Simple and Effective Paraphrastic Similarity from Parallel Translations

John Wieting[1], Kevin Gimpel[2], Graham Neubig[1], and Taylor Berg-Kirkpatrick[3]
[1]Carnegie Mellon University, Pittsburgh, PA, 15213, USA
[2]Toyota Technological Institute at Chicago, Chicago, IL, 60637, USA
[3]University of California San Diego, San Diego, CA, 92093, USA
{jwieting,gneubig}@cs.cmu.edu, kgimpel@ttic.edu, tberg@eng.ucsd.edu

Abstract

We present a model and methodology for learning paraphrastic sentence embeddings directly from bitext, removing the time-consuming intermediate step of creating paraphrase corpora. Further, we show that the resulting model can be applied to cross-lingual tasks where it both outperforms and is orders of magnitude faster than more complex state-of-the-art baselines.[1]

1 Introduction

Measuring sentence similarity is a core task in semantics (Cer et al., 2017), and prior work has achieved strong results by training similarity models on datasets of paraphrase pairs (Dolan et al., 2004). However, such datasets are not produced naturally at scale and therefore must be created either through costly manual annotation or by leveraging natural annotation in specific domains, like Simple English Wikipedia (Coster and Kauchak, 2011) or Twitter (Lan et al., 2017).

One of the most promising approaches for inducing paraphrase datasets is via manipulation of large bilingual corpora. Examples include bilingual pivoting over phrases (Callison-Burch et al., 2006; Ganitkevitch et al., 2013), and automatic translation of one side of the bitext (Wieting et al., 2017; Wieting and Gimpel, 2018; Hu et al., 2019). However, this is costly – Wieting and Gimpel (2018) report their large-scale database of sentential paraphrases required 10,000 GPU hours to generate.

In this paper, we propose a method that trains highly performant sentence embeddings (Pham et al., 2015; Hill et al., 2016; Pagliardini et al., 2017; McCann et al., 2017; Conneau et al., 2017) directly on bitext, obviating these intermediate steps and avoiding the noise and error propagation from automatic dataset preparation methods. This approach eases data collection, since bitext occurs naturally more often than paraphrase data and, further, has the additional benefit of creating cross-lingual representations that are useful for tasks such as mining or filtering parallel data and cross-lingual retrieval.

Most previous work for cross-lingual representations has focused on models based on encoders from neural machine translation (Espana-Bonet et al., 2017; Schwenk and Douze, 2017; Schwenk, 2018) or deep architectures using a contrastive loss (Grégoire and Langlais, 2018; Guo et al., 2018; Chidambaram et al., 2018). However, the paraphrastic sentence embedding literature has observed that simple models such as pooling word embeddings generalize significantly better than complex architectures (Wieting et al., 2016b). Here, we find a similar effect in the bilingual setting. We propose a simple model that not only produces state-of-the-art monolingual and bilingual sentence representations, but also encode sentences hundreds of times faster – an important factor when applying these representations for mining or filtering large amounts of bitext. Our approach forms the simplest method to date that is able to achieve state-of-the-art results on multiple monolingual and cross-lingual semantic textual similarity (STS) and parallel corpora mining tasks.[2]

Lastly, since bitext is available for so many language pairs, we analyze how the choice of language pair affects the performance of English paraphrastic representations, finding that using related languages yields the best results.

---

[1]Code and data to replicate results are available at https://www.cs.cmu.edu/~jwieting.

[2]In fact, we show that for monolingual similarity, we can devise random encoders that outperform some of this work.

2 Learning Sentence Embeddings

We first describe our objective function and then describe our encoder, in addition to several baseline encoders. The methodology proposed here borrows much from past work (Wieting and Gimpel, 2018; Guo et al., 2018; Grégoire and Langlais, 2018; Singla et al., 2018), but this specific combination has not been explored and, as we show in experiments, is surprisingly effective.

