Bilingual Distributed Phrase Representations for Statistical Machine Translation

Peyman Passban ppassban@computing.dcu.ie
Chris Hokamp chokamp@computing.dcu.ie
Qun Liu qliu@computing.dcu.ie
ADAPT Centre, School of Computing, Dublin City University, Ireland

Abstract

Phrase–based machine translation (PBMT) relies upon the phrase-table as the main resource for bilingual knowledge at decoding time. A phrase table in its basic form includes aligned phrases along with four probabilities indicating aspects of the co-occurrence statistics for each phrase pair. In this paper we add a new semantic similarity score as a statistical feature to enrich the phrase table. The new feature is inferred from a bilingual corpus by a neural network (NN), and estimates the semantic similarity of each source and target phrase pair. We observe a significant increase in system performance with the addition of the new feature. We evaluated our model on the English–French (En–Fr) and English–Farsi (En–Fa) language pairs. Experimental results show improvements for all translation directions of En↔Fr and En↔Fa.

1 Introduction

Phrase–based machine translation begins by segmenting a source sentence into phrases and looking up candidate translations for the sub-sentential phrases in a phrase table. The goal of PBMT is to assemble these *translation candidates* into the optimal target sequence. Finding the best source segmentation, translation candidates, and phrase ordering is a search problem which is typically formulated as a log-linear model using both dynamic and static features, whose weights are optimized via a search heuristic on held-out development data. The inverse phrase translation probability $\varphi(f|e)$, inverse lexical weighting $lex(f|e)$, direct phrase translation probability $\varphi(e|f)$ and direct lexical weighting $lex(e|f)$ are four of the standard static features used in phrase tables. These features are computed directly from the co-occurrence of aligned phrases in the training corpora. However, co-occurrence information alone cannot capture semantic information about phrases, especially when they are taken out of context. Therefore, many techniques have been proposed to enrich the feature list by including features which contain syntactic and/or semantic information (Banchs and Costa-jussà, 2011).

Most work evaluating the inclusion of semantic information into SMT decoders has focused upon adding dynamic features (those which must be computed at decoding time). However, dynamic features require the implementation of a new feature function which depends upon the hypothesis data structure of the particular decoder. The implementation of dynamic features typically requires significantly more engineering effort than simply augmenting the phrase table. In this paper, we add a new static feature and show how a good static feature alone can significantly boost translation quality. The basic idea behind our work is to use the vector representation, or *semantic embedding* of phrases, which is generated by an NN. The semantic features of the source and target languages are projected into a shared bilingual space,

which preserves both semantic and syntactic information about the phrases. The supervised approach to generating embeddings from neural networks optimizes the network to produce vectors which are good with respect to a specific objective. As an example, if the goal is to do the sentiment analysis or sentence clustering, a vector should reflect the polarity of a sentence — whether is positive or negative — or its closeness to a specific distribution (Kalchbrenner et al., 2014). In tasks like translation, a good vector should reflect semantic and syntactic information about original constituent (sentence, phrase or word) in addition to contextual knowledge from its surrounding words, and ideally some information which will make it easier to map into the target language.

Methods like word2vec[1] (Mikolov et al., 2013a) or that of (Le and Mikolov, 2014) produce general-purpose vector representations which can be leveraged for a variety of downstream applications. As the word and sentence vectors encode syntactic and semantic information, they are potentially useful for translation tasks. In pure neural MT engines, the word embeddings are trained as parameters of the model, which generally attempts to maximize the likelihoods (Kalchbrenner and Blunsom, 2013; Bahdanau et al., 2014), and then used directly to perform translation. Another line of work has tried to make use of distributed representations within the classic MT pipeline (Gao et al., 2013; Devlin et al., 2014).

In this work, we prepare an enhanced bilingual corpus which includes the source and target phrases extracted by the alignment model, along with the original source and target sentences, and a set of word pairs which are direct translations of one another. An NN is used to generate embeddings for the each of the phrases. The generated vectors reflect the similarity of phrases in the same language as well as their relevancy to the phrases of the other language. In the training data, monolingual distributional information for each language is contained in the sentences, while bilingual information is conveyed by the bilingual phrase pairs and word pairs.

