Machine Translation for the Portuguese language Using Words Distributed Representations

Fillipe Madureira
Programa de Pós-Graduação em Engenharia Elétrica (PROEE)
Universidade Federal de Sergipe (UFS). Brazil
fillipelgm@proee.ufs.br

Hendrik Macedo
Programa de Pós-Graduação em Ciência da Computação (PROCC)
Universidade Federal de Sergipe (UFS). Brazil
hendrik@dcomp.ufs.br

ABSTRACT: *Machine Translation is a highly demanding research field in natural language processing. The goal is to investigate a model which can automatically translate a document from one language to another. The mainstream approach for that is the Phrase-based Machine Translation (PBMT). In [9], a innovative method based on distributed representations was proposed to automate the process of generating dictionaries and phrase tables of PBMT-based solutions, that primarily rely on raw counts. In this paper, as in the Mikolov's, we have used distributed representations of words to perform automatic machine translation between languages but, unlike it, we have included the Portuguese language in the analysis which, actually, is the primary goal of our work. We have properly evaluated the performance in regards to Precision metrics. Results shows a precision of up to 89% if we consider translation between the Portuguese and the English.*

Keywords: Machine Translation, Word Embeddings, CBOW, Word2vec

Received: 18 June 2017, Revised 20 July 2017, Accepted 30 July 2017

1. Introduction

The goal of Machine Translation is to automatically translate a document from one language to another [1, 3, 2, 13, 12]. This task is incredibly difficult due to idiosyncrasies in each language, such as morphological and syntactic structures. Stylistic and cultural differences also impose difficulties. Such distinct characteristics between source and target languages are known as translation divergences.

A possible strategy to overcome the aforementioned difficulties is to use a neural network language model. These models have

become very popular in recent years due to their inherent capability to embed many syntactic and semantic regularities and patterns. In these models, words are represented as high dimensional real valued vectors. Mikolov et al. [11] explored some linguistic regularities by observing vectors offsets between words that shared some kind of relationship. Distributed representations of words in a vector space help learning algorithms to achieve better performance in natural language processing tasks by grouping similar words [10, 5, 7, 6].

In [9], Mikolov et al. proposed a method based on distributed representations that can automate the process of generat-ing dictionaries and phrase tables used by mainstream techniques of statistical machine translation that primarily rely on raw counts. The method is based on building monolingual language models of both the source and target language. These models are learned using the Skip-gram or Continuous Bag-of-Words (CBOW) proposed in [8]. The linguistic regularities and patterns which occur within the source language can be mapped to the vector space of the target lan-guage through a linear transformation. Thus, a simple vector multiplication can be used to estimate the translation of a word between two languages.

Mikolov et al. made an experiment with the publicly available corpora from WMT11[1]. They built monolingual data sets for English (EN), Spanish (SP) and Czech (CZ) languages and performed the following translations: EN!SP, SP!EN, EN!CZ and CZ!EN. They also made a large scale experiment using English and Spanish corpora with billions of words (Google News data sets), which were not made publicly available.

The goal of this paper is to take advantage of the distributed representations of words to perform machine translation for the Portuguese language and properly evaluate the performance.

2. Method

In this section we present the main aspects of our working method.

2.1 Selection of Corpora

We have chosen the Europarl parallel corpus[2] [4] for two reasons: (1) it contains texts in Portuguese language, which is obviously a primary requirement and, in addition, (2) it provides texts in both English and Spanish, which is important in order to compare the results of Mikolov's work. The Europarl parallel corpus is extracted from the proceedings of the European Parliament.

2.2 Preprocessing of Raw Texts

Each of the selected monolingual corpus was preprocessed using a Python script (instead of the preprocessing tools available at http://www.statmt.org/europarl/v7/tools.tgz). The preprocessing steps were the fol-lowing:

• Removal of punctuation and UTF-8 characters that do not belong to the set of alphabetic characters of each respective language

• Tokenization of text using Natural Language Toolkit (NLTK[3])

• Removal of any token containing numeric characters

• Lowercasing the text to discard named entities

2.3 Formation of Short Phrases of Words

We have formed short phrases of words just as Mikolov et al. did in [10, 9]. This allows us to represent common bigrams, trigrams or even greater n-grams as a single token. However, this is done only if the probability of the words' co-occurrence (above a predefined threshold) is greater than their isolated unigram probability. Such approach is robust enough to deal with the presence of stop words.

