-target model, and translating the pivot part in the target-pivot corpus, into the source language with a

pivot-source model. Finally combining the source sentences with the translated target sentences or combining the target sentences with the translated source sentences. Nevertheless, it is somehow difficult to build a high-quality translation system with a corpus created only by a machine translation system.

In this article our *SMT* pivoting experiments just rely on the first and the second methods.

3.1.1. Transfer Method

In the sentence translation pivot strategy, we first translate the *Persian* sentences into the *English* ones, and then translate these *English* sentences into the *Spanish* ones separately. We select the highest scoring sentence from the *Spanish* sentences.

$$\arg\max_{p} P(s|p) = \arg\max_{p} \sum_{e} P(s,e|p) = \arg\max_{p} \sum_{e} P(s|e,p)P(e|p) \qquad (5)$$

In this technique for assigning the best *Spanish* candidate sentence (*s*) to input the *Persian* sentence (*p*), we maximize the probability *P(s|p)* by defining hidden variable (*e*), which stands for the pivot language sentences, we gain:

$$\arg\max_{p} P(s|p) \approx \arg\max_{p} \sum_{e} P(s|e)P(e|p) \qquad (6)$$

In Equation (6), summation on all (*e*) sentences is difficult, so we replace it by maximization, and Equation (7) is an estimate of Equation (6):

$$\arg\max_{p} P(s|p) \approx \arg\max_{p} \max_{e} P(s|e)P(e|p) \qquad (7)$$

Instead of searching all the space of (*e*) sentences, we can just search a subspace of it. For simplicity we limit the search space in Equation (8). A good choice is (*e*) subspace produced by the (*n-best*) list output of the first *SMT* system (source-pivot):

$$\arg\max_{p} P(s|p) \approx \arg\max_{p} \max_{e \in n-best(s)} P(s|e)P(e|p) \qquad (8)$$

In fact each sentence (*p*) of the *Persian* test set is mapped to a subspace of total (*e*) space and search is done in this subspace for the best candidate sentence (*s*) of the second *SMT* system (pivot-target).

3.1.2. Multiplication Method

For applying the phrase translation pivot strategy, we directly construct the *Persian-Spanish* phrase translation table from the *Persian-English*, and the *English-Spanish* phrase-tables.

In this technique phrase (*p*) in the source-pivot phrase-table is connected to (*e*), and phrase (*e*) is associated with (*s*) in the pivot-target phrase-table. We link (*p*) and (*s*) in the new phrase-table for the source-target. For scoring the pair phrases of the new phrase-table, assuming *P(e|p)* as the score of the *Persian-English* phrases and *P(s|e)* as the score of the *English-Spanish* phrases, then the score of the new pair phrases (*p*) and (*s*), $P(s|p)$, in the *Persian-Spanish* phrase-table is:

$$P(s|p) = \sum_{e} P(s, e|p) \qquad (9)$$

(*e*) is a hidden variable and actually stands for the phrases of pivot language:

$$P(s|p) = \sum_{e} P(s|e, p) P(e|p) \qquad (10)$$

If we assume that, (*p*) and (*s*), are independent given (*e*):

$$P(s|p) \approx \sum_{e} P(s|e) P(e|p) \qquad (11)$$

For simplicity the summation on all the (*e*) phrases is replaced by maximization, then Equation (11) is approximated by:

$$P(s|p) \approx \max_{e} P(s|e) P(e|p) \qquad (12)$$

3.2. Pivoting Strategy for NMT

Considering $P(p|s;\ \theta_{sp})$ and $P(t|p;\ \theta_{pt})$ as the source-pivot and the pivot-target *NMT* models respectively, while giving two parallel corpora, the source-pivot parallel corpus (C_{sp}) and the pivot-target parallel corpus (C_{pt}). We employ the pivoing strategy in which the target sentence is generated for a source sentence after it is first translated to the pivot sentences. The crucial point is to jointly instruct two translation models, $P(p|s;\ \theta_{sp})$ and $P(t|p;\ \theta_{pt})$, heading at establishing the source-target translation path with the pivot sentences as the intermediate translations:

$$J(\theta_{sp}, \theta_{pt}) = \sum_{n=1}^{N_{sp}} \log P(p^{(n)} \mid s^{(n)}; \theta_{sp}) + \sum_{n=1}^{Npt} \log P(t^{(n)} \mid p^{(n)}; \theta_{pt}) + \lambda\, R(\theta_{sp}, \theta_{pt}) \qquad (13)$$

The source-pivot Likelihood, the pivot-target Likelihood, and the linking term, are the main objectives of our training model. In order to balance the significance between the Likelihoods and the linking term, (λ) is utilised. The linking term includes two sets of parameters; (θ_{sp}) and (θ_{pt}), for the source-pivot and the pivot-target translation models respectively. The linking term is controlled so as to allow two independently trained parameters from two different translation models to interact mutually. Replacing the linking term by any function with the parameters of these two included directional *NMT* models is feasible.

