Learning Monolingual Compositional Representations via Bilingual Supervision

Ahmed Elgohary and **Marine Carpuat**
Department of Computer Science
University of Maryland
College Park, MD 20742, USA
elgohary@cs.umd.edu, marine@cs.umd.edu

Abstract

Bilingual models that capture the semantics of sentences are typically only evaluated on cross-lingual transfer tasks such as cross-lingual document categorization or machine translation. In this work, we evaluate the quality of the monolingual representations learned with a variant of the bilingual compositional model of Hermann and Blunsom (2014), when viewing translations in a second language as a semantic annotation as the original language text. We show that compositional objectives based on phrase translation pairs outperform compositional objectives based on bilingual sentences and on monolingual paraphrases.

1 Introduction

The effectiveness of new representation learning methods for distributional word representations (Baroni et al., 2014) has brought renewed interest to the question of how to compose semantic representations of words to capture the semantics of phrases and sentences. These representations offer the promise of capturing phrasal or sentential semantics in a general fashion, and could in principle benefit any NLP applications that analyze text beyond the word level, and improve their ability to generalize beyond contexts seen in training.

While most prior work has focused either on composing words into short phrases (Mitchell and Lapata, 2010; Baroni and Zamparelli, 2010; Hermann et al., 2012; Fyshe et al., 2015), or on supervised task-specific composition functions (Socher et al., 2013; Iyyer et al., 2015; Rocktäschel et al., 2016; Iyyer et al., 2014; Tai et al., 2015, inter alia), Wieting et al. (2016) recently showed that a simple composition architecture (vector averaging) can yield sentence models that consistently perform well in semantic textual similarity tasks in a wide range of domains, and outperform more complex sequence models (Tai et al., 2015). Interestingly, these models are trained using PPDB, the paraphrase database (Ganitkevitch et al., 2013), which was learned from bilingual parallel corpora.

In bilingual settings, there are also a few examples of bilingual sentence models (Zou et al., 2013; Hermann and Blunsom, 2014; Lauly et al., 2014; Gouws et al., 2014). However, they have only been evaluated in cross-lingual transfer settings (e.g., cross-lingual document classification, or machine translation), which do not directly evaluate the quality of the sentence-level semantic representations learned.

In this work, we directly evaluate the usefulness of modeling semantic equivalence using compositional models of translated texts for detecting semantic textual similarity *in a single language*. For instance, in addition to using translated texts to model cross-lingual transfer from English to a foreign language, we can view English translations as a semantic annotation of the foreign text, and evaluate the usefulness of the resulting foreign representations. While learning representations in languages other than English is a pressing practical problem, this paper will focus on evaluating English sentence representations learned on English semantic similarity tasks to facilitate comparison with prior work.

Our results show that sentence representations learned using a bilingual compositional objective outperform representations learned using monolingual evidence, whether compositional or not. In addition, phrasal translations yield better representations than full sentence translations, even when applied to sentence-level tasks.

Table 1: Positive and negative examples for each of the 3 types of supervision considered

Bilingual Sentences	+	thus, in fact, we might say that he hurried ahead of the decision by our fellow member.	as que podramos decir , de hecho, que se adelant a la decisin de nuestro colega.
	-	thus, in fact, we might say that he hurried ahead of the decision by our fellow member.	seor presidente, la votacin sobre sellafield ha sido una novedad en el parlamento europeo .
English paraphrases	+	by our fellow member	by our colleague
	-	by our fellow member	of the committee's work
	+	slowly than anticipated	slowly than expected
Bilingual phrases	+	by our fellow member	de nuestro colega diputado
	-	by our fellow member	miles de personas de todo
	+	book and buy airline tickets	reserva y adquisicin de billetes
	+	the air fare advertised should show	el precio del billete anunciado debera indicar
	+	a book by the american writer noam	un libro del escritor norteamericano noam

2 Models

Inspired by the bilingual model of (Hermann and Blunsom, 2014), and paraphrase model of (Wieting et al., 2016), representations for multi-word segments are built with a simple bag-of-word additive combination of word representations, which are trained to minimize the distance between semantically equivalent segments.

