Baltic J. Modern Computing, Vol. 4 (2016), No. 2, pp. 129–140

Improving Phrase-Based SMT Using Cross-Granularity Embedding Similarity

Peyman PASSBAN, Chris HOKAMP, Andy WAY, Qun LIU

ADAPT Centre
School of Computing
Dublin City University
Dublin, Ireland

{ppassban,chokamp,away,qliu}@computing.dcu.ie

Abstract. The phrase–based statistical machine translation (PBSMT) model can be viewed as a log-linear combination of translation and language model features. Such a model typically relies on the phrase table as the main resource for bilingual knowledge, which in its most basic form consists of aligned phrases, along with four probability scores. These scores only indicate the co-occurrence of phrase pairs in the training corpus, and not necessarily their semantic relatedness. The basic phrase table is also unable to incorporate contextual information about the segments where a particular phrase tends to occur. In this paper, we define six new features which express the semantic relatedness of bilingual phrases. Our method utilizes both source and target side information to enrich the phrase table. The new features are inferred from a bilingual corpus by a neural network (NN). We evaluate our model on the English–Farsi (En–Fa) and English–Czech (En–Cz) pairs and observe considerable improvements in the all En↔Fa and En↔Cz directions.

Keywords: Statistical machine translation, phrase embeddings, incorporating contextual information.

1 Introduction

The process of PBSMT can be interpreted as a search problem where the score at each step of exploration is formulated as a log-linear model (Koehn, 2010). For each candidate phrase, the set of features is combined with a set of learned weights to find the best target counterpart of the provided source sentence. Because an exhaustive search of the candidate space is not computationally feasible, the space is typically pruned via some heuristic search, such as beam search (Koehn, 2010). The discriminative log-linear model allows the incorporation of arbitrary context-dependent and context-independent features. Thus, features such as those in Och and Ney (2002) or Chiang et al. (2009) can be combined to improve translation performance. The standard baseline bilingual features included in Moses (Koehn et al., 2007) by default are: the *phrase translation*

probability $\phi(e|f)$, *inverse phrase translation probability* $\phi(f|e)$, *direct lexical weighting* $lex(e|f)$ and *inverse lexical weighting* $lex(f|e)$.[1]

The scores in the phrase table are computed directly from the co-occurrence of aligned phrases in training corpora. A large body of recent work evaluates the hypothesis that co-occurrence information alone cannot capture contextual information as well as the semantic relations among phrases (see section 2). Therefore, many techniques have been proposed to enrich the feature list with semantic information. In this paper, we define six new features for this purpose. All of our features indicate the semantic relatedness of source and target phrases. Our features leverage contextual information which is lost by the traditional phrase extraction operations. Specifically, in both sides (source and target) we look for any type of constituents including phrases, sentences or even words which can fortify the semantic information about phrase pairs.

Our contributions in this paper are threefold: a) We define new semantic features and embed into PBSMT to enhance the translation quality. b) In order to define the new features we train bilingual phrase and sentence embeddings using an NN. Embeddings are trained in a joint distributed feature space which not only preserves monolingual semantic and syntactic information but also represents cross-lingual relations. c) We indirectly incorporate external contextual information using the neural features. We search in the source and target spaces and retrieve the closest constituent to the phrase pair in our bilingual embedding space.

The structure of the paper is as follows. Section 2 gives an overview of related work. Section 3 explains our pipeline and the network architecture in detail. In Section 4, experimental results are reported. We also have a separate section to discuss different aspects of embeddings and the model. Finally, in the last section we present our conclusions along with some avenues for future work.

2 Background

Several models such as He et al. (2008), Liu et al. (2008) and Shen et al. (2009) studied the use of contextual information for statistical machine translation (SMT). The idea is to go beyond the phrase level and enhance the phrase representation by taking surrounding phrases into account. This line of research is referred as discourse SMT (Hardmeier, 2014; Meyer, 2014). Because NNs can provide distributed representations for words and phrases, they are ideally suited to the task of comparing semantic similarity. Unsupervised models such as *Word2Vec*[2] (Mikolov et al., 2013a) or *Paragraph Vectors* (Le & Mikolov, 2014) have shown that distributional information is often enough to learn high-quality word and sentence embeddings.