Training. The training data consists of a sequence of parallel sentence pairs (s_i, t_i) in source and target languages respectively. For each sentence pair, we randomly choose a *negative* target sentence t_i' during training that is not a translation of s_i. Our objective is to have source and target sentences be more similar than source and negative target examples by a margin δ:

$$\min_{\theta_{\mathrm{src}}, \theta_{\mathrm{tgt}}} \sum_i \Big[\delta - f_\theta(s_i, t_i) + f_\theta(s, t_i')) \Big]_+ .$$

The similarity function is defined as:

$$f_\theta(s, t) = \cos\Big(g(s; \theta_{\mathrm{src}}), g(t; \theta_{\mathrm{tgt}}) \Big)$$

where g is the sentence encoder with parameters for each language $\theta = (\theta_{\mathrm{src}}, \theta_{\mathrm{tgt}})$. To select t_i' we choose the most similar sentence in some set according to the current model parameters, i.e., the one with the highest cosine similarity.

Negative Sampling. The described objective can also be applied to monolingual paraphrase data, which we explore in experiments. The choice of negative examples differs whether we are using a monolingual parallel corpus or a bilingual parallel corpus. In the monolingual case, we select from all examples in the batch except the current pair. However, in the bilingual case, negative examples are only selected from the sentences in the batch from the opposing language. To select difficult negative examples that aid training, we use the *mega-batching* procedure of Wieting and Gimpel (2018), which aggregates M mini-batches to create one mega-batch and selects negative examples therefrom. Once each pair in the mega-batch has a negative example, the mega-batch is split back up into M mini-batches for training.

Encoders. Our primary sentence encoder simply averages the embeddings of subword units generated by `sentencepiece` (Kudo and Richardson, 2018); we refer to it as SP. This means that the sentence piece embeddings themselves are the only learned parameters of this model. As baselines we explore averaging character trigrams (TRIGRAM) (Wieting et al., 2016a) and words (WORD). SP provides a compromise between averaging words and character trigrams, combining the more distinct semantic units of words with the coverage of character trigrams.

We also use a bidirectional LSTM encoder (Hochreiter and Schmidhuber, 1997), with LSTM parameters fully shared between languages, as well as LSTM-SP, which uses sentence pieces instead of words as the input tokens. For all encoders, when the vocabularies of source and target languages overlap, the corresponding encoder embedding parameters are shared. As a result, languages pairs with more lexical overlap share more parameters.

We utilize several regularization methods (Wieting and Gimpel, 2017) including dropout (Srivastava et al., 2014) and shuffling the words in the sentence when training the LSTM-SP. Additionally, we find that annealing the mega-batch size by increasing it during training improved performance by a significant margin for LSTM-SP.

3 Experiments

Experiments are split into two groups. First, we compare training on parallel data to training on back-translated parallel data. We evaluate these models on the 2012-2016 SemEval Semantic Textual Similarity (STS) shared tasks (Agirre et al., 2012, 2013, 2014, 2015, 2016), which predict the degree to which sentences have the same meaning as measured by human judges. The evaluation metric is Pearson's r with the gold labels. We use the small STS English-English dataset from Cer et al. (2017) for model selection. Second, we compare our best model, SP, on two semantic cross-lingual tasks: the 2017 SemEval STS task (Cer et al., 2017) which consists of monolingual and cross-lingual datasets and the 2018 Building and Using Parallel Corpora (BUCC) shared bitext mining task (Zweigenbaum et al., 2018).

3.1 Hyperparameters and Optimization

Unless otherwise specified, we fix the hyperparameters in our model to the following: mega-batch size to 120, margin δ to 0.4, annealing rate to

Model	En-En	En-Cs (1M)	En-Cs (2M)
LSTM-SP (20k)	66.7	65.7	66.6
SP (20k)	68.3	**68.6**	**70.0**
WORD	66.0	63.8	65.9
TRIGRAM	**69.2**	**68.6**	69.9

Table 1: Comparison between training on 1 million examples from a backtranslated English-English corpus (En-En) and the original bitext corpus (En-Cs) sampling 1 million and 2 million sentence pairs (the latter equalizes the amount of English text with the En-En setting). Performance is the average Pearson's r over the 2012-2016 STS datasets.