2.1 Three Views of Semantic Equivalence

The different types of semantic equivalence used for training are illustrated in Table 1.

Parallel Sentences occur naturally, and provide training examples that are more consistent with downstream applications. However, they can be noisy due to automatic sentence alignment and one-to-many mappings, and bag-of-word representations of sentence meaning are likely to be increasingly noisier as segments get longer.

Monolingual Paraphrases are invaluable resources, but rarely occur naturally , and creating paraphrase resources therefore requires considerable effort. Ganitkevitch et al. (2013) automatically-created paraphrase resources for many languages using parallel corpora.

Parallel Phrases or phrasal translations might provide a tighter definition of semantic equivalence than longer sentence pairs, but phrase pairs have to be extracted automatically based on word alignments, an automatic and noisy process.

2.2 Models and Learning Objectives

Our main model is based on the bilingual composition model of Hermann and Blunsom (2014), which learns a word embedding matrix W from a training set $\mathcal{X}$ of aligned sentence pairs $\langle x_1, x_2 \rangle$. Each of x_1 and x_2 is represented as a bag-of-words, i.e. a superset of column indices in W. Each aligned pair $\langle x_1, x_2 \rangle$ is augmented with k randomly selected sentences that are not aligned to x_1, and another k that are not aligned to x_2. Given this augmented example $\langle x_1, x_2, \bar{x}_1^1, ..., \bar{x}_1^k, \bar{x}_2^1, ..., \bar{x}_2^k \rangle$, the model training objective is defined as follows:

$$
\begin{aligned}
J_{bi}(W) = \frac{\lambda}{2}||W||_F^2 + \sum_{\langle x_1, x_2, \bar{x}_1, \bar{x}_2 \rangle} \sum_{i=1}^{k} \\
[\delta + ||g(x_1) - g(x_2)||^2 \\
- ||g(x_1) - g(\bar{x}_2^i)||^2]_h \\
[\delta + ||g(x_1) - g(x_2)||^2 \\
- ||g(x_2) - g(\bar{x}_1^i)||^2]_h \qquad (1)
\end{aligned}
$$

where $g(x) = \sum_{i \in x} W_{:i}$, $[.]_h$ is the hinge function (i.e. $[v]_h = \max(0, v)$) whose margin is given by δ and λ is a regularization parameter.

The paraphrase-based model of Wieting et al. (2016) shares the same structure as the bilingual model above, but differs in the nature of segments used to define semantic equivalence (sentence pairs vs. paraphrases), the distance function used (Euclidean distance vs. cosine similarity), as well as the negative sampling strategies, and word embeddings initialization and regularization. We

Table 2: Training conditions: three types of semantic equivalence for composed representations.

Condition	# examples	Avg. length	Provenance
Bilingual Sentences	1.9M	28	Europarl-v7 (Koehn, 2005)
Bilingual phrases	3M	5	+ Moses phrase extraction (Koehn et al., 2007)
Monolingual phrases	3M	3	PPDB XL (Ganitkevitch et al., 2013)

provide empirical comparisons with the Wieting et al. (2016) embeddings, and also define a simplified version of that objective, J_{pa}, to allow for controlled comparisons with J_{bi}.

J_{pa} uses random initialization and penalizes large values in W with a $||W||_F^2$ regularization term[1]. The choice of distance function (Euclidean distance or cosine similarity) and of the negative sampling strategy[2] are viewed as tunable hyperparameters.

3 Experiments

3.1 Evaluating Sentence Representations

Following Wieting et al. (2016), the models above are evaluated on the four Semantic Textual Similarity (STS) datasets (Agirre et al., 2012; Agirre et al., 2013; Agirre et al., 2014; Agirre et al., 2015), which provide pairs of English sentences from different domains (e.g., Tweets, news, webforums, image captions), annotated with human judgments of similarity on a 1 to 5 scale. Systems have to output a similarity score for each pair. Systems are evaluated using the Pearson correlation between gold and predicted rankings.

