Learning Bilingual Distributed Phrase Representations for Statistical Machine Translation

Chaochao Wang chaochaowang@foxmail.com
Deyi Xiong [*] dyxiong@suda.edu.cn
Min Zhang minzhang@suda.edu.cn
Soochow University, Suzhou, China

Chunyu KIT ctckit@cityu.edu.hk
Department of Linguistics and Translation, City University of Hong Kong

Abstract

Following the idea of using distributed semantic representations to facilitate the computation of semantic similarity between translation equivalents, we propose a novel framework to learn bilingual distributed phrase representations for machine translation. We first induce vector representations for words in the source and target language respectively, in their own semantic space. These word vectors are then used to create phrase representations via composition methods. In order to compute semantic similarity of phrase pairs in the same semantic space, we project phrase representations from the source-side semantic space onto the target-side semantic space via a neural network that is able to conduct nonlinear transformation between the two spaces. We integrate the learned bilingual distributed phrase representations into a hierarchical phrase-based translation system to validate the effectiveness of our proposed framework. Experiment results show that our method is able to significantly improve translation quality and outperform previous methods that only use word representations or linear semantic space transformation.

1 Introduction

Distributional semantic models provide a means to represent meanings of words under the assumption that words occurring in similar contexts tend to have the same meanings (Harris,1968). Various distributed representations have been successfully applied to various monolingual natural language processing tasks, such as word sense discrimination (Clark and Pulman,2007) and thesaurus compilation (Yang and Powers,2008). In this paper, we explore how to learn semantic representations of bilingual phrases, rather than monolingual words, in the context of statistical machine translation (SMT) to facilitate the computation of semantic similarity between translation equivalents at the phrase level. We also study whether semantic similarity scores calculated in terms of bilingual distributed phrase representations are complementary to phrase translation probabilities estimated by the conventional counting method in SMT Koehn et al. (2003).

Very recently we have witnessed some studies on learning bilingual distributed representations for SMT. Mikolov et al. (2013b) train two neural network (NN) based models to learn word embeddings in the source and the target language, respectively, and then map the embeddings from the source to the target language space using a transformation matrix that is learned

[*] Corresponding author

by minimizing the mapping cost on all word pairs. Zou et al. (2013) introduce bilingual word embeddings into phrase-based machine translation: Word representations are first learned from language A via an NN-based model, word embeddings in the parallel language B are then initialized according to A's embeddings and word alignments between A and B, and the final word representations of B are obtained by a further training process that optimizes a combined objective on bilingual data. Gao et al. (2013) extend distributional representations from the word level to the phrase level, adopting a fully connected neural network to transfer bag-of-words vector representations of raw phrases (in the source or the target language) to distributional representations in a language-independent low-dimensional semantic space and having the parameters of the neural network jointly learned with the feature weights of the log-linear model of phrase-based SMT.

Partially inspired by these previous studies, we propose a new framework to learn bilingual distributed phrase representations for SMT via semantic composition and bilingual projection. Our current work differs from the previous ones distinctively in several aspects.

- We learn bilingual phrase representations, instead of representations at the word level Mikolov et al. (2013b); Zou et al. (2013), so as to keep consistency with the SMT that uses phrases rather than words as basic translation units.

- We learn phrase representations from distributed word representations via semantic composition, instead of from raw phrases Gao et al. (2013) in order to avoid the data sparseness issue of directly learning phrase representations from data. Particularly, we empirically compare two different composition methods in our framework, namely, weighted vector addition (Mitchell and Lapata,2008) and recursive autoencoder Socher et al. (2011).

- Rather than jointly learning phrase representations with feature weights of the log-linear model of SMT Gao et al. (2013), we take a loose coupling strategy to simplify the learning process. We adopt a neural network to project phrase representations from the source onto the target language semantic space, in separation from the process of feature weight tuning in SMT.

- Rather than capturing only the linear transformation between the source and target language semantic space Mikolov et al. (2013b), our neural network for the bilingual projection can model both linear and nonlinear transformation between these two semantic spaces. We hope that the nonlinear transformation can better model the bilingual projection.