Our contributions in this paper are twofold. a) We extract a novel static feature from a bilingual corpus which boosts the translation quality. Most previous work has added new dynamic features which significantly increase computational overhead — our simple static feature can achieve comparable improvements with less effort. b) Our network extracts the vectors from a bilingual corpus. To the best of our knowledge, related research has modeled the source and target vectors separately in isolated spaces and focused upon finding a means to transform the source & target representations into comparable forms. We extracted vectors in the joint space and tried to capture information of the both target and source sides in a single vector. Inferred feature caused considerable improvements and showed that, a good static feature can perform as efficiently as a dynamic one. The structure of paper is as follows. Section 2 gives an overview of related work, and tries to show why word, phrase, and sentence vectors are useful for MT purposes. Section 3 explains the network architecture in detail. In Section 4, experimental results are reported for two language pairs: En–Fr and En–Fa. Finally, in the last section we present our conclusions and discuss some avenues for future work.

2 Background

Using vector representations for textual data (word, phrase, sentence, etc...) is not a new idea. The concept of continuous, distributed representations for text tokens was introduced by the Vector Space Model of Salton et al. (1975) and expanded by techniques such as like Latent Semantic Analysis (Deerwester et al., 1990), Latent Dirichlet Allocation (Blei et al., 2003), and Random Indexing (Sahlgren, 2005). Real valued, continuous representations are straightforward to use within Machine Learning models, and have contributed to the current state-of-the-art in many NLP tasks. However, a disadvantage of distributed representations is

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

that they are typically not decomposable — i.e. columns often do not correspond to intuitive features, thus models must be evaluated on a task-by-task basis.

The techniques mentioned in the previous paragraph build vectors from bag-of-words (BOW) representations, discarding structural dependencies and word order, so they cannot not reflect semantic information which depends upon syntactic structure. To compensate for these deficiencies, a more recent line-of-work uses neural networks to generate vectors which can encode local distributional information, as well as syntactic structure and word order in some cases. Hinton (1986) used NN for text modelling for the first time. Recently, neural networks have achieved state-of-the-art performance in many areas of NLP due both to the development of new learning algorithms and to the availability of computational resources such as GPUs. Word2vec (Mikolov et al., 2013a), and word2vec inspired works (Wolf et al., 2014) have been successfully applied to a wide variety of NLP tasks. The following paragraph focuses upon work that leverages these vectors for MT purposes.

A basic but successful application of NN-based word vectors for MT was reported in (Mikolov et al., 2013b). They project words of the source language into vectors and do the same with words of target language. Then try to find a transformation function which maps the source semantic space into the target semantic space using a small set of word pairs known to be high-quality translations. The model significantly reduces the volume of bilingual data required to train such systems and this is the main advantage of their approach. The model works on monolingual data and only needs a small number of parallel words to make the bridge between languages. The cross-lingual transformation allows an MT system 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 for the MT system. Garcia and Tiedemann (2014) and Dinu and Baroni (2014) are other examples of approaches which leverage NN-based word vectors for translation tasks. They focus upon exploiting similarities at the word level, but MT encompasses more than just word-level translation. To extend the application of text embeddings 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. The outcome is a set of vectors for phrases, and the similarity between each phrase pair vectors is used as a dynamic feature function in the log–linear model at decoding time. In another work Costa-jussa et al. (2014) tried to find the similarities among source sentences and incorporate source side contextual information into the decoding. Some other models try to re-score the phrase table or infer new phrase pairs to address the OOV word problem in order to improve translation quality (Alkhouli et al., 2014; Costa-Jussà and Banchs, 2011). Zhao et al. (2015) did the same using monolingual datasets and extended the phrase table.

3 Learning Embeddings for Phrases

Our model extends the document vectors of Le and Mikolov (2014) to bilingual texts. By including both monolingual and bilingual 'documents' into the training data, we learn a distributed representation for both languages simultaneously. In the method proposed by Le and Mikolov (2014), documents are treated as atomic units in order to learn an embedding with the same dimensionality as the vectors for the individual words in the model. We adapt this idea to sentences and phrases, where phrases are presented both as monolingual and as bilingual documents.