The Sentences Involving Compositional Knowledge (SICK) test set (Marelli et al., 2014) provides a complementary evaluation. It consists of sentence pairs annotated with semantic relatedness scores. While STS examples were simply drawn from existing NLP datasets, SICK examples were constructed to avoid non-compositional phenomena such as multiword expressions, named entities and world knowledge.

3.2 Experimental Conditions

At training time we learn word embeddings for each combination of objective (Section 2.2) and type of training examples (Table 2), using modified implementations of open-source implementations for J_{bi} (Hermann and Blunsom, 2014) and J_{pa} (Wieting et al., 2016). This results in six model configurations. Each was trained for 10 epochs using tuned hyperparameters.

At tuning time we use the SMT-europarl subset of STS-2012. We consider mini-batch sizes of $\{25, 50, 100\}$, $\delta \in \{1, 10, 100\}$ with Euclidean distance, $\delta \in \{0.4, 0.6, 0.8\}$ with cosine similarly, and $\lambda \in \{1, 10^{-3}, 10^{-5}, 10^{-7}, 10^{-9}\}$. In J_{bi}, we consider $k \in \{1, 5, 10, 15\}$, and in J_{pa} we tuned over the sampling strategy $\in \{MIX, MAX\}$ and the distance function used. To speed up tuning for J_{pa}, we follow Wieting et al. (2016), by limiting training to $100k$ pairs, and tuning to 5 epochs.

Tuning results confirmed the importance of negative sampling and distance function in our models: in J_{bi}, increasing k consistently helps the bilingual models, whereas the correlation score for monolingual models degrade for $k > 10$. In J_{pa}, MAX always outperforms MIX. Euclidean distance was consistently chosen for bilingual sentences and monolingual phrases, while cosine similarity was chosen for bilingual phrases.

At test time we construct sentence-level embeddings by averaging the representations of words in each sentence, and compute cosine similarity to capture the similarity between sentences.

4 Findings

Table 3 reports the Pearson correlation scores achieved for each approach and dataset.

Bilingual phrases yield the best models in controlled settings

Overall, the best representations are obtained using bilingual phrase pairs and the J_{bi} objective. They outperform all other compositional models for all tasks, except for one subset of STS-2015.

The best objective for a given type of training example varies: J_{pa} generally yields better

[1] In contrast, Wieting et al. (2016) initialize W with high-quality but resource intensive embeddings – they are trained using word-level PPDB paraphrases, tuned on SimLex-999, and regularized to penalize deviations from initial GloVe embeddings (Pennington et al., 2014).

[2] *MAX* (use the unaligned phrase of minimum distance) or *MIX* (use *MAX* with probability 0.5 and sample randomly otherwise)

Table 3: Pearson correlation scores obtained on the English STS sets (with per year averages) and on semantic-relatedness task (SICK). The left columns report results based on new representations learned in this work, while the 2 rightmost columns report reference results from prior work (Wieting et al., 2016).

