The 6th International Conference on Interdisciplinary research on
Computer science, Psychology, and Education (ICICPE' 2022)
December 26-28, 2022. Pattaya, Thailand.

Guidance by Semantic Search for Contrastive Learning of Sentence Embeddings

Dongsuk Oh[1,*], Suwan Kim[1,†], Heuiseok Lim[1,**]

[1] Computer Science and Engineering, Korea University
[†] {tndhks3837}@gmail.com
{inow3555, limhseok}@korea.ac.kr

Abstract. Universal sentence representations are an important open issue in natural language processing and must capture rich semantic information without task-specific fine-tuning. Various methods have been proposed for sentence embeddings. Still, the contrastive learning method that has the highest performance recently uses pre-trained language models. Sentence embeddings using contrastive learning is a method of learning to arrange sentences in a close space if the meanings are semantically similar and to place them farther apart if they are semantically dissimilar. In contrastive learning methods, unsupervised and supervised learning methods exist. In this paper, we propose an effective unsupervised learning method. In previous studies, the language model based on an unsupervised learning method learns by distinguishing the meaning of the sentence by itself. However, there is a limit to learning sentence representations only with information judged by one model because it can be learned biasedly. Therefore, in this paper, the performance of the existing model is improved by understanding the sentence pairs recommended from the semantic search. As a result, it shows a higher performance than the baseline in the STS tasks.

Keywords: Sentence Embeddings, Semantic Search, Contrastive Learning

1 Introduction

Learning universal sentence representations must understanding rich semantic information of sentences. Various methodologies using pre-trained language models have been proposed to improve the performance of sentence embeddings. Still, the method showing the highest performance among them is the pre-trained language model applying contrastive learning. Contrastive learning is a method of learning by placing samples that are semantically close to each other and placing samples that are not semantically close to each other [1]. There are the unsupervised methods for sentence embeddings using contrastive learning and supervised methods. The unsupervised method is a method in which the model learns by distinguishing

[*] This author is the first author.
[†] Work done as an intern at Computer Science and Engineering, Korea University.
[**] This author is corresponding author.

between positive and negative samples in a large sentence corpus. The supervised learning method is learning a corpus with positive and negative labels. However, building labeled data requires a lot of resources.

In this paper, sentence embeddings are performed using the unsupervised contrastive learning method. The unsupervised methods don't require resources to build labeled data, but it shows lower performance than supervised learning terms of performance. In addition, there is a limit in that learning is biased because it grasps the relationship between samples by itself without any guide. Therefore, we recommend sentence sample pairs for contrastive learning from semantic retrieval using the pre-trained language models. As a result, the model that received the semantic search guide performs better than the baseline in STS tasks.

2 Contrastive Learning with Semantic Search

Traditional retrieval engines look for literal matches with query sentences in a collection of text documents. These engines do not recognize synonyms, acronyms. Conversely, semantic search encodes the real values of query sentences and returns for sentences close to a vector space. The sentence embeddings of all sentences represent the [CLS] token. This architecture of semantic search is possible to overcome the disadvantages of the retrieval engine. We select the language model with the highest performance to use the semantic search model [13]. The sentence similarity calculation uses cosine similarity. And, the semantic search recommend sentence pairs for contrastive learning based on the language model. Then, the model learns universal sentence representation utilizing the contrastive learning method proposed by [2].

Equation (1) represents the training objective $\boldsymbol{l_i}$ for contrastive learning. When a sentence pair recommended from $\boldsymbol{D = (x_i, x_i^+)_{i=1}^m}$\$ is given from semantic search, $\boldsymbol{x_i}$ and $\boldsymbol{x_i^+}$ are sample pairs returned with high scores. And, it takes the cross-entropy objective with an in-batch negative [3,4]. $\boldsymbol{h_i}$ and $\boldsymbol{h_i^+}$ are representations of $\boldsymbol{x_i}$ and $\boldsymbol{x_i^+}$ through the [CLS] token of language models. where \$M\$ is the mini-batch, and $\boldsymbol{\tau}$ is the temperature hyperparameter. And, $\boldsymbol{sim}(*,*)$ is the cosine similarity.

$$\boldsymbol{l_i = \frac{e^{sim(h_i,h_i^+)/\tau}}{\sum_{j=1}^{M} e^{sim(h_i,h_j^+)/\tau}}} \tag{1}$$

3 Experiments

Our model evaluate in the STS dataset [5-11]. This data set consists of sentence pairs labeled with a similarity score between 0 and 5. Spearman's correlation evaluated the performance, and the SentEval tool was used for evaluation. The hardware environment performs in Google Colab Pro, and the parameter settings of the model are batch size:32 and learning rate:3e-5. The proposed method and the baseline are tested in the same parameter setting for a fair evaluation. The experimental results are shown in the Table 1. The model selects as $BERT_{base}$, and the baseline is compared with the unsupervised method proposed by [12]. Due to the hardware specifications, we experimented with a small batch size. Therefore, it may differ from the performance presented in the paper. Compared with the baseline, the overall average score shows a higher performance when learning the positive sample recommended from the semantic search. The experimental results show the performance difference when using information judged by itself in one language model and when using information suggested by a language model with better performance. As a result, learning positive samples by relying on random masks for dropouts of their language models shows lower performance. This method show limitation because it judges similar and dissimilar sentences by itself.