In order to do so, we used the word2phrase tool, which is part of the word2vec[4] tool. We run two iterations of word2phrase with thresholds of 200 and 100, respectively.

[1]http://www.statmt.org/wmt11/training-monolingual.tgz

2.4 Construction of the Language Models

The language models were constructed with the word2vec tool. We have used the CBOW architecture because it is much faster to train than Skip-bow. We trained row vectors of size in the range of 200 to 800 (in increments of 100). The window size was 10 and we used negative sampling as the training algorithm with 25 negative examples per each positive one. All of these values have been defined empirically.

2.5 Creation of Dictionaries between Languages and Optimization of the Translation Matrix

After the training of the language models, we took the 5000 most frequent words from a given source language and used Google Translate (GT) to find its translation to the target language (the word2vec tool organizes words in its output file according to their frequency). Then, we took those translations and looked for them in the target language model. For each word in the source language whose translation (given by GT) could be found in the target language we formed a pair.

As an example, suppose we want to translate the word "house" from English to Portuguese. GT returns the word "casa". If that word exists in the target language model vocabulary, we pair them. Ideally, we would have 5K words for training, but sometimes the words (or word phrases) produced by GT were not present in the target language vocabulary, which led us to discard such pairs. The vocabulary coverage was reported for each experiment. Paired words from both source and target language were used to learn the Translation Matrix.

Given a set of word pairs and their associated vector representations $\{x_i, z_i\}_{i=1}^{n}$, where $x_i \in \Re^{d1}$ is the distributed representation of the i-th word in the source language, and $z_i \in \Re^{d2}$ is the vector representation of its translation in the target language. The goal is to find the transformation matrix W such that Wx_i approximates z_i. The objective function is given by

