Retrofitting Contextualized Word Embeddings with Paraphrases

Weijia Shi[1]*, **Muhao Chen**[1,2]*, **Pei Zhou**[1,3] **and Kai-Wei Chang**[1]
[1]Department of Computer Science, University of California Los Angeles
[2]Department of Computer and Information Science, University of Pennsylvania
[3]Department of Computer Science, University of Southern California
{swj0419, kwchang}@cs.ucla.edu; muhao@seas.upenn.edu; peiz@usc.edu

Abstract

Contextualized word embedding models, such as ELMo, generate meaningful representations of words and their context. These models have been shown to have a great impact on downstream applications. However, in many cases, the contextualized embedding of a word changes drastically when the context is paraphrased. As a result, the downstream model is not robust to paraphrasing and other linguistic variations. To enhance the stability of contextualized word embedding models, we propose an approach to retrofitting contextualized embedding models with paraphrase contexts. Our method learns an orthogonal transformation on the input space, which seeks to minimize the variance of word representations on paraphrased contexts. Experiments show that the retrofitted model significantly outperforms the original ELMo on various sentence classification and language inference tasks.

1 Introduction

Contextualized word embeddings have shown to be useful for a variety of downstream tasks (Peters et al., 2018, 2017; McCann et al., 2017). Unlike traditional word embeddings that represent words with fixed vectors, these embedding models encode both words and their contexts and generate context-specific representations. While contextualized embeddings are useful, we observe that a language model-based embedding model, ELMo (Peters et al., 2018), cannot accurately capture the semantic equivalence of contexts. Specifically, in cases where the contexts of a word have equivalent or similar meanings but are changed in sentence formation or word order, ELMo may assign very different representations to the word. Table 1 shows two examples, where ELMo generates very different representations for the boldfaced words under semantic equivalent contexts.

Paraphrased contexts	L2	Cosine
How can I make **bigger** my arms? How do I make my arms **bigger**?	6.42	0.27
Some people believe earth is **flat**. Why? Why do people still believe in **flat** earth?	7.59	0.46
It is a very **small** window. I have a **large** suitcase.	5.44	0.26

Table 1: L2 and Cosine distances between embeddings of boldfaced words. The distance between the shared word in the paraphrases is even greater than the distance between *large* and *small* in random contexts.

Quantitatively, 28.3% of the shared words in the paraphrase sentence pairs on the MRPC corpus (Dolan et al., 2004) is larger than the average distance between *good* and *bad* in random contexts, and 41.5% of those exceeds the distance between *large* and *small*. As a result, the downstream model is not robust to paraphrasing and the performance is hindered.

Infusing the model with the ability to capture the semantic equivalence no doubt benefits semantic-oriented downstream tasks. Yet, finding an effective solution presents key challenges. First, the solution inevitably requires the embedding model to effectively identify paraphrased contexts. On top of that, the model needs to minimize the difference of a word's representations on paraphrased contexts, without compromising the varying representations on unrelated contexts. Moreover, the long training time prevents us from redesigning the learning objectives of contextualized embeddings and retraining the model.

To address these challenges, we propose a simple and effective paraphrase-aware retrofitting (PAR) method that is applicable to arbitrary pretrained contextualized embeddings. In particular, PAR prepends an orthogonal transformation layer to a contextualized embedding model. Without retraining the parameters of an existing model, PAR learns the transformation to minimize the differ-

* Both authors contributed equally to this work.

Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, pages 1198–1203,
Hong Kong, China, November 3–7, 2019.

ence of the contextualized representations of the shared word in paraphrased contexts, while differentiating between those in other contexts. We apply PAR to retrofit ELMo (Peters et al., 2018) and show that the resulted embeddings provide more robust contextualized word representations as desired, which further lead to significant improvements on various sentence classification and inference tasks.

2 Related Work

Contextualized word embedding models have been studied by a series of recent research efforts, where different types of pre-trained language models are employed to capture the context information. CoVe (McCann et al., 2017) trains a neural machine translation model and extracts representations of input sentences from the source language encoder. ELMo (Peters et al., 2018) pre-trains LSTM-based language models from both directions and combines the vectors to construct contextualized word representations. Recent studies substitute LSTMs with Transformers (Radford et al., 2018, 2019; Devlin et al., 2019). As shown in these studies, contextualized word embeddings perform well on downstream tasks at the cost of extensive parameter complexity and the long training process on large corpora (Strubell et al., 2019).