150,[3] dropout to 0.3, shuffling rate for BLSTM-SP to 0.3, and the size of the SentencePiece vocabulary to 20,000. For WORD and TRIGRAM, we limited the vocabulary to the 200,000 most frequent types in the training data. We optimize our models using Adam (Kingma and Ba, 2014) with a learning rate of 0.001 and trained the models for 10 epochs.

3.2 Back-Translated Text vs. Parallel Text

We first compare sentence encoders and sentence embedding quality between models trained on backtranslated text and those trained on bitext directly. As our bitext, we use the Czeng1.6 English-Czech parallel corpus (Bojar et al., 2016). We compare it to training on ParaNMT (Wieting and Gimpel, 2018), a corpus of 50 million paraphrases obtained from automatically translating the Czech side of Czeng1.6 into English. We sample 1 million examples from ParaNMT and Czeng1.6 and evaluate on all 25 datasets from the STS tasks from 2012-2016. Since the models see two full English sentences for every example when training on ParaNMT, but only one when training on bitext, we also experiment with sampling twice the amount of bitext data to keep fixed the number of English training sentences.

Results in Table 1 show two observations. First, models trained on En-En, in contrast to those trained on En-CS, have higher correlation for all encoders except SP. However, when the same number of English sentences is used, models trained on bitext have greater than or equal performance across all encoders. Second, SP has the best performance in the En-CS setting. It also has fewer parameters and is therefore faster to train than LSTM-SP and TRIGRAM. Further, it

is much faster at encoding new sentences at test time.

3.3 Monolingual and Cross-Lingual Similarity

We evaluate on the cross-lingual STS tasks from SemEval 2017. This evaluation contains Arabic-Arabic (Ar), Arabic-English, Spanish-Spanish (Es), Spanish-English, and Turkish-English STS datsets. These datasets were created by translating one or both pairs of an English STS pair into Arabic, Spanish, or Turkish.

Baselines. We compare to several models from prior work (Guo et al., 2018; Chidambaram et al., 2018). A fair comparison to other models is difficult due to different training setups. Therefore, we perform a variety of experiments at different scales to demonstrate that even with much less data, our method has the best performance. In the case of Schwenk (2018), we replicate their setting in order to do a fair comparison.[4]

As another baseline, we analyze the performance of averaging randomly initialized embeddings. We experiment with SP having SentencePiece vocabulary sizes of 20,000 and 40,000 as well as TRIGRAM. The embeddings have 300 dimensions and are initialized from a normal distribution with mean 0 and variance 1.

Results. The results are shown in Table 2. We make several observations. The first is that the 1024 dimension SP model has the top performance on 4 of the 6 STS datasets. Our results also show that performance increases substantially from using 1 million to using 10 million sentence pairs and is also correlated positively with dimension and size of the SP vocabulary. This result outperforms the baselines from the literature as well, all of which use deep architectures.[5] Our SP model trained on Europarl (EP) also surpasses the model from Schwenk (2018) which is trained on the same corpus. Since that model is based on many-to-many translation, Schwenk (2018) trains

---

[3] Annealing rate is the number of minibatches that are processed before the megabatch size is increased by 1.

[4] Two follow-up papers (Artetxe and Schwenk, 2018a,b) use essentially the same underlying model, but we compare to Schwenk (2018) because it was the only one of these papers where the model has been made available when this paper was written.

[5] Including a 3-layer transformer trained on a constructed parallel corpus (Chidambaram et al., 2018), a bidirectional gated recurrent unit network trained on a collection of parallel corpora using En-Es, En-Ar, and Ar-Es bitext (Espana-Bonet et al., 2017), and a 3 layer bidirectional LSTM trained on 9 languages in Europarl (Schwenk, 2018).