A large body of recent work has evaluated the use of embeddings in machine translation. A successful usecase was reported in (Mikolov et al., 2013b). They separately

[1] Although the features contributed by the language model component are as important as the bilingual features, we do not address them in this paper, since they traditionally only make use of the monolingual target language context, and we are concerned with incorporating bilingual semantic knowledge.

[2] http://code.google.com/p/word2vec/

project words of source and target languages into embeddings, then try to find a transformation function to map the source embedding space into the target space. The transformation function was approximated using a small set of word pairs extracted using an unsupervised alignment model trained with a parallel corpus. This approach allows the construction of a word-level translation engine with very large monolingual data and only a small number of bilingual word pairs. The cross-lingual transformation mechanism allows the engine to search for translations for OOV (out-of-vocabulary) words by consulting a monolingual index which contains words that were not observed in the parallel training data.The work by Garcia and Tiedemann (2014) is another model follows that the same paradigm.

However, machine translation (MT) is more than word-level translation. In Martínez et al. (2015) word embeddings were used in document-level MT to disambiguate the word selection. Tran et al. (2014) used bilingual word embeddings to compute the semantic similarity of phrases. To extend the application of text embedding beyond single words, Gao et al. (2013) proposed learning embeddings for source and target phrases by training a network to maximize the sentence-level BLEU score. Costa-jussa et al. (2014) worked at the sentence-level and incorporated the source side information into the decoding phase by finding the similarities between phrases and source embeddings. Some other models re-scored the phrase table (Alkhouli et al., 2014) or generated new phrase pairs in order to address the OOV word problem (Zhao et al., 2014).

Our network makes use of some ideas from existing models, but also extends the information available to the embedding model. We train embeddings in the joint space using both source and target side information simultaneously, using a model which is similar to that of Devlin et al. (2014) and Passban et al. (2015b). Similar to Gao et al. (2013) we make embeddings for phrases and sentences and add their similarity as feature functions to the SMT model.

3 Proposed Method

In order to train our bilingual embedding model, we start by creating a large bilingual corpus. Each line of the corpus may include:

- a source or target sentence,
- a source or target phrase,
- a concatenation of a phrase pair (source and target phrases which are each other's translation),
- a tuple of source and target words (each other's translation).

Sentences of the bilingual corpus are taken from the SMT training corpus. Accordingly, phrases and words are from the phrase tables and lexicons, generated by the alignment model and phrase extraction heuristic used by the SMT model. This means that the bilingual corpus is a very large corpus with size of $2 * |c| + 3 * |pt| + |bl|$ which $|c|$ indicates the number of source/target sentences, $|pt|$ is the size of the phrase table and $|bl|$ is the size of the bilingual lexicon.

By use of the concatenated phrases and bilingual tuples we try to score the quality of both sides of the phrase pair, by connecting phrases with other phrases in the same

language, and with their counterparts in the other language. Section 3.1 discusses how the network benefits from this bilingual property.

Each line of the bilingual training corpus has a dedicated vector (row) in the embeddings matrix. During training embeddings are updated. After training, we extract some information to enrich the phrase table. First we compute the semantic similarity between source and target phrases in phrase pairs. The similarity shows how semantically phrases are related to each other. The *Cosine* measure is used to compute the similarity:

$$similarity(E_s, E_t) = \frac{E_s.E_t}{||E_s|| \times ||E_t||}$$

where E_s and E_t indicate embeddings for the given source and target phrases, respectively. We map *Cosine* scores into the [0,1] range. This can be interpreted as a score indicating the semantic relatedness of the source and target phrases. The similarity between the source phrase and target phrase is the first feature and is referred as *sp2tp*.