	Monolingual Phrases		Bilingual Phrases		Bilingual Sentences		Reference Results	
	J_{bi}	J_{pa}	J_{bi}	J_{pa}	J_{bi}	J_{pa}	Paragram	GloVe
MSRpar	0.28	0.42	**0.54**	0.38	**0.54**	0.36	0.44	0.47
MSRvid	0.33	0.55	**0.71**	0.38	0.71	0.19	0.77	0.64
SMT-eur	0.39	0.41	0.49	0.46	0.47	0.47	0.48	0.46
SMT-news	0.40	0.50	**0.59**	0.40	0.58	0.38	0.63	0.50
OnWN	0.52	0.57	**0.64**	0.62	0.46	0.62	0.71	0.55
2012 Avg	0.39	0.49	**0.59**	0.45	0.54	0.41	0.61	0.53
headline	0.56	0.66	**0.70**	0.58	0.66	0.61	0.74	0.64
OnWN	0.55	0.53	**0.75**	0.34	0.48	0.25	0.72	0.63
FNWN	0.35	0.29	**0.41**	0.32	0.25	0.16	0.47	0.34
2013 Avg	0.49	0.49	**0.62**	0.41	0.46	0.34	0.58	0.42
deft forum	0.35	0.47	**0.51**	0.36	0.36	0.33	0.53	0.27
deft news	0.59	0.68	**0.77**	0.59	0.76	0.58	0.75	0.68
headline	0.56	0.63	**0.73**	0.58	0.67	0.58	0.72	0.60
images	0.58	0.73	**0.73**	0.59	0.66	0.49	0.80	0.61
OnWN	0.65	0.62	**0.80**	0.55	0.55	0.47	0.81	0.58
tweet news	0.59	0.66	**0.73**	0.64	0.56	0.69	0.77	0.51
2014 Avg	0.55	0.63	**0.71**	0.55	0.59	0.52	0.73	0.54
forums	0.35	0.42	**0.55**	0.48	0.50	0.45	0.66	0.31
students	0.66	0.66	**0.73**	**0.73**	0.65	0.69	0.77	0.63
headline	0.64	0.60	**0.79**	0.64	0.73	0.66	0.76	0.62
belief	0.46	**0.71**	0.68	0.67	0.48	0.61	0.77	0.41
images	0.52	0.71	**0.75**	0.62	0.67	0.56	0.82	0.68
2015 Avg	0.53	0.63	**0.70**	0.63	0.59	0.60	0.76	0.53
SICK	0.53	0.62	**0.66**	0.57	0.63	0.54	0.72	0.66

results with monolingual phrases, while J_{bi} performs better with bilingual examples. Bilingual phrases seem to benefit from larger number of randomly selected negative samples and from using the Euclidean distance rather than cosine similarity. The best bilingual compositional representations are better than non-compositional Glove embeddings (Pennington et al., 2014), but worse than compositional Paragram embeddings (Wieting et al., 2016). However, Paragram initialization requires large amounts of text and human word similarity judgments for tuning, while our models were initialized randomly.

Table 4: Undertrained word ratios (tokens seen fewer than 100 times during training) are uncorrelated with performance in Table 3.

Dataset	Monolingual Phrases	Bilingual Phrases	Bilingual Sentences
2012 Avg	0.15	0.17	0.09
2013 Avg	0.16	0.17	0.11
2014 Avg	0.19	0.22	0.11
2015 Avg	0.15	0.19	0.11
SICK	0.2	0.25	0.15

Bilingual sentences vs. bilingual phrases

Why do bilingual phrases outperform the bilingual sentences they are extracted from? In this section, we verify that this is not explained by systematic biases in the distribution of training examples.

First, Table 4 shows that bilingual sentences have the smallest ratios of undertrained words, and are therefore not penalized by rare words more than bilingual phrases[3].

Second, we see that the rankings are not biased due to memorization of the phrases seen during training. Rankings of models does not change when testing on unseen word sequences, as shown by SICK results with models trained using J_{bi} on a filtered training set that contains none of the bigrams observed at test time (Table 5).

Third, the advantage of bilingual phrases over bilingual sentences is not due to the larger number of training examples. $1.9M$ (and even $1M$) bilin-

[3]Further, more than 80% of words that appear in both bilingual sentences and bilingual phrases occur in 460 (in average) more bilingual sentences than in bilingual phrases. The remaining 20% were found to be the rare words (e.g. zazvorkova, woldesmayat, yellow-bellies) that hardly occur in test sets.

Table 5: Impact of memorization: Pearson correlation scores on SICK with training sets with and without filtering out training pairs that contain any bigrams that appear in SICK. Number of training pairs (# Pairs) is shown in millions.

	Not Filtered		Filtered	
	# Pairs	Score	# Pairs	Score
Monoling. Phrases	3M	0.53	2.3M	0.54
Bilingual Phrases	3M	**0.66**	2.1M	**0.65**
Bilingual Sentences	1.9M	0.63	0.47M	0.58

Table 6: Impact of training set size: Average Pearson correlation per test set with different numbers (in millions) of bilingual phrase pairs, compared to the full set of bilingual sentences and monolingually pretrained GloVe.