We integrate the learned bilingual distributed phrase representations into a hierarchical phrase-based SMT system (Chiang,2007) by calculating semantic similarity scores of bilingual phrases in terms of their representations. We also empirically compare combinations of learning methods for word representations, and phrase composition methods as well as bilingual projection strategies. Experiments on Chinese-to-English translation show that our best results outperform the baseline by 0.53 BLEU points, indicating the effectiveness of our approach.

The rest of the paper is organized as follows. Section 2 describes how we obtain word representations for the source and target language and vector representations of phrases via two different composition methods. Section 3 presents a nonlinear neural network to learn bilingual phrase representations by projecting source language phrase vectors onto the target language space and Section 4 introduces the integration of bilingual representations into SMT. Section 5 elaborates our large-scale experiments on Chinese-to-English translation and analyzes experimental results. Section 6 discusses related work in relation to ours and Section 7 concludes the paper with future directions of research.

2 Distributed Phrase Representation Acquisition via Semantic Composition

In this section, we introduce semantics-based vector representations of words and vector representations of phrases via two different composition methods.

2.1 Word Representations

Vectors of words are basic elements in our bilingual phrase representation learning framework. We employ two different models, namely a *point-wise mutual information* (PMI) based vector space model (Pado and Lapata,2007) and a neural language model Mikolov et al. (2013a), to derive word vectors.

PMI-based vector space word representations Vector space model provides an elegant way to represent the meaning of a word: each element in its vector denotes a degree that measures how frequently it co-occurs in a predefined context window with every other word in the vocabulary in question. A well-known measure for this is PMI, which estimates the strength of the relationship between a context word c and a target word t as follows:

$$pmi(c,t) = \log \frac{p(c,t)}{p(c)p(t)} \tag{1}$$

In order to get around certain unavoidable frequency bias, we use *positive point-wise mutual information* (PPMI) (Turney and Pantel,2010) to calculate the elements in a word. It is defined as:

$$ppmi(c,t) = \begin{cases} pmi(c,t) & if \quad pmi(c,t) > 0 \\ 0 & otherwise \end{cases} \tag{2}$$

Neural word representations Mikolov et al.(2013a) introduce an efficient neural language model to learn high-quality word embeddings from extremely large amounts of raw texts. We adopt their approach for learning word embeddings. After training the neural language model, we can obtain a word embedding matrix $M \in R^{n\times|V|}$, where each word in the vocabulary V corresponds to a vector $v \in R^n$ with n to denote vector size. Given this, the vector representation of the word assigned with index i in V can be retrieved simply by extracting the i^{th} column of M.

2.2 Composition Methods

Once having obtained vector representations for words, we can use them to construct those for phrases via various composition methods as phrases are composed of words. We explore two composition methods: one based on simple vector addition (Mitchell and Lapata,2008) and the other on a recursive autoencoder that takes the inner structure of a phrase into account Socher et al. (2011).

Weighted vector addition Given a phrase p that consists of two words w_1 and w_2, we obtain the vector $\overrightarrow{p}$ from its word vectors $\overrightarrow{w_1}$ and $\overrightarrow{w_2}$ by the following weighted vector addition:

$$\overrightarrow{p} = \alpha\overrightarrow{w_1} + \beta\overrightarrow{w_2} \tag{3}$$

where α and β are weights denoting the relative importance of each word in the composition. For a phrase with multiple words p = (w_1,w_2,...,w_n), we can use in a similar way to obtain the vector for p by summing over vectors of all its words,

$$\overrightarrow{p} = \sum_{i=1}^{n} \lambda_i \overrightarrow{w_i} \tag{4}$$

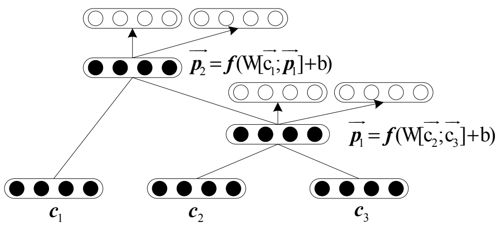
$\overrightarrow{p_2} = f(W[\overrightarrow{c_1}; \overrightarrow{p_1}]+b)$
$\overrightarrow{p_1} = f(W[\overrightarrow{c_2}; \overrightarrow{c_3}]+b)$
c_1
c_2
c_3

Figure 1: *The architecture of a recursive autoencoder, where the nodes with black dots are input word (or phrase) vectors and the nodes with circles are reconstructed vectors for computing reconstruction errors.*