Le and Mikolov (2014) create a new vector 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, the document vector is concatenated with the vectors for individual words to predict the surrounding words in the given unit of text. Intuitively, we expect document vectors to be representative of the semantic content of the entire unit of text, while word vectors are

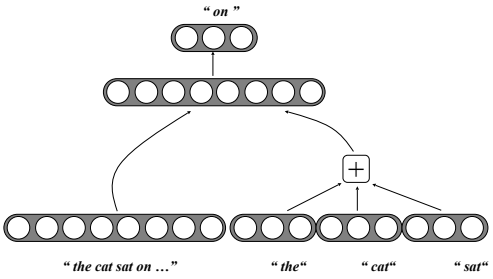

Figure 1: Network architecture for jointly learning sentence and word vectors

representative of all of the contexts where the word occurs. During training, word vectors and sentence/phrase vectors are updated until the cost is minimized. The model learns a semantic space where constituents with similar distributional tendencies tend to have similar vectors. More formally, given a sequence of tokens $c_1, c_2, ..., c_n$ the objective is to maximize the average log probability of a word given its context:

$$\frac{1}{n}\sum_{i=l}^{n-l} \log p(c_i|c_{i-l}, ..., c_{i+l})$$

Using the standard terminology for NN models, the training objective can be explained as follows. The cost function is defined according to the average log probability. Values for the probabilities can be calculated using a multiclass classifier such as softmax:

$$p(c_i|c_{i-l}, ..., c_{i+l}) = \frac{e^{y_{w_t}}}{\sum_j e^{y_j}}$$

where y_j is the output for the input word w_j in which:

$$y = b + Wh(c_{i-l}, ..., c_{i+l})$$

W and b are parameters for the Softmax function, and h represents the output of one or more hidden layers. The network is trained by stochastic gradient decent and back-propagation (Rumelhart et al., 1988) to obtain the set of constituent vectors C, and network parameters, W and b.

In our case, the network includes one hidden layer with 100 nodes which is fed by a bilingual corpus. The Corpus includes a) source and target sentences which are those we used to train our SMT engine, b) phrase pairs and c) bilingual lexicon. Both the phrase and lexicon sets are extracted by Moses (Koehn et al., 2007). Specifically, each line of corpus contains a sentence in one of the languages, a phrase pair, or a tuple of words.

As it has been shown in (Huang et al., 2012), word vectors can be affected by the word's surroundings as well as by the global structure of a text. Accordingly, by using a training corpus with some bilingual examples we expect to learn a semantic space which contains both languages. Each unique word has a specific vector representation. Clearly similar words in the same language would have similar vectors (Mikolov et al., 2013a). Words that are direct translations of each other (same meaning with different languages) should also have similar vectors. As the corpus contains the tuples of $< word_{L1}, word_{L2} >$, equal words are connected together. By the same logic, phrasal units are also connected together. During decoding, the sentences in a good translation pair should be built from similar sub-units, indicating the semantic compatibility of the constituent phrases. Table 1 shows the most similar phrases and words for two

examples. The items that were originally in Farsi have been translated into English, and are indicated with *italics*.

Query	we can't let them win
1	you could ever
2	what the bloody hell is that .
3	as you know it .
4	let her go .
5	he is lying . no i am not .
6	to be
7	*he won*
8	you !?
9	you got that .
10	*they are ahead*
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 toms dying
10	<*bitter* , angry >

Table 1: The top 10 most similar vectors for two queries: an English phrase, and a Farsi word. Recall that the index includes vectors for words in both languages, phrases in both languages, sentences in both languages, bilingual phrase pairs, and bilingual lexical pairs. Farsi words are indicated in *italics*.

The pipeline for adding our semantic feature to the phrase table is very straightforward. We have a set of vectors for the phrases of both languages. Each phrase is modelled with a 100-dimensional vector. The phrase table is scanned sequentially, and for each phrase pair, related vectors are fetched, then their similarity is estimated. To measure the similarities we use the cosine metric, which is:

$$similarity(v_1, v_2) = \frac{v_1.v_2}{||v_1|| * ||v_2||}$$

Values from the cosine similarity are in the range [-1,1].We map the similarity scores to the range [0,1] before adding them to the phrase table. After including the new feature, the weights for the log-linear model are learned by using held-out data.