In general, for many language pairs and domains, small corpora are pervasive. Given a test source sentence, it will be translated to the target sentence eventually through the pivoting technique. This translation path will be reinforced with the supply for parallel sentence pairs between the source and the target. The employed approach in the current study treats the pivot sentences as latent variables:

$$R(\theta_{sp}, \theta_{pt}) = \sum_{n=1}^{N_{st}} P(t^{(n)} \mid s^{(n)}; \theta_{sp}, \theta_{pt}) = \sum_{n=1}^{N_{st}} \sum_{p} P(t^{(n)}, p \mid s^{(n)}; \theta_{sp}, \theta_{pt})$$

$$= \sum_{n=1}^{N_{st}} \sum_{p} P(p \mid s^{(n)}; \theta_{ps}) P(t^{(n)} \mid p; \theta_{pt}) \qquad (14)$$

Where (p) is a latent pivot sentence. The intuition of Equation (14) is to maximize the translation probability of the target sentences given the source sentences via the pivot candidate translations. The source-pivot translation model first transforms the source sentences into the latent pivot sentences, from which, the pivot-target translation model aims to construct the target sentences. This training criterion conforms to the pivot translation strategy adopted by the test procedure [6].

The partial derivative of $J(\theta_{sp}, \theta_{pt})$ with respect to the parameters (θ_{sp}) of the source-pivot model is calculated as:

$$\frac{\partial J(\theta_{sp}, \theta_{pt})}{\partial \theta_{sp}} = \sum_{n=1}^{N_{sp}} \frac{\partial \log P(p^{(n)} \mid s^{(n)}; \theta_{sp})}{\partial \theta_{sp}} + \lambda \frac{R(\theta_{sp}, \theta_{pt})}{\theta_{sp}} \qquad (15)$$

The partial derivative with respect to the parameters (θ_{pt}) is similar to Equation (15). In our connection term, if we continue to expand the last term of Equation (15), a challenge emerges:

$$\frac{\sum_{p \in P_{(s)}} P(p \mid s; \theta_{sp}) P(t \mid p; \theta_{pt}) \frac{\partial \log P(p|s;\theta_{sp})}{\partial \theta_{sp}}}{\sum_{p \in P_{(s)}} P(p \mid s; \theta_{sp}) P(t \mid p; \theta_{pt})} \qquad (16)$$

Enumerating all of pivot candidate translations $p \in P(s)$ in Equation (16) is intractable because of the exponential search space for the pivot translations. As an alternative solution, the subset approximation is normally employed. In order to approximate the full space, we utilized a subset $P'(s) \subset P(s)$. In addition, we undertook two methods to generate $P'(s)$, sampling (k) translations from the full space and generating (*k-best*) list of candidate translations. The findings revealed that generating (*k-best*) list operates better.

Holding three parallel corpora including the source-pivot, the pivot-target, and the source-target, we still utilize mini-batch *stochastic gradient descent* algorithm in order to update the parameters. Though three mini-batches of parallel sentence pairs are randomly picked in each iteration from the source-pivot, the pivot-target and the source-target parallel corpora. Likelihood, in order to get the (*k-best*) pivot translations, decoding the source sentences of the source-target mini-batch is needed. Afterwards, the gradients for these batches are calculated and then collected for parameter updating purposes. The decision rules for the source-pivot and the pivot-target *NMT* models are respectively given by:

$$\theta_{sp}^* = \arg\max \left\{ \sum_{n=1}^{N_{(sp)}} \log P(p^{(n)} \mid s^{(n)}; \theta_{sp}) + \lambda R(\theta_{sp}, \theta_{pt}) \right\} \qquad (17)$$

$$\theta_{pt}{}^{*} = \arg\max \left\{ \sum_{n=1}^{N_{(pt)}} \log P(t^{(n)} \mid p^{(n)}; \theta_{pt}) + \lambda R(\theta_{sp}, \theta_{pt}) \right\} \quad (18)$$

4. EXPERIMENTAL FRAMEWORK

In this section, we present a set of experiments on both *SMT* and *NMT* frameworks including the pivot language technique to overcome the limitation of training resources scarsity. Then we present our results to compare the *Persian-Spanish* translation quality in either both aforementioned frameworks.