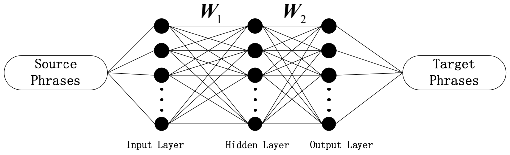

Figure 2: *The architecture of bilingual projection neural network that projects vector representations of source phrases to the semantic space of target language.*

Although weighted vector addition is a simple way for composition, it has proven effective in many tasks (Kartsaklis,2014). In our task, however, it cannot model word positions in a phrase. Therefore we only use it as a baseline to compare against a more advanced method: recursive autoencoder.

Recursive autoencoder (RAE) RAE is a neural network that can learn representations for large linguistic expressions such as phrases and sentences in a bottom-up fashion along a tree structure. Normally, word vectors learned via a distributional method can be input as leaf nodes of RAE. Figure 1 presents an illustration to visualize the architecture of RAE. Given a binary branch $p \rightarrow c_1 c_2$ where a child node is either a leaf or nonterminal node, the representation of p can be calculated as:

$$\overrightarrow{p} = f(W[\overrightarrow{c_1}; \overrightarrow{c_2}] + b) \tag{5}$$

where $[\overrightarrow{c_1}; \overrightarrow{c_2}]$ denotes the combination of the two child vectors, W and b are model parameters, and f is an element-wise activation function such as $sigmoid$. This will be used to further compute representations for larger structures. In order to judge how appropriately a parent vector computed this way can represent its children, we can reconstruct the children in a reconstruction layer as:

$$[\overrightarrow{c_1}'; \overrightarrow{c_2}'] = W' \overrightarrow{p} + b' \tag{6}$$

For each nonterminal node, we compute the Euclidean distance between its original child vectors and the reconstructed vectors as the reconstruction error of the node according to the following equation:

$$E_{rec}([\overrightarrow{c_1}; \overrightarrow{c_2}]) = \frac{1}{2}\left\|[\overrightarrow{c_1}; \overrightarrow{c_2}] - [\overrightarrow{c_1}'; \overrightarrow{c_2}']\right\|^2 \tag{7}$$

The parameters of an RAE can be learned by minimizing the reconstruction error over the entire tree.

In this paper, we adopt a greedy unsupervised RAE that is proposed in Socher et al. (2011) as an extension to the standard RAE described above. The unsupervised RAE can learn not only the representation of a phrase or sentence but also their tree structures in a greedy manner.

3 Learning Bilingual Phrase Representations

We use the methods introduced above to obtain phrase representations for the source and target language respectively, and adopt a nonlinear bilingual projection neural network to project the phrase representations in the source language onto the semantic vector space of the target language so as to calculate similarity scores of bilingual phrases in the same semantic space. The general architecture for this work is presented in Figure 2.

The adopted neural network for projection is a fully connected neural network with only one hidden layer. The projection can be formulated in the following equation:

$$\overrightarrow{p} = sigmoid(W_2(sigmoid(W_1\overrightarrow{x} + b_1)) + b_2) \tag{8}$$

where W_1 is the projection matrix from the input layer to the hidden layer, W_2 is the projection matrix from the hidden layer to the output layer, b_1 and b_2 are bias terms. In order to calculate the weights of the network, we need to calculate the squared error function as follows:

$$e = \frac{1}{2}\sum_i (t_i^m - p_i^m)^2 \tag{9}$$

where p_i^m is the vector calculated by the neural network according to the Eq.(8) and t_i^m the real vector of the corresponding target phrase. The weights can be trained via backpropagation by minimizing the error on the set of collected training instances $\{(\overrightarrow{s}, \overrightarrow{t})\}_1^n$ where $\overrightarrow{s}$ and $\overrightarrow{t}$ are vectors of the source and target side of a phrase pair (s,t). If we do not use any hidden layer in the projection neural network, the degenerated neural network will exactly learn Mikolov et al. (2013a)'s linear transformation matrix. Adding a hidden layer with nonlinear activation functions, we enable our projection neural network to model the nonlinear transformation between the semantic spaces of the source and target language. We will empirically compare the nonlinear against the linear projection in Section 5.Once the projection neural network is trained, we can learn projected representations of source phrases in the target semantic space using this neural network.