Retrofitting methods have been used to incorporate semantic knowledge from external resources into word embeddings (Faruqui et al., 2015; Yu et al., 2016; Glavaš and Vulić, 2018). These techniques are shown to improve the characterization of word relatedness and the compositionality of word representations. To the best of our knowledge, none of the previous approaches has been applied in contextualized word embeddings.

3 Paraphrase-Aware Retrofitting

Our method, illustrated in Figure 1, integrates the constraint of the paraphrased context into the contextualized word embeddings by learning the orthogonal transformation on the input space.

3.1 Contextualized Word Embeddings

We use $S = (w_1, w_2, \cdots, w_l)$ to denote a sequence of words of length l, where each word w belongs to the vocabulary V. We use boldfaced $\mathbf{w} \in \mathbb{R}^k$ to denote a k-dimensional input word embedding, which can be pre-trained or derived from a character-level encoder (e.g., the character-level

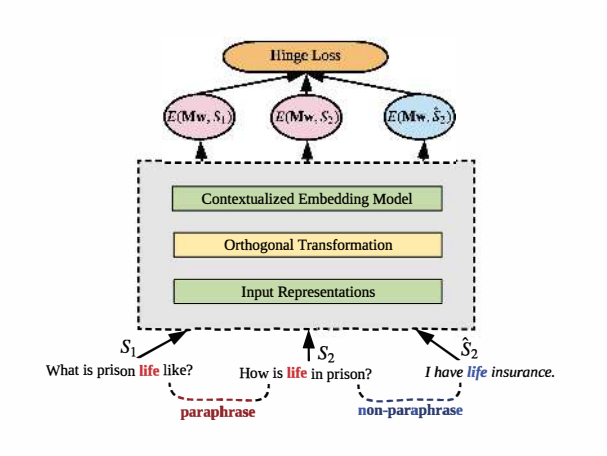

Figure 1: Learning framework of PAR

CNN used in ELMo (Peters et al., 2018)). A contextualized embedding model E takes input vectors of the words in S, and computes the context-specific representation of each word. The representation of word w specific to the context S is denoted as $E(\mathbf{w}, S)$.

3.2 Paraphrase-aware Retrofitting

PAR learns an orthogonal transformation $\mathbf{M} \in \mathbb{R}^{k \times k}$ to reshape the input representation into a specific space, where the contextualized embedding vectors of a word in paraphrased contexts are collocated, while those in unrelated contexts are differentiated. Specifically, given two contexts S_1 and S_2 that both contain a shared word w, the contextual difference of a input representation $\mathbf{w}$ is defined by the $L2$ distance,

$$d_{S_1,S_2}(\mathbf{w}) = \|E(\mathbf{w}, S_1) - E(\mathbf{w}, S_2)\|_2 .$$

Let P be the set of paraphrases on the training corpus, we minimize the following hinge loss (L_H).

$$\sum_{(S_1,S_2)\in P} \sum_{w \in S_1 \cap S_2} \left[d_{S_1,S_2}(\mathbf{Mw}) + \gamma - d_{\hat{S}_1,\hat{S}_2}(\mathbf{Mw}) \right]_+ .$$

$(S_1, S_2) \in P$ thereof is a pair of paraphrases in P. $(\hat{S}_1, \hat{S}_2) \notin P$ is a negative sample generated by randomly substituting either S_1 or S_2 with another sentence in the dataset that contains w. $\gamma > 0$ is a hyper-parameter representing the margin. The operator $[x]_+$ denotes $\max(x, 0)$.

The orthogonalization is realized by the following regularization term.

$$L_O = \left\| \mathbf{I} - \mathbf{M}^\top \mathbf{M} \right\|_F ,$$

where $\|\cdot\|_F$ denotes the Frobenius norm, and $\mathbf{I}$ is an identity matrix. The learning objective of PAR is then denoted as $L = L_H + \lambda L_O$ with a positive hyperparameter λ.

Orthogonalizing $\mathbf{M}$ has two important effects: (i) It preserves the word similarity captured by the original input word representation (Rothe et al., 2016); (ii) It prevents the model from converging to a trivial solution where all word representations collapse to the same embedding vector.