Our data resources in both *SMT* and *NMT* experiments are collected from the in-domain *Tanzil* parallel corpus [27]. In this corpus the *Persian-Spanish* part contains more than (*68K*) sentences and approximately (*3.51M*) words, the *Persian-English* part contains more than (*1M*) sentences and more than (*57M*) words, and the *English-Spanish* part contains more than (*133K*) sentences and approximately (*4.25M*) words. Table 1 shows our data resource statistics.

Table 1. Corpus Statistics.

Corpus	**Direction**	**Sentences**
Tanzil	Persian - English	1,028,996
Tanzil	English - Spanish	133,735
Tanzil	Persian - Spanish	68,601

The training part of our system involved of (*60K*) sentences. For the tuning and the testing steps we collected parallel texts from the *Tanzil* corpus, we extracted (*3K*) sentences for the tuning, and (*5K*) sentences for the testing.

4.1. SMT Systems Experiments and Results

"MOSES" package [21], is used for training our *SMT* systems. Through utilising *MOSES* decoder, we apply *fast-align* approach [12], for sentence alignment in our experiment. The employed language model for all *SMT* systems are *3-grams* and they are built using the *KenLM* toolkit [19]. We use the *BLEU* metric [24], in order to evaluate the systems performance. Table 2 presents the results of the *Persian-English*, the *English-Spanish*, and the *Persian-Spanish* direct translation systems.

Table 2. The BLEU scores of the Pe-En, the En-Es, and the Pe-Es direct SMT systems.

System	**Persian - English**	**English - Spanish**	**Persian - Spanish**
Direct	14.31	15.34	11.39

In the other portion of this experiment the two phrase-tables employed to shape a new table in the phrase pivoting method are extracted in turn from the *Persian-English* and the *English-Spanish* translation systems. Table 3 illustrates the results of the sentence translation pivoting and the phrase translation pivoting of the *Persian-Spanish* translation system through *English* as the intermediary language.

Table 3. The BLEU scores of the Pe-(En)-Es pivoting SMT systems.

System	**Phrase - Level**	**Sentence - Level**
Pivoting	13.55	12.78

According to the results, in the case of *Persian-Spanish* language pair, the pivot-based translation method is suitable for the scenario that there exist large amounts of source-pivot and pivot-target bilingual corpora and only a little source-target bilingual data. Thus we selected (*60K*) sentence pairs from the source-target bilingual corpora to simulate the lack of source-target data.

4.2. NMT Systems Experiments and Results

"MANTIS" package [9], is used as the attention-based *NMT* systems in our experiments. We have tried to analyze the *Persian-Spanish* language pair through *English* as the bridge language. For this language pair, we removed the empty lines and retain sentence pairs with no more than (50) words. In order to avoid the constitution of the tri-lingual corpora by the source-pivot and the pivot-target, the overlapping section of the pivot sentences from the source-pivot and the pivot-target corpora should be divided into two equal parts and also they should be combined separately with the non-overlapping parts. For the language modeling we used the *RNN* language model [23], separately. In order to use the Likelihood linking term, we set the sample size, (k), to (*40*), in order to avoid the weird segmentation fault error message. The hyper-parameter, (λ), to (*1.0*), and the threshold of gradient clipping to (*0.1*). The parameters for the source-pivot and the pivot-target translation models in Likelihood are initialized by pre-trained model parameters.

All the sentences of the corpora are encrypted by the *tokenize.perl* script and the development and the test data sets are from the *Tanzil* corpus as well as the training data set. The evaluation metric is *BLEU* [24], as calculated by the *multi-bleu.perl* script. We have used *English* as the pivot language and followed Likelihood linking term that jointly train the source-pivot and the pivot-target translation models. We have tried to show a comparison between translation quality for the source-pivot, the pivot-target, and the source-target directions. The source-target translation results are obtained by translating pivot sentences. Table 4 shows a comparison results on the *Persian-Spanish* translation task from the *Tanzil* corpus.

Table 4. The BLEU scores comparing the direct with the Likelihood NMT system.

System	**Persian - English**	**English - Spanish**	**Persian - Spanish**
Direct	14.17	14.88	11.19
Likelihood	14.31	15.02	12.93

The results show that, the *BLEU* scores of the Likelihood method are better than the standard direct training. Our analysis points out that, the Likelihood strategy improves the translation performance on the *Persian-Spanis*h translation task up to (1.74) *BLEU* scores (in comparison with the direct translation approach), by introducing the source-target parallel corpus to maximize $P(t|s;\ \theta_{sp},\ \theta_{pt})$ with (p) as the latent variables makes the source-pivot and the pivot-target translation models improved collaboratively. As we have showed, this approach improves translation quality of both pivot and target sentences.