4 Integrating Bilingual Representations into SMT

A straightforward way to integrate bilingual phrase representations into SMT is to calculate semantic similarity between representations of this kind for translation equivalents. Given a phrase pair (s,t), let $(\overrightarrow{s}, \overrightarrow{t})$ denote their vector representations on the source and the target language semantic space and $(p(\overrightarrow{s}), \overrightarrow{t})$ the learned bilingual distributed phrase representations, where $p(\overrightarrow{s})$ is the projected vector representation of source phrase s obtained by our projection neural network as presented above. The semantic similarity between s and t can then be calculated as follows:

$$Sim(p(\overrightarrow{s}), \vec{t}) = \frac{p(\overrightarrow{s}) \bullet \vec{t}}{\|p(\overrightarrow{s})\| \times \|\vec{t}\|} \tag{10}$$

Given a source sentence c, we can build a new semantic similarity model based on our learned bilingual phrase representations according to the following equation:

$$M_{Sim} = \sum_{(s,t)\in P} Sim(p(\overrightarrow{s}), \overrightarrow{t}) \tag{11}$$

System	NIST 06	NIST 08
Baseline	30.23	23.21
PPMI	30.37	23.36
Neural	30.46	23.40

Table 1: Results of integrating word representations acquired by the two methods integrated into hierarchical phrase-based SMT

where P denotes all possible phrase pairs that are in use to translate the source sentence c and have distributed vector representations. The semantic similarity can be used as a feature in a log-linear model and can also be integrated into any SMT system that uses bilingual phrase pairs during decoding. In this paper, we integrate this new model into a hierarchical phrase-based SMT system without loss of generality.

Rules in the hierarchical phrase-based SMT can be classified into two types: 1) phrase rules that only contain terminals and 2) non-terminal rules with at least one non-terminal. For phrase rules, the similarity score can be easily calculated according to Eq. (10) in a preprocessing step. As for non-terminal rules, we compute their similarity scores in two steps. Let us take a specific non-terminal rule X $\rightarrow <$ X_1 选举 委员会 X_2 举行 选举,X_1 the election committee X_2 hold election,0-0 1-2 2-3 3-4 4-5 5-6 $>$ as an example to show how we compute their similarity. First, we can find phrase pairs (“选举 委员会”,“the election committee”) and (“举行 选举”,“hold election”) via word alignments. The similarity values of these two phrase pairs are estimated according to Eq. (10). In order to ensure decoding speed, the semantic similarities of these phrases are also calculated in a preprocessing step so that they can be quickly retrieved during decoding. Second, we sum up all similarity values, including the similarity values of phrases within nonterminals X_1 and X_2, according to Eq. (11) by means of dynamic programming.

5 Experiments

We have carried out a number of experiments on Chinese-to-English translation to validate the effectiveness of the proposed framework for learning bilingual phrase representations. Various combinations of different word representation models, composition methods, and projection strategies presented above are tested. Particularly, we intend to explore answers to the following questions:

- Which word representation is better, PMI-based vector space representation or neural representation?
- Would phrase representations be better than word representations when used to calculate semantic similarity scores? Furthermore, would RAE provide more efficient phrase representations than weighted vector addition?
- Is it necessary to project phrase representations in a non-linear fashion?

5.1 Experiment Setup

Our baseline system is hierarchical phrase-based system (Chiang,2007), where translation candidates are scored by a set of features. Our training data consists of 4.1M sentence pairs with 98.9M Chinese words and 112.6M English words from LDC corpora, including LDC 2003E07, LDC 2003E14, LDC 2004E12, LDC 2004T07, LDC 2005T06 and LDC 2005T10. We use the NIST evaluation set of 2005 (NIST 05) as development set, and sets of NIST 06/NIST 08 as our test sets.

System		NIST 06	NIST 08
Baseline		30.23	23.21
Word Representations		30.46	23.40
Phrase Representations	weighted vector addition	30.59	23.45^{+}
	RAE	30.76^{*}	23.51^{+}

Table 2: Results of integrating word and phrase representations into hierarchical phrase-based SMT with "+" and "*" to mark the statistically significant performance improvement over the baseline with $p < 0.05$ and $p < 0.01$

System	NIST 06	NIST 08
Baseline	30.23	23.21
Linear Projection	30.48	23.42^{+}
Nonlinear Projection	30.76^{*}	23.51^{+}

Table 3: Comparison of linear and nonlinear projection, with "+" and "*" to mark the statistically significant performance improvement over the baseline with $p < 0.05$ and $p < 0.01$.