Among source-side embeddings (word, phrase or sentence embeddings) we search for the close match to the source phrase. There might be a word, phrase or sentence on the source side which can enhance the source phrase representation and ease its translation. If the closest match belongs to a phrase, probably that is a paraphrased form of the original phrase and if the closest match belongs to a word, probably that is a keyword which could enhance the word selection quality. We refer to this source-side similarity score as *sp2sm*.

We also look for the closest match of the source phrase on the target side. As we jointly learn embeddings, structures that are each other's translation should have close embeddings. We compute the similarity of the closest target match to the source phrase (*sp2tm*). We compute the same similarities for the target phrase, namely the similarity of the target phrase with the closest target match (*tp2tm*) and the closest source match (*tp2sm*). The source and target matches may preserve other type of semantic similarity (*sm2tm*), therefore these features should add more information about the overall quality of the phrase pair. All new features are added to the phrase table and used in the tuning phase to optimise the translation model. Figure 1 tries to clarify the relation among different matches and phrases.

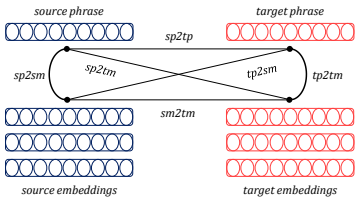

Fig. 1. *sp, tp, sm* and *tm* stand for *source phrase, target phrase, source match* and *target match*, respectively. The embeddings size for all types of embedding are the same. The source/target-side embedding could belong to a source/target word, phrase or sentence. The labels of arrows indicate the Cosine similarity between two embeddings which is mapped into the [0,1] range.

3.1 Learning Embeddings

Our network is an extension of Le and Mikolov (2014) and Passban et al. (2015b). In those methods, documents (words, phrases, sentences and any other chunks of text) are treated as atomic units in order to learn embeddings in the same semantic space as the space used for the individual words in the model. The model includes an embedding for each document which in our case may be a monolingual sentence, a monolingual phrase, a bilingual phrase pair or a bilingual word pair. During training, at each iteration a random target word (w^t) is selected from the input document to be predicted at the output layer by using the context and document embeddings. The context embedding is made by averaging embeddings of adjacent words around the target word. Word and document embeddings are updated during training until the cost is minimized. The model learns an embedding space in which constituents with similar distributional tendencies are close to each other. More formally, given a sequence of $S_i = w_1, w_2, ..., w_n$ the objective is to maximize the log probability of the target word given the context and document vector:

$$\frac{1}{n}\sum_{j=1}^{n}\log p(w_j^t|C_i^{w^t}, D_i)$$

where $w_j^t \in S_i$ is randomly selected at each iteration. D_i is the document embedding for S_i and C^{w^t} indicates the context embedding which is the mean of embeddings for m preceding and m following words around the target word.

As previously mentioned, S_i could be a monolingual sentence or phrase, in which case w^t and adjacent words are from the same language. In other words, the context includes m words before and m words after the target word. S_i also could be a concatenation of source and target phrases. In that case context words are selected from both languages, i.e. m words from the source (the side from which the target word is selected) and m words from the target side. Finally S_i could be a pair of source and target words where C^{w^t} is made using the target word's translation. The word on one side is used to predict the word on the opposite side. In the proposed model m is the upper bound.

Table 1. Context vectors for different input documents. w^t is **better** and $m = 5$. Italics are in Farsi.

D_1	know him **better** than anyone
C_1^{better}	[know, him, than, anyone]$_s$
D_2	know him **better** than anyone . *āv rā bhtr āz hrks myšnāsy*
C_2^{better}	[know, him, than, anyone]$_s$ + [*āv*, *rā*, *bhtr*, *āz*, *hrks*]$_t$
D_3	**better** . *bhtr*
C_3^{better}	[*bhtr*]$_t$

Table 1 illustrate some examples of the context window. The examples are selected from the En–Fa bilingual corpus (see Section 4).[3] In C_1 the context window includes 2 words before **better** and 2 words after. In this case the target word and all other context words are from the same language (indicated by a 's' subscript). In the second example the input document is a concatenation of English and Farsi phrases, so C_2 includes m (or fewer) words from each side (indicated with different subscripts). In the final example the input document is a word tuple where the target word's translation is considered as its context.