4 Experimental Results

To study the impact of our new feature, we selected two datasets. One is the TEP++ corpus (Passban et al., 2015) which is a collection of 600K parallel En–Fa sentences. Farsi is a particularly interesting language for new MT research because it is both low resource and morphologically rich. The state-of-the-art MT quality for Farsi is not advanced relative to languages with

more training data available. The other dataset is the WMT[2] Fr-En corpus. The WMT corpora are frequently used as the de facto standard test datasets for SMT systems.

In all of the experiments Moses (Koehn et al., 2007) is used to build the SMT engines. BLEU (Papineni et al., 2002) is the evaluation metric and the feature weights are tuned by MERT (Och, 2003). All language models are built using SRILM (Stolcke et al., 2002). The improvements in translation quality are statistically significant according to the results of paired bootstrap re-sampling (Koehn, 2004).[3] The experimental setup is shown in the Table 2 and results from the baseline and extended systems are in Table 3.

Language pair	En–Fr	En–Fa
Dataset	Europarl corpus (Koehn, 2005) version 7 is used as a bilingual training set. As a dev set 2000 sentences of news-test2014 were used and the test set is the test set of the WMT2015 shared translation task.	The training, test and dev sets are subsets of the TEP++ corpus consisting of 575191, 2000 and 1000 sentences respectively. All the sentences have been selected randomly.
Language model	5-gram language models trained on the monolingual part of the training sets.	
MT engine	statistical phrase based engine with default Moses configuration	

Table 2: Experimental setup

As the results show, the new feature leads to improvements in all directions. We anticipated these improvements, as the positive impact of distributed phrase representation has already been shown by Gao et al. (2013). However, our work consider two new aspects of the problem. We train the vectors in a shared bilingual space, and show that proposed model can generate similar vectors for similar/equivalent constituents, even for languages pairs such as En-Fa, which are not typologically similar. We also show that a simple but efficient static feature can improve translation quality.

	En–Fr	Fr–En	Fa–En	En–Fa
Baseline	29.91	27.31	29.21	21.03
Extended	30.62	27.95	29.72	21.44
Improvement	+0.71	+0.64	+0.51	+0.41

Table 3: Results for base-line and extended systems

The new semantic similarity feature causes, on average, +0.56 enhancement in terms of BLEU for all of the directions of En↔Fr and En↔Fa, and as the size of training data increases the method provides even better performance. Improvements for the En–Fr pair demonstrate that achievements of the model are valid for large datasets, and improvements for the En–Fa pair show that the model can be used to translate distant language pairs. The word order and structure of the Farsi and English languages are quite different from each other, and Farsi is a morphologically rich language, making translation more difficult than for closely related language pairs such as En-Es.

[2]http://www.statmt.org/wmt15/translation-task.html

[3]We used ARK research group codes for statitical significance testing for 1000 samples with parameter of 0.05, http://www.ark.cs.cmu.edu/MT/

5 Conclusion and Future work

In this work we presented a new bilingual semantic similarity feature obtained from a neural network that is trained on a bilingual corpus, and computes the distributed representation of phrases in a shared semantic space. Each phrase is projected into a vector and the similarity of the vectors for each phrase pair is estimated. The similarity score for the phrase pair is added as a new phrase table feature, and the MT engine is tuned according to the default features in addition to new one. This augmentation of the information in the phrase table provides improvements in translation quality.

The method is quite straightforward and does not impose any significant overhead to the baseline SMT pipeline. Distributed vector representations preserve the semantic information of the constituents as well as their order and structural dependencies. The bilingual examples in the training data create dependencies between the equivalent constituents from different languages. As the model connects the phrases of two different languages to each other, it implicitly includes contextual information about the phrase pair into the MT process. Our next goal is to incorporate information from the source and target sides at decoding time. Although our model provides a global measure of the quality of a phrase pair, we cannot use the current framework to do tasks like disambiguation, because our features are static. We hope to incorporate the knowledge from paragraphs and text segments that the source and target phrases are extracted from, and compare this information to the context of the phrase at decoding time in order to provide a dynamic means of computing cross-lingual similarity.