$$\min_{W} \sum_{i=1}^{n} \|Wx_i - z_i\|^2 \tag{1}$$

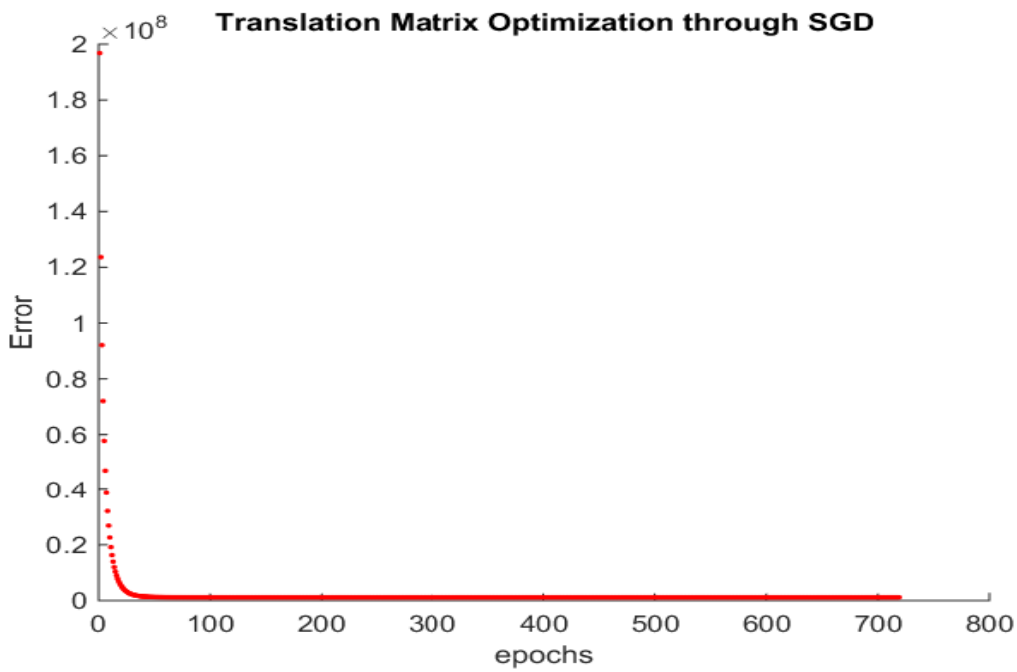

Figure 1. Training of the translation matrix for xi and zi of dimensionality 300

[2]http://www.statmt.org/europarl/
[3]http://www.nltk.org/
[4]Available at: https://code.google.com/archive/p/word2vec/source/default/source

which Mikolov et al. solved with stochastic gradient descent (SGD). Optimizing the matrixW through SGD, however, is time consuming, taking over 700 epochs to minimize the cost function, as can be seen in Figure 1.

Since the transformation is linear and the objective function is quadratic, Equation 1 can be solved instantaneously with the Moore-Penrose pseudoinverse as in

$$X^{\dagger} = (X^T X)^{-1} X^T \tag{2}$$

which leads to the optimization of the Translation Matrix as in

$$W = X^{\dagger} Z \tag{3}$$

where X and Z are the respective collection of vectors from the source and target language.

2.6 Creating the Test Set and Predicting Translations

The test set was created following the same procedure used for the training set but, in this case, we used the subsequent 1000 words from the source language instead. We reported the test vocabulary coverage in each experiment as well as we did with the training set.

At the prediction time, for any given test word and its continuous vector representation x, we mapped it to the target language space by computing $z = Wx$. Then we find the word whose representation is closest to z in the target language space using cosine similarity as the distance metric.