Table 1. Performance of sentence embedding on all STS tasks (Spearman's correlation).

Model	STS12	STS13	STS14	STS15	STS16	STS-B	STS-R	Avg
BERT	67.58	79.25	72.40	**80.50**	77.97	77.59	71.30	75.23
w/Semantic Search	**71.52**	**80.65**	**72.62**	79.50	**79.21**	**79.20**	**72.27**	**76.42**

4 Conclusion

In this paper, the performance of sentence embedding is improved by contrast learning using the semantic search method. The model selected $BERT_{base}$, which shows better performance than the baseline. The proposed method shows higher performance because it utilizes information from an external model. However, it is impossible to learn a sentence representation higher than the baseline unless a model better than the encoder that learns the actual sentence representation is used. Therefore, in the future, when learning sentence representation, the multi-task learning method will be developed so that the weights of the external language model can also be updated.

Acknowledgments. This work was supported by Institute of Information & communications Technology Planning & Evaluation(IITP) grant funded by the Korea government(MSIT) (No. 2020-0-00368, A Neural-Symbolic Model for Knowledge Acquisition and Inference Techniques). In addition, This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education(NRF-2021R1A6A1A03045425)

References

1. R. Hadsell, S. Chopra, and Y. LeCun, "Dimensionality reduction by learning an invariant mapping," 2006, IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06), Vol. 2, pp.1735–1742, 2006.
2. T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," International conference on machine learning, pp. 1597–1607, 2020.
3. T. Chen, Y. Sun, Y. Shi, and L. Hong, "On sampling strategies for neural network-based collaborative filtering," Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 767–776, 2017.
4. M. Henderson, R. Al-Rfou, B. Strope, Y.-H. Sung, L. Luk´acs, R. Guo, S. Kumar, B. Miklos, and R. Kurzweil, "Efficient natural language response suggestion for smart reply," arXiv preprint arXiv:1705.00652, 2017.
5. D. Cer, M. Diab, E. Agirre, I. Lopez-Gazpio, and L. Specia, "Semeval-2017 task 1: Semantic textual similarity-multilingual and cross-lingual focused evaluation," arXiv preprint arXiv:1708.00055, 2017.
6. E. Agirre, D. Cer, M. Diab, and A. Gonzalez-Agirre, "Semeval-2012 task 6: A pilot on semantic textual similarity," * SEM 2012: 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 (SemEval 2012), pp. 385–393, 2012.
7. E. Agirre, D. Cer, M. Diab, A. Gonzalez-Agirre, and W. Guo, "* sem 2013 shared task: Semantic textual similarity," Second joint conference on lexical and computational semantics (* SEM), volume 1: proceedings of the Main conference and the shared task: semantic textual similarity, pp. 32–43, 2013.
8. E. Agirre, C. Banea, C. Cardie, D. M. Cer, M. T. Diab, A. Gonzalez-Agirre, W. Guo, R. Mihalcea, G. Rigau, and J. Wiebe, "Semeval-2014 task 10: Multilingual semantic textual similarity." SemEval@ COLING, pp. 81–91, 2014.
9. E. Agirre, C. Banea, C. Cardie, D. Cer, M. Diab, A. Gonzalez-Agirre, W. Guo, I. Lopez-Gazpio, M. Maritxalar, R. Mihalcea et al., "Semeval-2015 task 2: Semantic textual similarity, english, spanish and pilot on interpretability," Proceedings of the 9th international workshop on semantic evaluation (SemEval 2015), pp. 252–263, 2015.
10. E. Agirre, C. Banea, D. Cer, M. Diab, A. Gonzalez Agirre, R. Mihalcea, G. Rigau Claramunt, and J. Wiebe, "Semeval-2016 task 1: Semantic textual similarity, monolingual and cross-lingual evaluation," SemEval-2016. 10th International Workshop on Semantic Evaluation; 2016 Jun 16-17; San Diego, CA. Stroudsburg (PA): ACL; 2016. p. 497-511., 2016.
11. M. Marelli, S. Menini, M. Baroni, L. Bentivogli, R. Bernardi, and R. Zamparelli, "A sick cure for the evaluation of compositional distributional semantic models," Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), pp. 216–223, 2014.
12. T. Gao, X. Yao, and D. Chen, "Simcse: Simple contrastive learning of sentence embeddings," Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pp. 6894–6910, 2021.
13. https://www.sbert.net/docs/pretrained_models.html