4 Experiment

Our method can be integrated with any contextualized word embedding models. In our experiment, we apply PAR on ELMo (Peters et al., 2018) and evaluate the quality of the retrofitted ELMo on a broad range of sentence-level tasks and the adversarial SQuAD corpus.

4.1 Experimental Configuration

We use the officially released 3-layer ELMo (original), which is trained on the 1 Billion Word Benchmark with 93.6 million parameters. We retrofit ELMo with PAR on the training sets of three paraphrase datasets: (i) *MRPC* contains 2,753 paraphrase pairs; (ii) *Sampled Quora* contains randomly sampled 20,000 paraphrased question pairs (Iyer et al., 2017); and (iii) *PAN* training set (Madnani et al., 2012) contains 5,000 paraphrase pairs.

The orthogonal transformation $\mathbf{M}$ is initialized as an identity matrix. In our preliminary experiments, we observed that SGD optimizer is more stable and less likely to quickly overfit the training set than other optimizers with adaptive learning rates (Reddi et al., 2018; Kingma and Ba, 2015). Therefore, we use SGD with the learning rate of 0.005 and a batch size of 128. To determine the terminating condition, we train a Multi-Layer Perceptron (MLP) classifier on the same paraphrase training set and terminate training based on the paraphrase identification performance on a set of held-out paraphrases. The sentence in the dataset is represented by the average of the word embeddings. λ is selected from $\{0.1, 0.5, 1, 2\}$ and γ from $\{1, 2, 3, 4\}$ based on validation set. The best margin γ and epochs ζ by early stopping are $\{\gamma = 3, \zeta = 20\}$ on MRPC, $\{\gamma = 2, \zeta = 14\}$ on PAN, and $\{\gamma = 3, \zeta = 10\}$ on Sampled Quora, with $\lambda = 1$ in all settings.

4.2 Evaluation

We use the SentEval framework (Conneau and Kiela, 2018) to evaluate the sentence embeddings on a wide range of sentence-level tasks. We consider two baselines models: (1) ELMo (all layers) constructs a 3,074-dimensional sentence embedding by averaging the hidden states of all the language model layers. (2) ELMo (top layers) encodes a sentence to a 1,024 dimensional vector by averaging the representations of the top layer. We compare these baselines with four variants of PAR built upon ELMo (all layers) that trained on different paraphrase corpora.

4.3 Task Descriptions

Sentence classification tasks. We evaluate the sentence embedding on four sentence classification tasks including two sentiment analysis (MR (Pang and Lee, 2004), SST-2 (Socher et al., 2013)), product reviews (CR (Hu and Liu, 2004)), and opinion polarity (MPQA (Wiebe et al., 2005)). These tasks are all binary classification tasks. We employ a MLP with a single hidden layer of 50 neurons to train the classifer, using a batch size of 64 and Adam optimizer.

Sentence inference tasks We consider two sentence inference tasks: paraphrase identification on MRPC (Dolan et al., 2004) and the textual entailment on SICK-E (Marelli et al., 2014). MRPC consists of pairs of sentences, where the model aims to classify if two sentences are semantically equivalent. The SICK dataset contains 10,000 English sentence pairs annotated for relatedness in meaning and entailment. The aim of SICK-E is to detect discourse relations of entailment, contradiction and neutral between the two sentences. Similar to the sentence classification tasks, we apply a MLP with the same hyperparameters to conduct the classification.

Semantic textual similarity tasks. Semantic Textual Similarity (STS-15 (Agirre et al., 2015) and STS-16 (Agirre et al., 2016)) measures the degree of semantic relatedness of two sentences based on human-labeled scores from 0 to 5. We report the Pearson correlation between cosine similarity of two sentence representations and normalized human-label scores.