5. Related Works

In the case of low-resourced language pairs, some researchers introduce a pivot language to bridge source and target languages in *SMT*, such as the case of *Catalan-English* with no parallel corpus [11]. Some researchers investigated the *SMT* system with pivot language method. One example is Hartley et al. (2007) who used the *Russian* language as a pivot for translating from *Ukrainian* to *English*. From their experience, we figured out that, it is possible to achieve better translation quality with pivot language approach [18]. Habash and Hu (2007) compared two approaches for *Arabic-Chinese* language pair with direct *MT* system through *English* as a pivot language. Their researches indicate that using *English* as a pivot language in either approaches leads to better results than the direct translation from *Arabic* to *Chinese* [17]. Going in the same direction, Al-Hunayti et al. (2010) presented a comparison between two common pivot strategies; phrase translation and sentence translation, in order to enhance *Danish-Arabic SMT* system. This approach showed that the sentence pivoting overtakes the phrase pivoting when common parallel corpora are not available [1].

Firat et al. (2016) proposes a multiway, multilingual *NMT* model that enables zero-resourced *MT*. In order to find tune parameters of the low-resourced language pairs using trained parameters on the high-resourced language pairs [14]. Zoph et al. (2016) adopted a transfer learning method. The aim was to build a source-target *NMT* model. Because of limited quantity, quality, and coverage for parallel corpora, additional data resource have come under scrutiny lately [31]. For example, Zhang and Zong (2016) proposed two approaches to incorporate the source side monolingual corpora; One is to employ self-training algorithm to generate parallel corpora from monolingual corpora. The other adopts multi-task learning framework to enhance the encoder network of *NMT* [30]. On the other hand, Cheng et al. (2016) introduced an auto-encoder framework to reconstruct monolingual sentences using the source-target and the target-source *NMT* models [7]. Researchers such as Gulccehre et al. (2015) proposed to incorporate the target side monolingual corpora as the language model for *NMT* [16]. As Sennrich et al. (2016) pairs the target monolingual corpora with its corresponding translations then merges them with parallel data for retraining the source-target model [25].

6. Conclusion

In this article, we have tried to analyze the behavior of the pivot (bridge) language technique on both statistical and neural machine translation systems for the *Persian-Spanish*, which is a resource poor language pair.

In the first case, we have compared two common pivoting translation methods comprising the phrase-level combination, and the sentence-level combination, for the *Persian-Spanish SMT* by employing *English* as an intermediary language. By organizing controlled experiments, we have assessed the performances of these two methods against the performance of directly trained *SMT* system. The results revealed that utilizing *English* as a bridging language in either approaches gives better results than by the direct translation approach from *Persian* to *Spanish*.

In the second case, we have presented a joint training method for the *Persian-Spanish NMT* via *English* as a bridge language. The connection term in our joint training objective makes the *Persian-English* and the *English-Spanish* translation models interact better. So that the experiments confirm that, this approach achieves significant improvements.

Acknowledgements

We would like to express our sincere gratitude to Shekoofeh Dadgostar for all her support. We have benefited from her erudition and thoughtful comments which truly enriched this article.

REFERENCES

[1] Al-Hunaity, Mossab & Maegaard, Bente & Hansen, Dorte, (2010) “Using English as a Pivot Language to Enhance Danish-Arabic Statistical Machine Translation”.

[2] Babych, Bogdan & Hartley, Anthony & Sharoff, Serge & Mudraya, Olga, (2007) “Assisting Translators in Indirect LexicalTransfer”, ACL.

[3] Bahdanau, Dzmitry & Cho, Kyunghyun & Bengio, Yoshua, (2015) “Neural machine translation by jointly learning to align and translate”, ICLR.

[4] Berger, Adam.L & Della Pietra, Stephen & Della Pietra, Vincent.J, (1996) “A Maximum Entropy Approach to Natural Language Processing”, Journal of Computational Linguistics.

[5] Bertoldi, Nicola & Barbaiani, Madalina & Federico, Marcello Cattoni, Roldano, (2008) “Phrase-based statistical machine translationwith pivot languages”, IWSLT.

[6] Cheng, Y.ong & Liu, Yang & Yang, Qian & Sun, Maosong & Xu, Wei, (2016) “Neural Machine Translation with Pivot Languages”, ACL.

[7] Cheng, Yong & Xu, Wei & He, Zhongjun & He, Wei & Wu, Hua & Sun, Maosong & Liu, Yang, (2016) “Semi-Supervised Learning for Neural Machine Translation”, ACL.