Word alignments of training data were obtained by running GIZA++ (Och,2003b) in both directions of our bilingual language source and applying refinement rule grow-diag-final-and Koehn et al. (2003). A 4-gram language model was trained on the Xinhua section of Gigaword by SRILM toolkit Stolcke et al. (2002). We also extracted SCFG rules from the word-aligned training data. The translation performance was measured by case-insensitive BLEU Papineni et al. (2002). We used minimum error rate training (MERT) (Och.2003a) to tune the log-linear feature weights. As MERT is normally instable, we ran the tuning process three times for all our experiments and presented the average BLEU scores on the three MERT runs as suggested by Clark et al (2011).

The open source toolkit DISSECT[1] was applied to obtain PMI-based vector space word representations with a context window of 5 words, and Word2Vec[2] to acquire neural word representations, with each word represented as a 50-dimensional vector. When adopted Word2Vec, we just set the context window of size 5 and using continuous bag-of-words model. DISSECT was also adpoted to train weights in semantic composition of weighted vector addition. Unsupervised greedy RAE was trained in the way following Socher et al. (2011). In the bilingual projection neural network, 50 hidden units were used in the hidden layer.

5.2 PMI-Based Word Representations vs. Neural Word Representations

Our first series of experiments were carried out to compare PMI-based vector space word representations obtained by DISSECT against neural word representations obtained by Word2Vec. Note that we did not perform semantic composition in this series of experiments as we focus on word representations. Source word vector representations were projected onto the target semantic space via the projection neural network described in Section 3. Experimental results are presented in Table 1, from which we find both PMI-based vector space and neural word representations can improve translation quality in terms of BLEU. Since vector space representations obtained by neural word representations are better than PMI-based word representations, we used neural word representations in experiments hereafter.

[1] http://clic.cimec.unitn.it/composes/toolkit/index.html

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

Example 1	Source	兰州 物价局 就 牛肉面 限价 **作出 解释 ：** 只 因 涨幅 过 大 。
	Reference	Lanzhou Price Bureau **gives explanation of** price controls on beef noodles : It is only because the raises have been too large .
	Baseline	Lanzhou explained beef noodles reduce : only because of the excessive increase .
	NWR	Lanzhou explained that beef noodles reduce only because of excessive price .
	PRR	Lanzhou **gives explanation of** beef noodles reduce : only because of the excessive raises .
Example 2	Source	**高 收入 是** 许多 人 从事 小时 工 兼职 的 重要 原因 之一 。
	Reference	**High wages are** one of the major reasons for many people to get second jobs as hourly workers .
	Baseline	High income many people engage in hourly workers outside one of the major causes .
	NWR	High income is one major cause many people engage in hourly workers outside .
	PRR	**High wages are** one important reason many people engaged in hours for part-time workers .
Example 3	Source	禽流感 病毒 踪迹 **不断** 在 欧洲 **出现**， 造成 人心惶惶 。
	Reference	Panic strikes as signs of bird flu virus **continue to emerge** in Europe .
	Baseline	Traces of avian flu virus keeps on Europe, causing panic .
	NWR	Traces of avian flu virus continued there, causing panic in Europe .
	PRR	Traces of avian flu **continue to emerge**, causing panic in Europe .

Table 4: Translation examples to illustrate the advantage of RAE composition based phrase representations over others. NWR = neural word representations. PPR = phrase representations with RAE composition.

5.3 Comparison of Different Composition Methods

The second series of experiments were aimed at investigating whether we should learn semantic representations for phrases and at examining which composition method, i.e., either weighted vector addition or recursive autoencoder, is a better approach to create phrase vector representations from word representations. Table 2 presents the experimental results, from which we can draw the following observations:

- Integrating bilingual distributed phrase representations leads to a substantial improvement up to 0.53 BLEU points over the baseline.

- Phrase representations gave a better performance than word representations by up to 0.3 BLEU points.