Semantic relatedness tasks The semantic relatedness tasks include SICK-R (Marelli et al., 2014) and the STS Benchmark dataset (Cer et al., 2017), which comprise pairs of sentences annotated with

Method	Classification (Acc %)				Relatedness \ Similarity (ρ)				Inference (Acc %)	
	MPQA	MR	CR	SST-2	SICK-R	STS-15	STS-16	STS-Benchmark	MRPC	SICK-E
ELMo (all layers)	89.55	79.72	85.11	86.33	0.84	0.69	0.64	0.65	71.65	81.86
ELMo (top layer)	89.30	79.36	84.13	85.28	0.81	0.67	0.63	0.62	70.20	79.64
ELMo-PAR (*MRPC*)	92.61	**83.40**	**87.01**	86.83	**0.87**	0.70	0.66	0.64	-	82.89
ELMo-PAR (*Sampled Quora*)	**93.76**	81.14	85.52	88.71	0.83	0.71	0.66	0.69	73.22	81.51
ELMo-PAR (*PAN*)	92.13	83.11	85.73	88.56	0.85	**0.73**	0.67	**0.70**	**74.86**	83.37
ELMo-PAR (*PAN+MRPC+Quora*)	93.40	82.26	86.39	**89.26**	0.86	**0.73**	**0.68**	0.67	-	**84.46**

Table 2: Performance on downstream applications. We report accuracy for classification and inference tasks, and Pearson correlation for relatedness and similarity tasks. We do not report results of ELMo-PAR on MRPC when MRPC is used in training the model. The baseline results are from (Perone et al., 2018).

Model	AddOneSent		AddSent	
	EM	F1	EM	F1
BiSAE	47.7	53.7	36.1	41.7
BiSAE-PAR(MRPC)	**51.6**	**57.9**	**40.8**	**47.1**

Table 3: Exact Match and F1 on Adversarial SQuAD.

semantic scores between 0 and 5. The goal of the tasks is to measure the degree of semantic relatedness between two sentences. We learn tree-structured LSTM (Tai et al., 2015) to predict the probability distribution of relatedness scores.

Adversarial SQuAD The Stanford Question Answering Datasets (SQuAD) (Rajpurkar et al., 2016) is a machine comprehension dataset containing 107,785 human-generated reading comprehension questions annotated on Wikipedia articles. Adversarial SQuAD (Jia and Liang, 2017) appends adversarial sentences to the passage in the SQuAD dataset to study the robustness of the model. We conduct evaluations on two Adversarial SQuAD datasets: AddOneSent which adds a random human-approved sentence, and AddSent which adds grammatical sentences that look similar to the question. We train the Bi-Directional Attention Flow (BiDAF) network (Seo et al., 2017) with self-attention and ELMo embeddings on the SQuAD dataset and test it on the adversarial SQuAD datasets.

4.4 Result Analysis

The results reported in Table 2 show that PAR leads to 2% $\sim$ 4% improvement in accuracy on sentence classification tasks and sentence inference tasks. It leads to 0.03 $\sim$ 0.04 improvement in Pearson correlation (ρ) on semantic relatedness and textual similarity tasks. The improvements on sentence similarity and semantic relatedness tasks shows that ELMo-PAR is more stable to the semantic-preserving modifications but more sensitive to subtle yet semantic-changing perturbations. PAR model trained on the combined corpus (*PAN+MRPC+Sampled Quora*) achieves the best improvement across all these tasks, showing the model benefits from a larger paraphrase corpus. Besides sentence-level tasks, Table 3 shows that the proposed PAR method notably improves the performance of a downstream question-answering task. For AddSent, ELMo-PAR achieves 40.8% in EM and 47.1% in F1. For AddOneSent, it boosts EM to 51.6% and F1 to 57.9%, which clearly shows that the proposed PAR method enhances the robustness of the downstream model combined with ELMo.

4.5 Case Study

Shared word distances We compute the average embedding distance of shared words in paraphrase and non-paraphrase sentence pairs from test sets of MRPC, PAN, and Quora. Results are listed in Table 4. Table 5 shows the ELMo-PAR embedding distance for the shared words in the examples in Table 1. Our model effectively minimizes the embedding distance of the shared words in the paraphrased contexts and maximize such distance in the non-paraphrased contexts.

5 Conclusion

We propose a method for retrofitting contextualized word embeddings, which leverages semantic equivalence information from paraphrases. PAR learns an orthogonal transformation on the input space of an existing model by minimizing the contextualized representations of shared words on paraphrased contexts without compromising the varying representations on non-paraphrased contexts. We demonstrate the effectiveness of this method applied to ELMo by a wide selection of semantic tasks. We seek to extend the use of PAR to other contextualized embeddings (Devlin et al., 2019; McCann et al., 2017) in future work.