As shown in Huang et al. (2012), word vectors can be affected by the word's surrounding as well as by the global structure of a text. Each unique word has a specific vector representation and clearly similar words in the same language would have similar vectors (Mikolov et al., 2013a). By use of the bilingual training corpus and our proposed architecture we tried to expand the monolingual similarities to the bilingual setting, resulting in an embedding space which contains both languages. Words that are direct translations of each other should have similar/close embeddings in our model. As the corpus contains tuples of $< word_{L_1}, word_{L_2} >$, embeddings for words which tend to be translations of one another are trained jointly. Phrasal units are also connected together by the same process. Since the bigger blocks encompass the embeddings for words and phrasal units they should also have representations which are similar to the representations of their constituents.

3.2 Network Architecture

In the input layer we have an embedding matrix. Each row in the matrix is dedicated to one specific line in the bilingual corpus. During training embeddings are tuned and updated. The network has only one hidden layer. A *Softmax* layer is placed on top of the hidden layer to map values to class probabilities. *Softmax* is a vector-valued function which maps its input values to the [0,1] range. The output values from the *Softmax* can be interpreted as class probabilities for the given input. The *Softmax* function is formulated as follows:

$$P(w_j^t | C_i^{w^t} \bullet D_i) = \frac{\exp(h_j.w_j + a_j)}{\sum_{j' \in \mathcal{V}} \exp(h_j.w_{j'} + a_{j'})}$$

Intuitively, we are estimating the probability of selecting the j-th word as the target word from the i-th training document. The input for the *Softmax* layer is $h = W(C_i^{w^t} \bullet D_i) + b$, where W is a weight matrix between the input layer and the hidden layer, b is a bias vector and $\bullet$ indicates the concatenation function. w_j is the j-th column of another weight matrix (between the hidden layer and the *Softmax* layer) and a_j is a bias term. The output of *Softmax*, $V \in \mathbb{R}^{|\mathcal{V}|}$, is the distribution probability over classes which are words in our setting. The j-th cell in V is interpreted as the probability of selecting the j-th word from the target vocabulary $\mathcal{V}$ as the target word. Based on *Softmax* values the word with the highest probability is selected and the error is computed accordingly. The network parameters are optimized using stochastic gradient descent and

[3] We used the DIN transliteration standard to show the Farsi alphabets; https://en.wikipedia.org/wiki/Persian_alphabet

back-propagation (Rumelhart et al., 1988). All parameters of the model are randomly initialized over a uniform distribution in the [-0.1,0.1] range. Weight matrices, bias values and word embeddings are all network parameters which are tuned during training. The embedding size in our model is 200. Figure 2 illustrates the whole pipeline.

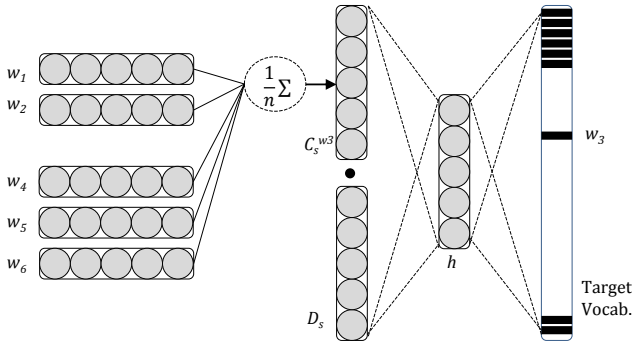

Fig. 2. Network architecture. The input document is $S = w_1\ w_2\ w_3\ w_4\ w_5\ w_6$ and the target word is w_3.