Model	Data	N	Dim.	Ar-Ar	Ar-En	Es-Es	Es-En	En-En	Tr-En
Random TRIGRAM	OS	1M	300	67.9	1.8	77.3	2.8	73.7	18.6
Random SP (20k)	OS	1M	300	61.9	17.5	68.8	6.5	67.0	23.1
Random SP (40k)	OS	1M	300	58.3	16.1	68.2	10.4	66.6	22.2
SP (20k)	OS	1M	300	70.0	64.9	82.0	71.6	82.2	68.9
SP (20k)	OS	10M	300	72.4	74.9	85.0	77.0	83.3	74.6
SP (20k)	OS	10M	1024	75.4	73.2	85.0	**78.2**	83.2	75.7
SP (80k)	OS	10M	1024	**76.3**	**75.0**	**86.2**	**78.2**	**84.0**	**77.5**
SP (20k)	EP	2M	300	-	-	78.6	53.5	78.9	-
SP (20k)	EP	2M	1024	-	-	80.6	54.0	80.0	-
Schwenk (2018)	EP	18M	1024	-	-	64.4	40.8	66.0	-
Espana-Bonet et al. (2017)	MIX	32.8M	2048	59	44	78	49	76	-
Chidambaram et al. (2018)	MIX	470M/500M	512	-	-	64.2	58.7	-	-
2017 STS 1st Place	-	-	-	**75.4**	**74.9**	**85.6**	**83.0**	**85.5**	**77.1**
2017 STS 2nd Place	-	-	-	75.4	71.3	85.0	81.3	85.4	74.2
2017 STS 3rd Place	-	-	-	74.6	70.0	84.9	79.1	85.4	73.6

Table 2: Comparison of our models with those in the literature and random encoder baselines. All models are evaluated on cross-lingual STS where performance is measured in Pearson's r (%). N refers to the number of examples in the training data. OS stands for OpenSubtitles, EP for Europarl, and MIX for a variety of domains.

on nine (related) languages in Europarl. We only train on the splits of interest (En-Es for STS and En-De/En-Fr for the BUCC tasks) in our experiments

Secondly, random encoders, especially random TRIGRAM, perform strongly in the monolingual setting. In fact, the random encoders are competitive or outperform all three models from the literature in these cases. For cross-lingual similarity, however, random encoders lag behind because they are essentially measuring the lexical overlap in the two sentences and there is little lexical overlap in the cross-lingual setting, especially for distantly related languages like Arabic and English.

3.4 Mining Bitext

Lastly, we evaluate on the BUCC shared task on mining bitext. The task consists of finding the gold *aligned* parallel sentences given two large corpora in two distinct languages. Typically, only about 2.5% of the sentences are aligned. Following Schwenk (2018), we train our models on Europarl and evaluate on the publicly available BUCC data.

Results in Table 3 on the French and German mining tasks demonstrate the proposed model outperforms Schwenk (2018), although the gap is substantially smaller than on the STS tasks. The reason for this is likely the domain mismatch between the STS data (image captions) and the training data (Europarl). We suspect that the deep NMT encoders of Schwenk (2018) overfit to the domain more than the simpler SP model, and the BUCC task uses news data which is closer to Europarl than image captions.

Model	En-De	En-Fr
Schwenk (2018)	76.1	74.9
SP (20k)	76.9	76.3
SP (40k)	**77.5**	**76.8**

Table 3: F1 scores for bitext mining on BUCC.

Model	Dim	Sentences/Sec.
Schwenk (2018)	1024	2,601
Chidambaram et al. (2018)	512	3,049
SP (20k)	300	855,571
SP (20k)	1024	683,204

Table 4: A comparison of encoding times for our model compared to two models from prior work.

4 Analysis

We next conduct experiments on encoding speed and analyze the effect of language choice.