	Bilingual Phrases				Sent.	
	$0.5M$	$1M$	$1.9M$	$3M$	*1.9M*	GloVe
2012	0.55	0.58	0.59	0.59	0.54	0.53
2013	0.59	0.61	0.61	0.62	0.46	0.42
2014	0.69	0.71	0.71	0.71	0.59	0.54
2015	0.68	0.69	0.70	0.70	0.59	0.53
SICK	0.62	0.64	0.65	0.66	0.63	0.66

gual phrase pairs still outperform the $1.9M$ bilingual sentence pairs on all subsets (See Table 6).

Taken together, these additional results support our initial intuition that the main advantage of bilingual phrases over bilingual sentences is that phrase pairs have stronger semantic equivalence than sentence pairs, since phrase pairs are shorter and are constructed by identifying strongly aligned subsets of sentence pairs.

Monolingual vs. bilingual phrases

Based on the analysis thus far, we hypothesize that paraphrase pairs with overlapping tokens make the compositional training objective less useful. Around 40% of the paraphrase training pairs differ only by one token. With Euclidean distance in the training objective, overlapping tokens cancel each other out of the composition term. For example, the pair $\langle$*healthy and stable, healthy and steady*$\rangle$ yields the compositional term

$$||(healthy + and + stable) - (healthy + and + steady)||_2 = ||stable - steady||_2$$

In contrast, overlap cannot occur in the bilingual setting, and all words within bilingual phrases contribute to the compositional objective. Furthermore, bilingual pairs provide a more explicit semantic signal as translations can disambiguate polysemous words (Diab, 2004; Carpuat and Wu, 2007) and help discover synonyms by pivoting (Callison-Burch, 2007; Yao et al., 2012).

All these factors might contribute to the ability of training with bilingual phrases of taking advantage of larger number of negative samples k.

5 Conclusion

We conducted the first evaluation of compositional representations learned using bilingual supervision on monolingual textual similarity tasks.

Phrase and sentence representations are constructed by composing word representations using a simple additive composition function. We considered two training objective that encourage the resulting representations to distinguish English-Spanish segment pairs that are semantically equivalent or not. The resulting English sentence representations consistently outperform compositional models trained to detect monolingual paraphrases on five different English semantic textual similarity tasks from SemEval.

Bilingual phrase pairs are consistently the best evidence of semantic equivalence in our experiments. They yield better results than the sentence pairs they are extracted from, despite the noise introduced by the automatic extraction process.

Furthermore the composed representations outperform non-compositional word representations derived from monolingual co-occurrence statistics. While sizes of monolingual vs. bilingual corpora are not directly comparable, it is remarkable that representations learned with only 500k bilingual phrase pairs outperform GloVe embeddings trained on 840B tokens.

Since our best models still underperform Paragram vectors, which require a more sophisticated initialization process, we will turn to improving our initialization strategies in future work. Nevertheless, current results provide further evidence of the usefulness of compositional text representations, even with a simple bag-of-word additive composition function, and of bilingual translation pairs as a strong signal of semantic equivalence.

References

Eneko Agirre, Mona Diab, Daniel Cer, and Aitor Gonzalez-Agirre. 2012. Semeval-2012 task 6: A pilot on semantic textual similarity. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation*, pages 385–393.

Eneko Agirre, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, and Weiwei Guo. 2013. sem 2013 shared task: Semantic textual similarity, including a pilot on typed-similarity. In *In* SEM 2013: The Second Joint Conference on Lexical and Computational Semantics.*

Eneko Agirre, Carmen Banea, Claire Cardie, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Weiwei Guo, Rada Mihalcea, German Rigau, and Janyce Wiebe. 2014. Semeval-2014 task 10: Multilingual semantic textual similarity. In *Proceedings of the 8th international workshop on semantic evaluation (SemEval 2014)*, pages 81–91.

Eneko Agirre, Carmen Banea, Claire Cardie, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Weiwei Guo, Inigo Lopez-Gazpioa, Montse Maritxalar, Rada Mihalcea, et al. 2015. Semeval-2015 task 2: Semantic textual similarity, english, spanish and pilot on interpretability. In *Proceedings of the 9th international workshop on semantic evaluation (SemEval 2015)*, pages 252–263.