4 Experimental Results

We evaluated our new features on two language pairs: En–Fa and En–Cz. Both Farsi and Czech are morphologically rich languages; therefore, translation to/from these languages can be more difficult than it is for languages where words tend to be discrete semantic units. Farsi is also a low-resource language, so we are interested in working with these pairs. For the En–Fa pair we used the TEP++ corpus (Passban et al., 2015a) and for Czech we used the Europarl[4] corpus (Koehn, 2005). TEP++ is a collection of 600,000 parallel sentences. We used 1000 and 2000 sentences for testing and tuning, respectively and the rest of the corpus for training. From the Czech dataset we selected the same number of sentences for training, testing and tuning. The baseline system is a PBSMT engine built using Moses (Koehn et al., 2007) with the default configuration. We used MERT (Och, 2003) for tuning. In the experiments we trained 5-gram language models on the monolingual parts of the bilingual corpora using SRILM (Stolcke et al., 2002). We used BLEU (Papineni et al., 2002) as the evaluation metric. We added our features to the phrase table and tuned the translation models. Table 2 shows the impact of each feature. We also estimated the translation quality in the presence of the all features (we run MERT for each row of Table 2). Bold numbers are statistically significant according to the results of paired bootstrap re-sampling with p=0.05 for 1000 samples (Koehn, 2004). Arrows indicate whether the new features increased or decreased the quality over the baseline.

[4] http://www.statmt.org/europarl/

Table 2. Impact of the proposed features.

Feature	En–Fa	↑↓	Fa–En	↑↓	En–Cz	↑↓	Cz–En	↑↓
Baseline	21.03	0.00	29.21	0.00	28.35	0.00	39.63	0.00
sp2tp	21.46	**0.43** ↑	29.71	**0.50** ↑	28.72	**0.37** ↑	40.34	**0.71** ↑
sp2sm	21.32	0.29 ↑	29.74	**0.53** ↑	28.30	0.05 ↓	39.76	0.13 ↑
sp2tm	21.40	**0.37** ↑	29.56	**0.35** ↑	28.52	0.17 ↑	39.79	0.16 ↑
tp2tm	20.40	0.63 ↓	29.56	0.35 ↑	28.00	0.35 ↓	39.68	0.05 ↑
tp2sm	21.93	**0.90** ↑	29.26	0.05 ↑	28.94	**0.59**↑	39.81	0.18 ↑
sm2tm	21.18	0.15 ↑	30.08	**0.87** ↑	28.36	0.01 ↑	39.99	**0.36** ↑
All	21.84	**0.81** ↑	30.26	**1.05** ↑	29.01	**0.66** ↑	40.24	**0.61** ↑

Results show that the new features are useful and positively affect the translation quality. Some of the features such as *sp2tp* are always helpful regardless of the translation direction and language pair. This feature is the most important feature among others. The *sm2tm* feature always works effectively in translating into English and the *tp2sm* feature is effective when translating from English. In the presence of all features results are significantly better than the baseline system in all cases. Some of the features are not as strong as the others (*tp2tm*) and some of them behave differently based on the language (*sp2tm*).

5 Discussion

Numbers reported in in Section 4 indicate that the proposed method and features result in a significant enhancement of translation quality, but it cannot be decisively claimed that they are always helpful for all languages and settings. Therefore we tried to study the impact of features not only quantitatively but also qualitatively. We mainly focus on three issues in this section. First we show how the features change SMT translations. Then we show ability of the network in capturing cross-lingual similarities and finally we discuss the way we learn embeddings.

Based on our investigation, the new features seem to help the model determine the quality of a phrase pair. As an example for the English phrase *"but I'm your teammate"* in the phrase table, the corresponding Farsi target phrase is *"āmā mn hm tymyt hstm"* which is the exact translation of the source phrase. The closest match in the source side is *"we played together"* and in the target side is *"Ben mn ānjā bāzy krdm"* (meaning *"I played in that team"*). These retrieved matches indicate that this is a high-quality phrase. By comparing the outputs we recognized that before adding our features the word *"your"* was not translated. In translation into Farsi, possessives sometimes are not translated and the verb implicitly shows them, but the best translation is a translation including possessives. The translation of *"your"* appeared in the output after adding our features.