4.1 Encoding Speed

In addition to outperforming more complex models (Schwenk, 2018; Chidambaram et al., 2018), the simple SP models proposed here are much faster at encoding sentences. Since implementations to encode sentences are publicly available several baselines, we are able to test their encoding speed and compare. To do so, we randomly select 128,000 English sentences from the English-Spanish Europarl corpus. We then encode these sentences in batches of 128 on an Nvidia Quadro GP100 GPU. The number of sentences encoded per second is shown in Table 4, showing that SP is hundreds of times faster.
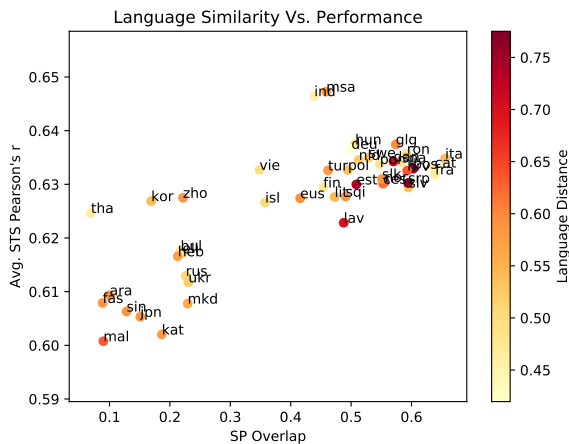
Figure 1: Plot of average performance on the 2012-2016 STS tasks compared to SP overlap and language distance as defined by Littell et al. (2017).

4.2 Does Language Choice Matter?

We next investigate the impact of the non-English language in the bitext when training English paraphrastic sentence embeddings. We took all 46 languages with at least 100k parallel sentence pairs in the 2016 OpenSubtitles Corpus (Lison and Tiedemann, 2016) and made a plot of their average STS performance on the 2012-2016 English datasets compared to their SP overlap[6] and language distance.[7] We segmented the languages separately and trained the models for 10 epochs using the 2017 STS task for model selection.

The plot, shown in Figure 1, shows that SentencePieces (SP) overlap is highly correlated with STS score. There are also two clusters in the plot, languages that have a similar alphabet to English and those that do not. In each cluster we find that performance is negatively correlated with language distance. Therefore, languages similar to English yield better performance. The Spearman's correlations (multiplied by 100) for all languages and these two clusters are shown in Table 5. When choosing a language to pair up with English for learning paraphrastic embeddings, ideally there will be a lot of SP overlap. However, beyond or below a certain threshold (approximately 0.3 judging by the plot), the linguistic distance to English is more predictive of performance. Of the

---

[6]We define SP overlap as the percentage of SPs in the English corpus that also appear in the non-English corpus.

[7]We used the feature distance in URIEL (Littell et al., 2017) which accounts for a number of factors when calculating distance like phylogeny, geography, syntax, and phonology.

Model	SP Ovl.	Lang. Distance
All Lang.	71.5	-22.8
Lang. (SP Ovl. ≤ 0.3)	23.6	-63.8
Lang. (SP Ovl. > 0.3)	18.5	-34.2

Table 5: Spearman's $\rho \times 100$ between average performance on the 2012-2016 STS tasks compared to SP overlap (SP Ovl.) and language distance as defined by Littell et al. (2017). We included correlations for all languages as well as those with low and high SP overlap with English.

factors in URIEL, syntactic distance was the feature most correlated with STS performance in the two clusters with correlations of -56.1 and -29.0 for the low and high overlap clusters respectively. This indicates that languages with similar syntax to English helped performance. One hypothesis to explain this relationship is that translation quality is higher for related languages, especially if the languages have the same syntax, resulting in a cleaner training signal.