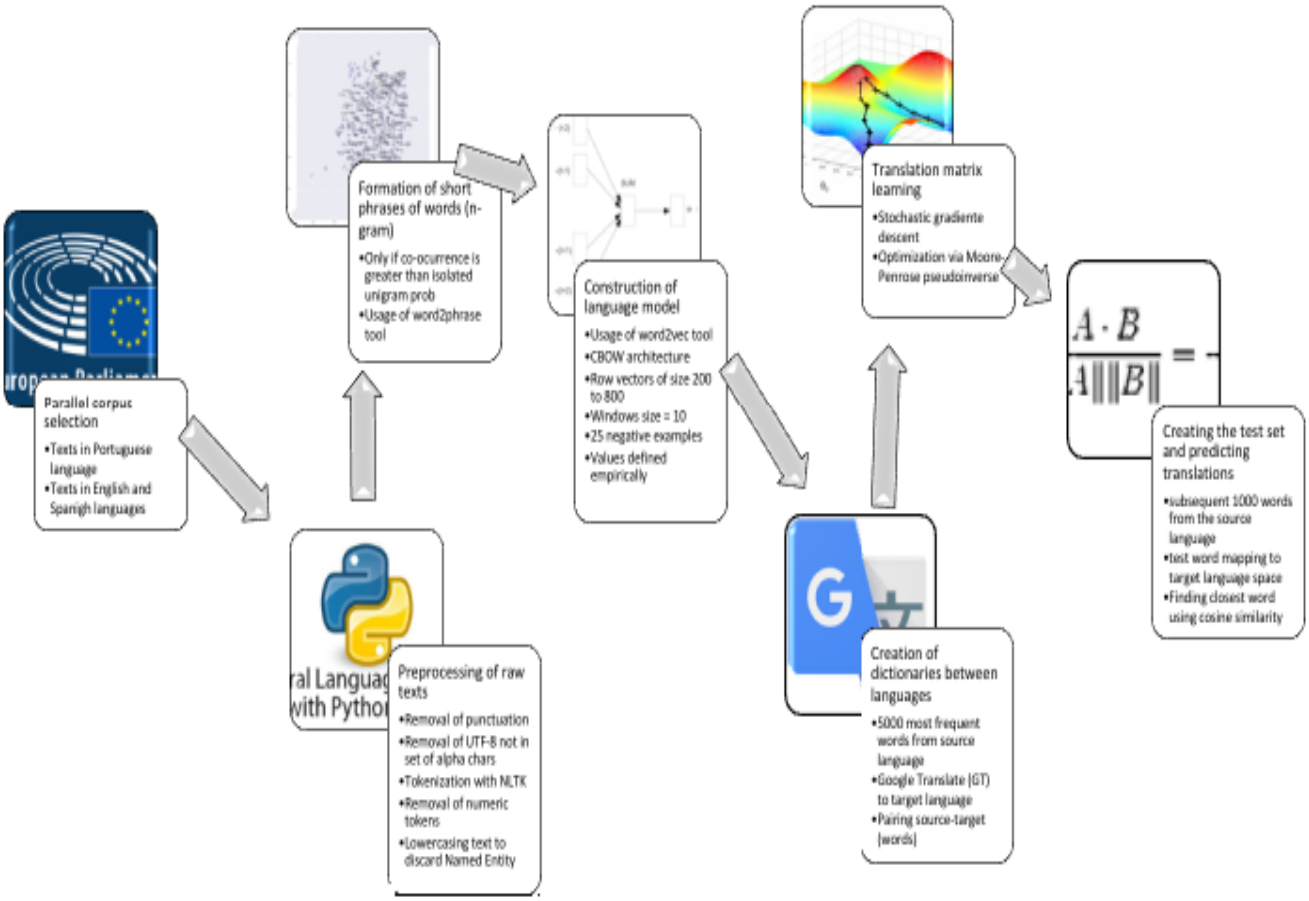

Figure 2. Architecture of the proposed method.

3. Results and Discussion

We gathered some data about the Europarl corpora used in this work. They are shown in Table 1.

Language	Training Tokens	Vocabulary Size
English	45,601,784	72,660
Spanish	46,911,534	108,643
Portuguese	45,509,054	105,253

Table 1. Information about the corpora texts

We varied the vectors size for both the source and target languages and calculated Precision@1 and Precision@5. For each test word, we computed 5 translation candidates, based on cosine similarity score. Precision@1 only counts a successful translation whether the translation candidate whose score is the highest matches the dictionary entry given by GT, whilst at Precision@5 we count a successful translation whether any of the candidates matches the dictionary entry.

Information about the translation and test sets coverage are available in Table 2. This information can help when evaluating the results, as it is expected that the bigger the training set is, the better the performance in testing and validation.

Language Pair	Training Coverage (%)	Test Coverage (%)
EN→ES	87.8	82.1
EN→PT	61.98	76.8
ES→EN	79.24	71.9
ES→PT	87.22	84.4
PT→EN	79.58	73.8
PT→ES	90.58	87.3

Table 2. Coverage of training and testing for all translation directions performed in the experiments