- The RAE composition outperforms the weighted vector addition by up to 0.17 BLEU points.

5.4 Nonlinear vs. Linear Projection

As mentioned in Section 3, a linear variation of the bilingual projection can be derived by removing the hidden layer. In the third series of experiments, we investigated whether linear transformation is sufficient to project source language representations onto the target language semantic space. Our experimental results presented in Table 3 show that the nonlinear projection outperforms the linear one by 0.28 BLEU points. This suggests that the former is effective than the latter for transformation between the semantic spaces of the source and the target language even though they are learned by the same method from the same sets of data.

5.5 Translation Examples

Experimental results presented in the last four subsections show that the nonlinearly projected compositional phrase representations based on RAE give a better performance than the others. In this section, we examine a number of translation examples extracted from the test set, as presented in Table 4, to see the difference that the proposed method makes. These examples illustrate that the decoder equipped with the proposed semantic composition and bilingual nonlinear projection is able to select better translations for both continuous (Example 1 & 2) and non-continuous phrases (Example 3).

6 Related Work

Previous studies related to our research can be categorized into three strands as follows:

Distributed representations in monolingual settings Various methods are explored to learn distributed vector representations for words and phrases. Among them, vector space models are widely used, creating a vector to represent the co-occurrence relations between a target word and its contextual words (Bullinaria and Levy,2007; Pado and Lapata,2007). Topic models can be also used to construct distributed representations for words over topics that are learned from data Xiao et al. (2012). Recently, a variety of deep neural networks are applied to learn neural representations for both words and phrases in a continuous semantic space (Bengio et al.2003; Collobert and Weston,2008; Turian et al.2010; Socher et al.2012). All these methods can be used to create monolingual word representations for use in our framework to underlay the composition and projection operations.

Distributed representations for SMT In addition to the already mentioned three methods (Mikolov et al.2013b; Zou et al.2013a; Gao et al.2013) in the Introduction section, very recently Zhang et al. (2014) have proposed a bilingually-constrained recursive autoencoder in this strand, which extends the traditional semi-supervised recursive autoencoders Socher et al. (2011) to learn semantic phrase representations. They learn representations of one language with constraints from the counterpart language and share learned representations for phrases in the other language while we learn representations for the source and target language separately.

Semantic similarity models Our work is also related to semantic similarity models used in various NLP tasks. Bullinaria et al. (2007) carried out a word-based semantic similarity task to exam the degree of correlation between human judgments for two individual words and vector based similarity values. Xiao et al. (2012) introduced a topic similarity model to measure the similarity of translation rules to a document in terms of topics. We differ from them in that we calculate semantic similarity scores based on bilingual phrase representations learned via semantic composition and bilingual projection.

7 Conclusion and Future Work

We have presented a flexible framework above, which learns bilingual distributed phrase representations for machine translation. In this framework, vector representations of phrases are obtained by weighted vector addition or recursive autoencoder composition over words, which are represented as PMI-based vectors or continuous-valued vectors. We adopt a bilingual projection neural network to build nonlinear transformations between the source and the target language semantic space that are separately learned.

We integrate learned bilingual phrase representations into a hierarchical phrase-based SMT system. Our experimental results suggest the following:

- A semantic similarity model built on phrase representations is better than one built on word representations.
- Recursive autoencoder is superior to simple weighted vector addition in creating phrase vector representations from word vectors via composition.
- Nonlinear transformation is effective than linear transformation between the source and target language semantic space.

Our future work along this direction is to build stronger RAEs to construct vector representations for sentences.

Acknowledgment

The work was sponsored by the National Natural Science Foundation of China (Grants No. 61403269, 61432013 and 61333018) and Natural Science Foundation of Jiangsu Province (Grant No. BK20140355). We would like to thank three anonymous reviewers for their insightful comments.

References

Baroni, M.; and Zamparelli, R. 2010. Nouns are Vectors, adjectives are matrices: Representing adjective-noun constructions in Semantic space. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics*, 1183-1193.

Bullinaria, J. A.; and Levy, J. P. 2007. Extracting semantic representations from word co-occurrence statistics: A computational study. *Behavior research methods*, 39(3): 510-526.

Bengio, Y.; Ducharme, R.; Vincent, P.; Jauvin, C. 2006. Neural probabilistic language models. *Journal of Machine Learning Research*, 3:1137-1155.