We also hypothesize that having high SP overlap is correlated with improved performance because the English SP embeddings are being updated more frequently during training. To investigate the effect, we again learned segmentations separately for both languages then prefixed all tokens in the non-English text with a marker to ensure that there would be no shared parameters between the two languages. Results showed that SP overlap was still correlated (correlation of 24.9) and language distance was still negatively correlated with performance albeit significantly less so at -10.1. Of all the linguistic features, again the syntactic distance was the highest correlated at -37.5.

5 Conclusion

We have shown that using automatic dataset preparation methods such as pivoting or back-translation are not needed to create higher performing sentence embeddings. Moreover by using the bitext directly, our approach also produces strong paraphrastic cross-lingual representations as a byproduct. Our approach is much faster than comparable methods and yields stronger performance on cross-lingual and monolingual semantic similarity and cross-lingual bitext mining tasks.

References

Eneko Agirre, Carmen Banea, Claire Cardie, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Weiwei Guo, Inigo Lopez-Gazpio, Montse Maritxalar, Rada Mihalcea, German Rigau, Larraitz Uria, and Janyce Wiebe. 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)*.

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)*.

Eneko Agirre, Carmen Banea, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Rada Mihalcea, German Rigau, and Janyce Wiebe. 2016. SemEval-2016 task 1: Semantic textual similarity, monolingual and cross-lingual evaluation. *Proceedings of SemEval*, pages 497–511.

Eneko Agirre, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, and Weiwei Guo. 2013. * sem 2013 shared task: Semantic textual similarity. In *Second Joint Conference on Lexical and Computational Semantics (* SEM), Volume 1: Proceedings of the Main Conference and the Shared Task: Semantic Textual Similarity*, volume 1, pages 32–43.

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*. Association for Computational Linguistics.

Mikel Artetxe and Holger Schwenk. 2018a. Margin-based parallel corpus mining with multilingual sentence embeddings. *arXiv preprint arXiv:1811.01136*.

Mikel Artetxe and Holger Schwenk. 2018b. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *arXiv preprint arXiv:1812.10464*.

Ondřej Bojar, Ondřej Dušek, Tom Kocmi, Jindřich Libovický, Michal Novák, Martin Popel, Roman Sudarikov, and Dušan Variš. 2016. CzEng 1.6: Enlarged Czech-English Parallel Corpus with Processing Tools Dockered. In *Text, Speech, and Dialogue: 19th International Conference, TSD 2016*, number 9924 in Lecture Notes in Computer Science, pages 231–238, Cham / Heidelberg / New York / Dordrecht / London. Masaryk University, Springer International Publishing.

Chris Callison-Burch, Philipp Koehn, and Miles Osborne. 2006. Improved statistical machine translation using paraphrases. In *Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, pages 17–24, New York City, USA.

Daniel Cer, Mona Diab, Eneko Agirre, Inigo Lopez-Gazpio, and Lucia Specia. 2017. SemEval-2017 Task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 1–14, Vancouver, Canada.

Muthuraman Chidambaram, Yinfei Yang, Daniel Cer, Steve Yuan, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. 2018. Learning cross-lingual sentence representations via a multi-task dual-encoder model. *arXiv preprint arXiv:1810.12836*.

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. 2017. Supervised learning of universal sentence representations from natural language inference data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 670–680, Copenhagen, Denmark.

William Coster and David Kauchak. 2011. Simple English Wikipedia: a new text simplification task. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2*, pages 665–669.

Bill Dolan, Chris Quirk, and Chris Brockett. 2004. Unsupervised construction of large paraphrase corpora: Exploiting massively parallel news sources. In *Proceedings of COLING*.

Cristina Espana-Bonet, Adám Csaba Varga, Alberto Barrón-Cedeño, and Josef van Genabith. 2017. An empirical analysis of nmt-derived interlingual embeddings and their use in parallel sentence identification. *IEEE Journal of Selected Topics in Signal Processing*, 11(8):1340–1350.

Juri Ganitkevitch, Benjamin Van Durme, and Chris Callison-Burch. 2013. PPDB: The Paraphrase Database. In *Proceedings of HLT-NAACL*.