ES (target) \ EN (source)	200	300	400	500	600	700	800
800	P@1 = 74.06% P@5 = 86.97%	P@1 = 75.88% P@5 = 86.24%	P@1 = 76.25% P@5 = 87.45%	P@1 = 76.13% P@5 = 87.33%	P@1 = 75.64% P@5 = 87.45%	P@1 = 76.74% P@5 = 86.97%	P@1 = 75.64% P@5 = 86.72%
700	P@1 = 73.93% P@5 = 86.60%	P@1 = 75.76% P@5 = 86.72%	P@1 = 76.25% P@5 = 86.60%	P@1 = 76.00% P@5 = 87.70%	P@1 = 75.40% P@5 = 87.45%	P@1 = 75.76% P@5 = 87.21%	P@1 = 75.40% P@5 = 86.24%
600	P@1 = 74.30% P@5 = 87.45%	P@1 = 76.25% P@5 = 87.33%	P@1 = 76.49% P@5 = 87.09%	P@1 = 75.27% P@5 = 87.82%	P@1 = 76.00% P@5 = 87.58%	P@1 = 76.15% P@5 = 86.97%	P@1 = 75.64% P@5 = 87.45%
500	P@1 = 75.27% P@5 = 86.85%	P@1 = 76.49% P@5 = 86.97%	P@1 = 76.49% P@5 = 87.21%	P@1 = 77.10% P@5 = 87.70%	P@1 = 76.49% P@5 = 87.45%	P@1 = 77.10% P@5 = 86.97%	P@1 = 75.15% P@5 = 86.85%
400	P@1 = 76.49% P@5 = 86.97%	P@1 = 76.98% P@5 = 86.72%	P@1 = 76.98% P@5 = 87.45%	P@1 = 77.71% P@5 = 87.70%	P@1 = 76.37% P@5 = 87.70%	P@1 = 76.61% P@5 = 87.21%	P@1 = 75.64% P@5 = 87.33%
300	P@1 = 76.74% P@5 = 87.09%	P@1 = 77.47% P@5 = 87.70%	P@1 = 78.56% P@5 = 87.70%	P@1 = 77.22% P@5 = 88.19%	P@1 = 77.83% P@5 = 87.58%	P@1 = 77.95% P@5 = 87.09%	P@1 = 77.10% P@5 = 87.45%
200	P@1 = 78.68% P@5 = 87.33%	P@1 = 78.44% P@5 = 86.72%	P@1 = 79.42% P@5 = 87.09%	P@1 = 78.68% P@5 = 87.21%	P@1 = 78.32% P@5 = 86.97%	P@1 = 78.81% P@5 = 86.48%	P@1 = 78.08% P@5 = 86.85%

Table 3. Accuracy at P@1 and P@5 for various vectors sizes for the language pair EN→ES

Table 3 contains the accuracy results obtained from the multiple translation from English to Spanish experiments. As Mikolov et al. observed in [9], usually, the best results occur when the word vectors trained on the source language are larger than the word vectors trained on the target language. That observation holds only if we observe each column of the table individually, though. The best results for P@1 and P@5 were highlighted.

Table 4 contains the accuracy results obtained from the multiple translation from English to Portuguese experiments. For this language pair, the results were very good, considering the training coverage. The behaviour of the results is very similar to the ones obtained for EN!SP. Once again, the best results for P@5 and P@1 were highlighted.

PT (target)							
800	P@1 = 75.39% P@5 = 87.76%	P@1 = 76.82% P@5 = 87.89%	P@1 = 75.39% P@5 = 88.41%	P@1 = 76.17% P@5 = 87.50%	P@1 = 76.95% P@5 = 87.24%	P@1 = 75.52% P@5 = 87.11%	P@1 = 74.87% P@5 = 87.11%
700	P@1 = 75.52% P@5 = 87.76%	P@1 = 75.65% P@5 = 88.41%	P@1 = 76.04% P@5 = 89.32%	P@1 = 75.13% P@5 = 87.11%	P@1 = 76.04% P@5 = 88.28%	P@1 = 74.61% P@5 = 87.63%	P@1 = 74.61% P@5 = 87.89%
600	P@1 = 75.39% P@5 = 88.15%	P@1 = 76.04% P@5 = 87.89%	P@1 = 76.95% P@5 = 88.80%	P@1 = 78.69% P@5 = 87.24%	P@1 = 77.21% P@5 = 87.76%	P@1 = 74.09% P@5 = 86.59%	P@1 = 74.22% P@5 = 87.37%
500	P@1 = 76.95% P@5 = 89.19%	P@1 = 76.95% P@5 = 88.93%	P@1 = 77.08% P@5 = 89.32%	P@1 = 77.34% P@5 = 88.28%	P@1 = 77.08% P@5 = 87.89%	P@1 = 75.26% P@5 = 87.89%	P@1 = 75.13% P@5 = 87.50%
400	P@1 = 77.34% P@5 = 88.93%	P@1 = 77.73% P@5 = 88.28%	P@1 = 77.73% P@5 = 89.32%	P@1 = 77.34% P@5 = 88.67%	P@1 = 77.99% P@5 = 88.02%	P@1 = 77.34% P@5 = 88.67%	P@1 = 75.78% P@5 = 88.15%
300	P@1 = 77.34% P@5 = 89.45%	P@1 = 78.52% P@5 = 88.93%	P@1 = 78.91% P@5 = 89.84%	P@1 = 77.47% P@5 = 87.89%	P@1 = 77.34% P@5 = 87.63%	P@1 = 76.43% P@5 = 87.63%	P@1 = 76.56% P@5 = 88.54%
200	P@1 = 79.17% P@5 = 89.45%	P@1 = 78.65% P@5 = 89.58%	P@1 = 79.04% P@5 = 89.19%	P@1 = 78.78% P@5 = 88.93%	P@1 = 78.26% P@5 = 88.54%	P@1 = 77.86% P@5 = 88.02%	P@1 = 76.95% P@5 = 88.54%
Size	200	300	400	500	600	700	800
	EN (source)						