Collobert, R.; and Weston, J. 2008. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th international conference on Machine Learning*, 160-167.

Chiang, D. 2007. Hierarchical phrase-based translation. *Computational Linguistics*, 33(2):201-228.

Clark, S.; and Pulman, S. 2007. Combining Symbolic and Distributional Models of Meaning. In *Proceedings of the Twenty-Second AAAI Conference on Artificial Intelligence (AAAI-07)*, 52-55. Spring Symposium: Quantum Interaction.

Clark, J.H.; Dyer, C.; Lavie, A.; and Smith, N. A. 2011. Better hypothesis testing for statistical machine translation: Controlling for optimizer instability. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 176-181.

Coecke, B.; Sadrzadeh, M.; and Clark, S. 2010. Mathematical Foundations for Distributed Compositional Model of Meaning. *Lambek Festschrift. Linguistic Analysis*, 36:345-384.

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

Harris, Z. 1968. Mathematical Structures of Language. *Wiley*.

Huang, P. S.; He, X.; Gao, J.; Deng, L.; Acero, A.; and Heck, L. 2013. Learning deep structured semantic models for web search using clickthrough data. In *Proceedings of the 22nd ACM international conference on Conference on information and knowledge management. ACM*, 2333-2338.

Kartsaklis, D. 2014. Compositional operators in distributional semantics. *Springer Science Reviews*, 1-17.

Koehn, P.; Och, F. J.; and Marcu, D. 2003. Statistical phrase-based translation. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, 58-54.

Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G.; and Dean, J.; 2013a. Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems*, 3111-3119.

Mikolov, T.; Le, Q. V.; and Sutskever, I. 2013b. Exploiting similarities among languages for machine translation. arXiv preprint arXiv:1309.4168.

Mitchell, J.; and Lapata, M.; 2008. Vector-based Models of Semantic Composition. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics*,236-244.

Och, F. J. 2003a. Minimum error rate training in statistical machine translation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, 160-167. Sapporo, Japan: Association for Computational Linguistics.

Och, F. J.; and Ney, H. 2003b. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19-51.

Pado', S.; and Lapata, M. 2007. Dependency-based Construction of Semantic Space Models. *Computational Linguistics*, 33(2): 161-199.

Papineni, K,; Roukos, S.; Ward, T.; and Zhu, W. J. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics*, 311-318.

Socher, R.; Pennington, J.; Huang, E. H.; Andrew, Y. Ng.; and Manning, D. C. 2011. Semi-supervised recursive autoencoders for predicting sentiment distributions. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics*, 151-161.

Socher, R.; Huval, B.; Manning, C. D.; and Andrew, Y. Ng. 2012. Semantic compositionality through recursive matrix-vector spaces. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. Association for Computational Linguistics*, 1201-1211.

Stolcke, A. 2002. Srilm-an extensible language modeling toolkit. In *proceedings of the 7th International Conference on Spoken Language Processing*, 901-904.

Turian, J.; Ratinov, L.; Bengio, Y. 2010. Word representations: a simple and general method for semi-supervised learning. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics*, 384-394.

Turney, P. D.; and Pantel, P. 2010. From frequency to meaning: Vector space models of semantics. *Journal of artificial intelligence research*, 37(1): 141-188.

Xiao, X.; Xiong, D.; Zhang, M.; Liu, Q.; Lin, S. 2012. A topic similarity model for hierarchical phrase-based translation. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1. Association for Computational Linguistics*, 750-758.

Yang,D.; and Powers, D. M. 2008. Automatic thesaurus construction. In *Proceedings of the Thirty-First Australasian conference on Computer science-Volume 74. Australian Computer Society*, Inc., 147-156.

Zou, W. Y.; Socher, R.; Cer, D.; and Manning, D. C. 2013. Bilingual Word Embeddings for Phrase-Based Machine Translation. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, 1393-1398.

Zhang, J.; Liu, S.; Li, M.; Zhou, M.; Zong, C. 2014. Bilingually-constrained phrase embeddings for machine translation. In *Proceedings of the 52th Annual Meeting on Association for Computational Linguistics. Association for Computational Linguistics.* Baltimore, Maryland, USA, 111-121.