Francis Grégoire and Philippe Langlais. 2018. Extracting parallel sentences with bidirectional recurrent neural networks to improve machine translation. *arXiv preprint arXiv:1806.05559*.

Mandy Guo, Qinlan Shen, Yinfei Yang, Heming Ge, Daniel Cer, Gustavo Hernandez Abrego, Keith Stevens, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. 2018. Effective parallel corpus mining using bilingual sentence embeddings. *arXiv preprint arXiv:1807.11906*.

Felix Hill, Kyunghyun Cho, and Anna Korhonen. 2016. Learning distributed representations of sentences from unlabelled data. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8).

J Edward Hu, Rachel Rudinger, Matt Post, and Benjamin Van Durme. 2019. ParaBank: Monolingual bitext generation and sentential paraphrasing via lexically-constrained neural machine translation. In *Proceedings of AAAI*.

Diederik Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Taku Kudo and John Richardson. 2018. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. *arXiv preprint arXiv:1808.06226*.

Wuwei Lan, Siyu Qiu, Hua He, and Wei Xu. 2017. A continuously growing dataset of sentential paraphrases. In *Proceedings of EMNLP*, Copenhagen, Denmark. Association for Computational Linguistics.

Pierre Lison and Jörg Tiedemann. 2016. Opensubtitles2016: Extracting large parallel corpora from movie and tv subtitles.

Patrick Littell, David R. Mortensen, Ke Lin, Katherine Kairis, Carlisle Turner, and Lori Levin. 2017. Uriel and lang2vec: Representing languages as typological, geographical, and phylogenetic vectors. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 8–14. Association for Computational Linguistics.

Bryan McCann, James Bradbury, Caiming Xiong, and Richard Socher. 2017. Learned in translation: Contextualized word vectors. In *Advances in Neural Information Processing Systems*, pages 6297–6308.

Matteo Pagliardini, Prakhar Gupta, and Martin Jaggi. 2017. Unsupervised learning of sentence embeddings using compositional n-gram features. *arXiv preprint arXiv:1703.02507*.

Nghia The Pham, Germán Kruszewski, Angeliki Lazaridou, and Marco Baroni. 2015. Jointly optimizing word representations for lexical and sentential tasks with the c-phrase model. 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: Long Papers)*.

Holger Schwenk. 2018. Filtering and mining parallel data in a joint multilingual space. *arXiv preprint arXiv:1805.09822*.

Holger Schwenk and Matthijs Douze. 2017. Learning joint multilingual sentence representations with neural machine translation. *arXiv preprint arXiv:1704.04154*.

Karan Singla, Dogan Can, and Shrikanth Narayanan. 2018. A multi-task approach to learning multilingual representations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 214–220.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: A simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1).

John Wieting, Mohit Bansal, Kevin Gimpel, and Karen Livescu. 2016a. Charagram: Embedding words and sentences via character n-grams. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1504–1515.

John Wieting, Mohit Bansal, Kevin Gimpel, and Karen Livescu. 2016b. Towards universal paraphrastic sentence embeddings. In *Proceedings of the International Conference on Learning Representations*.

John Wieting and Kevin Gimpel. 2017. Revisiting recurrent networks for paraphrastic sentence embeddings. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2078–2088, Vancouver, Canada.

John Wieting and Kevin Gimpel. 2018. ParaNMT-50M: Pushing the limits of paraphrastic sentence embeddings with millions of machine translations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 451–462. Association for Computational Linguistics.

John Wieting, Jonathan Mallinson, and Kevin Gimpel. 2017. Learning paraphrastic sentence embeddings from back-translated bitext. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 274–285, Copenhagen, Denmark.

Pierre Zweigenbaum, Serge Sharoff, and Reinhard Rapp. 2018. Overview of the third bucc shared task: Spotting parallel sentences in comparable corpora. In *Proceedings of 11th Workshop on Building and Using Comparable Corpora*, pages 39–42.