The proposed model is expected to learn the cross-lingual similarities along with the monolingual relations. To study this feature Table 3 shows two samples. Results in Table 3 show the proposed model can capture cross-lingual relations. It is also able to

model similarities in different granularities. It has word level, phrase level and sentence level similarities. Retrieved instances are semantically related to the given queries.

Table 3. The top 10 most similar vectors for the given English query. Recall that the retrieved vectors could belong to words, phrases or sentences in either English or Farsi and word or phrase pairs. The items that were originally in Farsi have been translated into English, and are indicated with *italics*.

Query	***sadness***
1	<*apprehension*, nervous>
2	*emotion*
3	<*ill*,sick>
4	pain
5	<*money*,money>
6	*benignity*
7	<*may he was punished*,punished harshly>
8	is really gonna hurt
9	i know tom ' s dying
10	<*bitter*,angry>

Tang et al. (2015) proposed that a sentence embedding could be generated by averaging/concatenating embeddings of the words in that sentence. In our case the model by Tang et al. was not as beneficial as ours for both Farsi and Czech. As an example if the *sp2tp* is computed using their model, it degrades the En–Fa direction's BLEU from 21.03 to 20.97 and its improvement for the Fa–En direction is only +0.11 points (almost 5 times less than ours). Our goal is not to compare our model to that of Tang et al.. We only performed a simple comparison on the most important feature to see the difference. Furthermore, according to discussions from Le and Mikolov (2014) document vectors (such as ours) work better than averaging/concatenating vectors. Our model also contains both source and target side information in word and phrase embeddings. Averaging cannot provide such rich information. Our results are aligned with Devlin et al. (2014), who showed the impact of using both source and target side information.

6 Conclusion and Future work

In this work we proposed a novel neural network model which learns word, phrase, and sentence embeddings in a bilingual space. Using embeddings we define six new features which are incorporated into an SMT phrase table. Our results show that the new semantic similarity features enhance translation performance across all of the languages we evaluated. In future work, we hope to directly include the distributed semantic representation into the phrase table, allowing on-line incorporation of semantic information into the translation model features.

Acknowledgement

We would like to thank the three anonymous reviewers and Rasul Kaljahi for their valuable comments and the Irish Center for High-End Computing (www.ichec.ie) for providing computational infrastructures. This research is supported by Science Foundation Ireland through the CNGL Programme (Grant 12/CE/I2267) in the ADAPT Centre (www.adaptcentre.ie) at Dublin City University.

References

Alkhouli, T., Guta, A., & Ney, H. (2014). Vector space models for phrase-based machine translation. *Syntax, Semantics and Structure in Statistical Translation*.

Chiang, D., Knight, K., & Wang, W. (2009). 11,001 new features for statistical machine translation. In *Proceedings of human language technologies: The 2009 annual conference of the north american chapter of the association for computational linguistics* (pp. 218–226). Boulder, Colorado.

Costa-jussa, M., Gupta, P., Rosso, P., & Banchs, R. (2014). English-to-hindi system description for wmt 2014: Deep sourcecontext features for moses. In *Proceedings of the ninth workshop on statistical machine translation, baltimore, maryland, usa. association for computational linguistics.*

Devlin, J., Zbib, R., Huang, Z., Lamar, T., Schwartz, R., & Makhoul, J. (2014). Fast and robust neural network joint models for statistical machine translation. In *Proceedings of the 52nd annual meeting of the association for computational linguistics* (Vol. 1, pp. 1370–1380).

Gao, J., He, X., Yih, W., & Deng, L. (2013). Learning semantic representations for the phrase translation model. *CoRR*, *abs/1312.0482*.

Garcia, E. M., & Tiedemann, J. (2014). Words vector representations meet machine translation. *Syntax, Semantics and Structure in Statistical Translation*, 132.

Hardmeier, C. (2014). *Discourse in statistical machine translation*. Unpublished doctoral dissertation.