[8] Cho, Kyunghyun & Merrienboer, Bert.V & Gulccehre, Caglar & Bougares, Fethi & Schwenk, Holger & Bengio, Yoshua, (2014) “Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation”.

[9] Cohn, Trevor & Hoang, Cong.D.V & Vymolova, Ekaterina & Yao, Kaisheng & Dyer, Chris & Haffari, Gholamreza, (2016) “Incorporating Structural Alignment Biases into an Attentional Neural Translation Model”, NAACL-HLT.

[10] Cohn, Terevor & Lapata, Mirella, (2007) “Machine Translation by Triangulation: Making Effective Use of Multi-Parallel Corpora”.

[11] DeGispert, Adria & Marino, Jose.B, (2006) “Catalan-English statistical machine translation without parallel corpus: bridging through Spanish”.

[12] Dyer, Chris & Chahuneau, Victor & Smith, Noah.A, (2013) “simple, Fast, and Effective Reparameterization of IBM Model 2”.

[13] El-Kholey, Ahmed & Habash, Nizar, (2013) “Language Independent Connectivity Strength Features for Phrase Pivot Statistical Machine Translation”.

[14] Firat, Orhan & Sankaran, Baskaran & Al-Onaizan, Yaser & Yarman Vural, Fatos.T & Cho, Kyughyun, (2016) “Zero-Resource Translation with Multi-Lingual Neural Machine Translation”.

[15] Graves, Alex & Wayne, Greg & Danihelka, Ivo, (2014) “ Neural Turing Machines”.

[16] Gulccehre, Caglar & Firat, Orhan & Xu, Kelvin & Cho, Kyunghyun & Barrault, Loic & Lin, Huei-Chi & Bougares, Fethi & Schwenk, Holger & Bengio, Yoshua, (2015) “On Using Monolingual Corpora in Neural Machine Translation”.

[17] Habash, Nizar & Hu, Jun, (2007) “Improving Arabic-Chinese Statistical Machine Translation using English as Pivot Language”.

[18] Hartley, Anthony & Babych, Bogdan & Sharoff, Serge, (2007) “Translating from under-resourced languages Comparing direct transfer against pivot translation”.

[19] Heafield, Kenneth & Pouzyrevsky, Ivan & Clark, Jonathan.H & Koehn, Philipp, (2013) “Scalable Modified Kneser-Ney Language Model Estimation”, ACL.

[20] Kalchbrenner, Nal & Blunsom, Phil, (2013) “Recurrent Continuous Translation Models”.

[21] Koehn, Philipp & Hoang, Hieu & Birch, Alexandra & Callison-Burch, Chris & Federico, Marcello & Bertoldi, Nicola & Cowan, Brooke & Shen, Wade & Moran, Christine & Zens, Richard & Dyer,

Chris & Bojar, Ondrej & Constantin, Alexandra & Herbst, Evan, (2007) “Moses: Open Source Toolkit for Statistical Machine Translation”, ACL.

[22] Koehn, Philipp & Och, Franz.J & Marcu, Daniel, (2003) “Statistical Phrase-Based Translation”, NAACL-HLT.

[23] Mikolov, Tomas & Karafiat, Martin & Burget, Lukas & Cernocky, Jan & Khudanpur, Sanjeev, (2010) “Recurrent neural network based language model”, Interspeech.

[24] Papineni, Kishore & Roukos, Salim & Ward, Todd Zhu, Wei.J, (2002) “BLEU: a Method for Automatic Evaluation of Machine Translation”.

[25] Sennrich, Rico & Haddow, Barry & Brich, Alexandra, (2016) “Improving Neural Machine Translation Models with Monolingual Data”.

[26] Sutskever, Ilya & Vinyals, Oriol & Le, Quoc.V, (2014) “Sequence to Sequence Learning with Neural Networks”.

[27] Tiedemann, Jorg, (2012) “Parallel Data, Tools and Interfaces in OPUS”, LREC.

[28] Utiyama, Masao & Isahara, Hitoshi, (2007) “A comparison of pivot methods for phrase-based statistical machine translation”, NAACL-HLT.

[29] Wu, Hua & Wang, Haifeng, (2007) “Pivot language approach for phrase-based statistical machine translation”, ACL.

[30] Zhang, Jiajun & Zong, Chengqing, (2016) “Exploiting source-side monolingual datain neural machine translation”, EMNLP.

[31] Zoph, Barret & Yuret, Deniz & May, Jonathan & Knight, Kevin, (2016) “Transfer learning for low-resource neural machine translation”, EMNLP.