Model	Paraphrase			Non-paraphrase		
	MRPC	**Quora**	**PAN**	**MRPC**	**Quora**	**PAN**
ELMo(all layers)	3.35	3.17	4.03	3.97	4.42	6.26
ELMo-PAR(*PAN+MRPC+Quora*)	1.80	1.34	1.21	4.73	5.49	6.54

Table 4: Averaging L2 distance for the shared word in paraphrased and non-paraphrased contexts.

Paraphrased contexts	**L2**	**Cosine**
How can I make **bigger** my arms? How do I make my arms **bigger**?	2.75	0.14
Some people believe earth is **flat**. Why? Why do people still believe in **flat** earth?	3.29	0.16
It is a very **small** window. I have a **large** suitcase.	5.84	0.30

Table 5: L2 and Cosine distance between embeddings of boldfaced words after retrofitting.

.

6 Acknowledgement

This work was supported in part by National Science Foundation Grant IIS-1760523. We thank reviewers for their comments.

References

Eneko Agirre, Carmen Banea, Claire Cardie, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Weiwei Guo, Inigo Lopez-Gazpio, 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*.

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. In *Proceedings of the 10th International Workshop on Semantic Evaluation*.

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*.

Alexis Conneau and Douwe Kiela. 2018. Senteval: An evaluation toolkit for universal sentence representations. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*.

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

Manaal Faruqui, Jesse Dodge, Sujay K Jauhar, Chris Dyer, Eduard Hovy, and Noah A Smith. 2015. Retrofitting word vectors to semantic lexicons. In *NAACL*.

Goran Glavaš and Ivan Vulić. 2018. Explicit retrofitting of distributional word vectors. In *ACL*.

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*.

Shankar Iyer, Nikhil Dandekar, and Kornél Csernai. 2017. First quora dataset release: Question pairs. *data.quora.com*.

Robin Jia and Percy Liang. 2017. Adversarial examples for evaluating reading comprehension systems. In *EMNLP*.

Diederik P Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. *ICLR*.

Nitin Madnani, Joel Tetreault, and Martin Chodorow. 2012. Re-examining machine translation metrics for paraphrase identification. In *NAACL*.

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. In *Proceedings of the 8th international workshop on semantic evaluation*.

Bryan McCann, James Bradbury, Caiming Xiong, and Richard Socher. 2017. Learned in translation: Contextualized word vectors. In *NIPS*.

Bo Pang and Lillian Lee. 2004. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *ACL*.

Christian S Perone, Roberto Silveira, and Thomas S Paula. 2018. Evaluation of sentence embeddings in downstream and linguistic probing tasks. *arXiv preprint arXiv:1806.06259*.

Matthew Peters, Waleed Ammar, Chandra Bhagavatula, and Russell Power. 2017. Semi-supervised sequence tagging with bidirectional language models. In *ACL*.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *NAACL*.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. In *EMNLP*.

Sashank J Reddi, Satyen Kale, and Sanjiv Kumar. 2018. On the convergence of adam and beyond. In *ICLR*.

Sascha Rothe, Sebastian Ebert, and Hinrich Schütze. 2016. Ultradense word embeddings by orthogonal transformation. In *NAACL*.

Minjoon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi. 2017. Bidirectional attention flow for machine comprehension. In *ICLR*.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *EMNLP*.

Emma Strubell, Ananya Ganesh, and Andrew McCallum. 2019. Energy and policy considerations for deep learning in nlp. In *ACM*.

Kai Sheng Tai, Richard Socher, and Christopher D Manning. 2015. Improved semantic representations from tree-structured long short-term memory networks. In *ACL-IJCNLP*.

Janyce Wiebe, Theresa Wilson, and Claire Cardie. 2005. Annotating expressions of opinions and emotions in language. *Language resources and evaluation*.

Zhiguo Yu, Trevor Cohen, Byron Wallace, Elmer Bernstam, and Todd Johnson. 2016. Retrofitting word vectors of mesh terms to improve semantic similarity measures. In *Proceedings of the Seventh International Workshop on Health Text Mining and Information Analysis*.