Marco Baroni and Roberto Zamparelli. 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*, pages 1183–1193.

Marco Baroni, Georgiana Dinu, and Germán Kruszewski. 2014. Don't count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of the ACL*.

Chris Callison-Burch. 2007. *Paraphrasing and Translation*. Ph.D. thesis, University of Edinburgh.

Marine Carpuat and Dekai Wu. 2007. Improving Statistical Machine Translation using Word Sense Disambiguation. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL 2007)*, pages 61–72, Prague, June.

Mona Diab. 2004. Relieving the data acquisition bottleneck in word sense disambiguation. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics*.

Alona Fyshe, Leila Wehbe, Partha P Talukdar, Brian Murphy, and Tom M Mitchell. 2015. A compositional and interpretable semantic space. In *Proc. of NAACL*.

Juri Ganitkevitch, Benjamin Van Durme, and Chris Callison-Burch. 2013. Ppdb: The paraphrase database. In *HLT-NAACL*, pages 758–764.

Stephan Gouws, Yoshua Bengio, and Greg Corrado. 2014. Bilbowa: Fast bilingual distributed representations without word alignments. *arXiv preprint arXiv:1410.2455*.

Karl Moritz Hermann and Phil Blunsom. 2014. Multilingual models for compositional distributed semantics. In *Association for Computational Linguistics (ACL), 2014.*

Karl Moritz Hermann, Phil Blunsom, and Stephen Pulman. 2012. An unsupervised ranking model for noun-noun compositionality. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation*, pages 132–141.

Mohit Iyyer, Jordan L Boyd-Graber, Leonardo Max Batista Claudino, Richard Socher, and Hal Daumé III. 2014. A neural network for factoid question answering over paragraphs. In *EMNLP*, pages 633–644.

Mohit Iyyer, Varun Manjunatha, Jordan Boyd-Graber, and Hal Daumé III. 2015. Deep unordered composition rivals syntactic methods for text classification. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, volume 1, pages 1681–1691.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, 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.

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

Stanislas Lauly, Hugo Larochelle, Mitesh Khapra, Balaraman Ravindran, Vikas C Raykar, and Amrita Saha. 2014. An autoencoder approach to learning bilingual word representations. In *Advances in Neural Information Processing Systems*, pages 1853–1861.

Marco Marelli, Luisa Bentivogli, Marco Baroni, Raffaella Bernardi, Stefano Menini, and Roberto Zamparelli. 2014. Semeval-2014 task 1: Evaluation of

compositional distributional semantic models on full sentences through semantic relatedness and textual entailment. *SemEval-2014*.

Jeff Mitchell and Mirella Lapata. 2010. Composition in distributional models of semantics. *Cognitive science*, 34(8):1388–1429.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *EMNLP*, volume 14, pages 1532–1543.

Tim Rocktäschel, Edward Grefenstette, Karl Moritz Hermann, Tomáš Kočiskỳ, and Phil Blunsom. 2016. Reasoning about entailment with neural attention. In *In International Conference on Learning Representations (ICLR)*.

Richard Socher, Alex Perelygin, Jean Y Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the conference on empirical methods in natural language processing (EMNLP)*, volume 1631, page 1642.

Kai Sheng Tai, Richard Socher, and Christopher D Manning. 2015. Improved semantic representations from tree-structured long short-term memory networks. *arXiv preprint arXiv:1503.00075*.

John Wieting, Mohit Bansal, Kevin Gimpel, and Karen Livescu. 2016. Towards universal paraphrastic sentence embeddings. In *In International Conference on Learning Representations (ICLR)*.

Xuchen Yao, Benjamin Van Durme, and Chris Callison-Burch. 2012. Expectations of word sense in parallel corpora. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 621–625, Montréal, Canada, June. Association for Computational Linguistics.

Will Y Zou, Richard Socher, Daniel M Cer, and Christopher D Manning. 2013. Bilingual word embeddings for phrase-based machine translation. In *EMNLP*, pages 1393–1398.