Table 4. Accuracy at P@1 and P@5 for various vectors sizes for the language pair EN→PT

The results presented in Tables 3 and 4 show that each language pair (considering the direction of the translation) has an optimal vectors' sizes configuration. Table 5 contains a summary of the best results found in each language pair.

Language Pair	P@1 (%)	Vectors' Size	P@5 (%)	Vectors' Size
EN→ES	79.42	400 - 200	88.19	500 - 300
ES→EN	75.66	400 - 200	87.62	500 - 200
PT→ES	71.48	600 - 200	85.22	400 - 700
ES→PT	70.97	400 - 200	82.58	400 - 700
EN→PT	79.17	200 - 200	89.84	400 - 300
PT→EN	73.17	400 - 200	89.02	500 - 400

Table 5. Summary of the best results for each translation experiment

Considering only the pairs EN→SP or SP→EN, the results were better than those presented in [9], including those produced with vectors trained on large corpora. Perhaps, the preprocessing or the way the dictionaries were created might have influenced the results. However, a simple analysis of the actual results shows that the procedure followed in this work is consistent. Several translation examples for all language pairs were randomly selected and are displayed in Tables 6, 7, 8, 9, 10 and 11. The vectors dimensions were selected in order to maximize the P@5 measure.

English Word	Computed Translation Candidates (Top 5)	Dictionary Translation (SP)
rape	prostitución esclavitud violación población civil intimidación	violación
cold war	guerra fría geopolítico unión soviética relaciones transatlánticas comunismo	guerra fría
albanian	albanés serbio belgrado serbios bosnia herzegovina	albanés

Table 6. Examples of translations from English to Spanish

SpanishWord	Computed Translation Candidates (Top 5)	Dictionary Translation (EN)
instrumentos financieros	financial instruments horizontal effective facilitate complement	financial instruments
declara	declares issued describes respected called	declares
consiguió	was reach got helped meant	got

Table 7. Examples of translations from Spanish to English

The translation candidates are ranked accordingly to their cosine similarity score. It can be seen that several words that are semantically related to the proper answer are suggested, which denotes the capability of generalization and context inference of distributed representations.

EnglishWord	Computed Translation Candidates (Top 5)	Dictionary Translation (PT)
assistants	assistentes salário promoções empregador intérpretes	assistentes
absolutely necessary	absolutamente necessário urgentemente necessário quadro legislativo legislar preferível	absolutamente necessário
commerce	comércio atrai turistas gerentes artistas	comércio

Table 8. Examples of translations from English to Portuguese

Portuguese Word	Computed Translation Candidates (Top 5)	Dictionary Translation (EN)
vírus	virus bovine vaccination tuberculosis feathers	virus
equidade	equity guarantee based sexes solidarity	equity
guerra civil	civil war terrible civilian population somalia oppression	civil war

Table 9. Examples of translations from Portuguese to English

4. Conclusion

In this paper, we have used the distributed representations of words, as proposed by Mikolov et al. [9] to perform automatic machine translation between the Portuguese, the English and the Spanish languages. We have properly evaluated the performance.