Acknowledgement

We would like to thank the three anonymous reviewers and Rasul Kaljahi for their valuable comments. 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., and Ney, H. (2014). Vector space models for phrase-based machine translation. *Syntax, Semantics and Structure in Statistical Translation*, page 1.

Bahdanau, D., Cho, K., and Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. In *Proceedings of International Conference on Learning Representations*, San Diego, USA.

Banchs, R. E. and Costa-jussà, M. R. (2011). A semantic feature for statistical machine translation. In *Proceedings of the fifth workshop on syntax, semantics and structure in statistical translation*, pages 126–134, Portland, Oregon, USA.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022.

Costa-jussa, M., Gupta, P., Rosso, P., and 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. Association for Computational Linguistics*, Baltimore, Maryland, USA.

Costa-Jussà, M. R. and Banchs, R. E. (2011). A vector-space dynamic feature for phrase-based statistical machine translation. *Journal of Intelligent Information Systems*, 37(2):139–154.

Deerwester, S. C., Dumais, S. T., Landauer, T. K., Furnas, G. W., and Harshman, R. A. (1990). Indexing by latent semantic analysis. *JAsIs*, 41(6):391–407.

Devlin, J., Zbib, R., Huang, Z., Lamar, T., Schwartz, R., and 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*, volume 1, pages 1370–1380, Baltimore, USA.

Dinu, G. and Baroni, M. (2014). Improving zero-shot learning by mitigating the hubness problem. *arXiv preprint arXiv:1412.6568*.

Gao, J., He, X., Yih, W.-t., and Deng, L. (2013). Learning semantic representations for the phrase translation model. *arXiv preprint arXiv:1312.0482*.

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

Hinton, G. E. (1986). Learning distributed representations of concepts. In *Proceedings of the Eighth Annual Conference of the Cognitive Science Society*, Amherst, USA.

Huang, E. H., Socher, R., Manning, C. D., and 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*, pages 873–882. Association for Computational Linguistics.

Kalchbrenner, N. and Blunsom, P. (2013). Recurrent continuous translation models. In *EMNLP*, pages 1700–1709.

Kalchbrenner, N., Grefenstette, E., and Blunsom, P. (2014). A convolutional neural network for modelling sentences. In *Proceedings of Association for Computational Linguistics (ACL 2014)*, Baltimore, USA.

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

Koehn, P. (2005). Europarl: A parallel corpus for statistical machine translation. In *MT summit*, volume 5, pages 79–86.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., 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*, pages 177–180. Association for Computational Linguistics.

Le, Q. V. and Mikolov, T. (2014). Distributed representations of sentences and documents. In *Proceedings of ICML*, Beijing, China.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013a). Efficient estimation of word representations in vector space. In *ICLR Workshop*, Scottsdale, AZ, USA.

Mikolov, T., Le, Q. V., and Sutskever, I. (2013b). Exploiting similarities among languages for machine translation.

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*, pages 160–167, Sapporo, Japan.

Papineni, K., Roukos, S., Ward, T., and 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*, pages 311–318. Association for Computational Linguistics.

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

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

Sahlgren, M. (2005). An introduction to random indexing. In *In Methods and Applications of Semantic Indexing Workshop at the 7th International Conference on Terminology and Knowledge Engineering, TKE 2005*.

Salton, G., Wong, A., and Yang, C.-S. (1975). A vector space model for automatic indexing. *Communications of the ACM*, 18(11):613–620.

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

Wolf, L., Hanani, Y., Bar, K., and Dershowitz, N. (2014). Joint word2vec networks for bilingual semantic representations. *International Journal of Computational Linguistics and Applications*, 5(1):27–44.

Zhao, K., Hassan, H., and Auli, M. (2015). Learning translation models from monolingual continuous representations. In *Proceedings of the Ninth Workshop on Statistical Machine Translation. Association for Computational Linguistics*, Baltimore, Maryland, USA.