He, Z., Liu, Q., & Lin, S. (2008). Improving statistical machine translation using lexicalized rule selection. In *Proceedings of the 22nd international conference on computational linguistics - volume 1* (pp. 321–328). Stroudsburg, PA, USA: Association for Computational Linguistics. Retrieved from http://dl.acm.org/citation.cfm?id=1599081.1599122

Huang, E. H., Socher, R., Manning, C. D., & Ng, A. Y. (2012). Improving word representations via global context and multiple word prototypes. In *Proceedings of the 50th annual meeting of the association for computational linguistics: Long papers-volume 1* (pp. 873–882).

Koehn, P. (2004). Statistical significance tests for machine translation evaluation. In *Emnlp* (pp. 388–395).

Koehn, P. (2005). Europarl: A parallel corpus for statistical machine translation. In *Mt summit* (Vol. 5, pp. 79–86).

Koehn, P. (2010). *Statistical machine translation* (1st ed.). New York, NY, USA: Cambridge University Press.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., et al. (2007). Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th annual meeting of the acl on interactive poster and demonstration sessions* (pp. 177–180).

Le, Q. V., & Mikolov, T. (2014). Distributed representations of sentences and documents. *CoRR*, *abs/1405.4053*.

Liu, Q., He, Z., Liu, Y., & Lin, S. (2008). Maximum entropy based rule selection model for syntax-based statistical machine translation. In *Proceedings of the conference on empirical methods in natural language processing* (pp. 89–97). Stroudsburg, PA, USA: Association for Computational Linguistics. Retrieved from http://dl.acm.org/citation.cfm?id=1613715.1613729

Martínez, E., España Bonet, C., Márquez Villodre, L., et al. (2015). Document-level machine translation with word vector models. In *Proceedings of the 18th annual conference of the european association for machine translation (eamt)* (pp. 59–66). Antalya, Turkey.

Meyer, T. (2014). *Discourse-level features for statistical machine translation*. Unpublished doctoral dissertation, École Polytechnique Fédérale de Lausanne.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013a). Efficient estimation of word representations in vector space. *CoRR*, *abs/1301.3781*.

Mikolov, T., Le, Q. V., & Sutskever, I. (2013b). Exploiting similarities among languages for machine translation. *CoRR*, *abs/1309.4168*.

Och, F. J. (2003). Minimum error rate training in statistical machine translation. In *Proceedings of the 41st annual meeting on association for computational linguistics - volume 1* (pp. 160–167). Sapporo, Japan.

Och, F. J., & Ney, H. (2002). Discriminative training and maximum entropy models for statistical machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics* (pp. 295–302). Philadelphia, Pennsylvania.

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics* (pp. 311–318).

Passban, P., Hokamp, C., & Liu, Q. (2015b). Bilingual distributed phrase representation for statistical machine translation. In *Proceedings of mt summit xv* (pp. 310–318).

Passban, P., Way, A., & Liu, Q. (2015a). Benchmarking SMT performance for Farsi unisng the TEP++ corpus. In *Proceedings of the 18th annual conference of the European Association for Machine Translation (eamt)* (pp. 82–88). Antalya, Turkey.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1988). Learning representations by back-propagating errors. *Cognitive modeling*, *5*, 3.

Shen, L., Xu, J., Zhang, B., Matsoukas, S., & Weischedel, R. (2009). Effective use of linguistic and contextual information for statistical machine translation. In *Proceedings of the 2009 conference on empirical methods in natural language processing: Volume 1 - volume 1* (pp. 72–80). Singapore.

Stolcke, A., et al. (2002). SRILM-an extensible language modeling toolkit. In *Interspeech.*

Tang, D., Qin, B., & Liu, T. (2015). Document modeling with gated recurrent neural network for sentiment classification. In *Proceedings of the 2015 conference on empirical methods in natural language processing* (pp. 1422–1432).

Tran, K. M., Bisazza, A., & Monz, C. (2014). Word translation prediction for morphologically rich languages with bilingual neural networks. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (pp. 1676–1688).

Zhao, K., Hassan, H., & Auli, M. (2014). Learning translation models from monolingual continuous representations.

Received May 2, 2016 , accepted May 5, 2016