Portuguese Word	Computed Translation Candidates (Top 5)	Dictionary Translation (SP)
comparativamente	reducido inferior gasto aumentar aumentado	relativamente
desculpa	excusa intolerable golpear arma extremistas	excusa
mencionados	mencionado dicho discutido tratada debatido	mencionado

Table 10. Examples of translations from Portuguese to Spanish

SpanishWord	Computed Translation Candidates (Top 5)	Dictionary Translation (PT)
prevén	prever disposi ç ões estabelecer acordo aplicado	prever
descubierto	descoberto transportado aconteceu alimentado grave	descoberto
crearÃaÛ	criará oferecerá gerar beneficiará aumentar	criará

Table 11. Examples of translations from Spanish to Portuguese

We have been able to improve some training aspects of Mikolov´s work, which showed to provide better results as well. As with any NLP task, preprocessing step is critical and could be the cause of some noticed discrepancies in the results. Regardless, experimentation results are very promising, with translation precision around 85% if we consider translations involving the Portuguese language, which Mikolov´s work does not consider.

For future work, we propose training the model on a large scale corpora and also increasing the training set to evaluate if it increases the overall accuracy.

Aknowledgment

The authors thank the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) for granting a research scholarship to Hendrik T. Macedo [Type/Level: DT-II, Process 310446/2014-7] and for the financial support [Universal 14/2012, Process 483437/2012-3].

References

[1] Bertoldi, Nicola., Simianer, Patrick., Cettolo, Mauro., Wäschle, Katharina., Federico, Marcello., Riezler, Stefan. (2014). Online adaptation to post-edits for phrase-based statistical machine translation. *Machine Translation*, 28 (3- 4) 309–339.

[2] Choi, Heeyoul., Cho, Kyunghyun., Bengio, Yoshua. (2017). Context-dependent word representation for neural machine translation. Computer Speech Language, 45. 149 – 160.

[3] Devlin, Jacob., Zbib, Rabih., Huang, Zhongqiang., Lamar, Thomas., M Schwartz, Richard., Makhoul, John. (2014). Fast and robust neural network joint models for statistical machine translation. *In:* ACL (1) 1370–1380.

[4] Koehn, Philipp. (2005). Europarl: A parallel corpus for statistical machine translation. *In:* MT summit, volume 5, p. 79–86.

[5] Le, Quoc., Mikolov, Tomas. (2014). Distributed representations of sentences and documents. *In*: Proceedings of the 31st International Conference on Machine Learning (ICML-14), p. 1188–1196.

[6] LeCun, Yann., Bengio, Yoshua., Hinton, Geoffrey. (2015). Deep learning. *Nature*, 521 (7553) 436–444.

[7] Levy, Omer., Goldberg, Yoav., Dagan, Ido. (2015). Improving distributional similarity with lessons learned from word embeddings. *Transactions of the Association for Computational Linguistics*, 3. 211–225.

[8] Mikolov, Tomas., Chen, Kai., Corrado, Greg., Dean, Jeffrey. (2013). Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.

[9] Mikolov, Tomas.,V Le, Quoc., Sutskever, Ilya. (2013). Exploiting similarities among languages for machine translation. arXiv preprint arXiv:1309.4168.

[10] Mikolov, Tomas., Sutskever, Ilya., Chen, Kai.,Corrado, Greg S., Dean, Jeff. (2013). Distributed representations of words and phrases and their compositionality. *In:* Advances in neural information processing systems, p. 3111– 3119.

[11] Mikolov, Tomas., Wen-tau Yih, Scott., Zweig, Geoffrey. (2013). Linguistic regularities in continuous space word representations. *In*: Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT-2013). Association for Computational Linguistics.

[12] Salami, Shahram., Shamsfard, Mehrnoush., Khadivi, Shahram. (2016). Phrase-boundary model for statistical machine translation. Computer Speech Language, 38. 13 – 27.

[13] Zhang, M., Liu, Y., Luan, H., Sun, M. (2016). Listwise ranking functions for statistical machine translation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24 (8) 1464–1472.