

DefSent: Sentence Embeddings using Definition Sentences

Hayato Tsukagoshi Ryohei Sasano Koichi Takeda

Graduate School of Informatics, Nagoya University
tsukagoshi.hayato@e.mbox.nagoya-u.ac.jp,
{sasano,takedasu}@i.nagoya-u.ac.jp

Abstract

Sentence embedding methods using natural language inference (NLI) datasets have been successfully applied to various tasks. However, these methods are only available for limited languages due to relying heavily on the large NLI datasets. In this paper, we propose DefSent, a sentence embedding method that uses definition sentences from a word dictionary, which performs comparably on unsupervised semantics textual similarity (STS) tasks and slightly better on SentEval tasks than conventional methods. Since dictionaries are available for many languages, DefSent is more broadly applicable than methods using NLI datasets without constructing additional datasets. We demonstrate that DefSent performs comparably on unsupervised semantics textual similarity (STS) tasks and slightly better on SentEval tasks to the methods using large NLI datasets. Our code is publicly available at <https://github.com/hpprc/defsent>.

1 Introduction

Sentence embeddings represent sentences as dense vectors in a low dimensional space. Recently, sentence embedding methods using natural language inference (NLI) datasets have been successfully applied to various tasks, including semantic textual similarity (STS) tasks. However, these methods are only available for limited languages due to relying heavily on the large NLI datasets. In this paper, we propose DefSent, a sentence embedding method that uses definition sentences from a word dictionary. Since dictionaries are available for many languages, DefSent is more broadly applicable than the methods using NLI datasets without constructing additional datasets.

Defsent is similar to the model proposed by Hill et al. (2016) in that it generates sentence embeddings so that the embeddings of a definition sen-

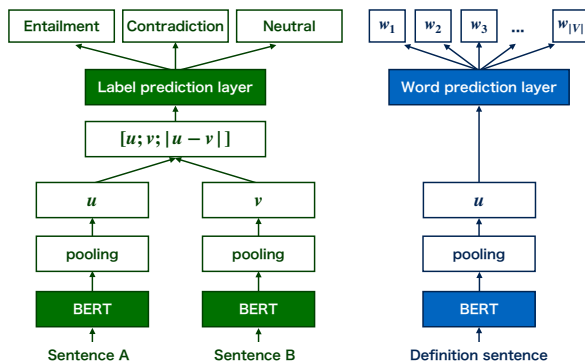


Figure 1: Sentence-BERT (left) and DefSent (right).

tence and the word it represents are similar. However, while Hill et al. (2016)’s model is based on recurrent neural network language models, DefSent is based on pre-trained language models such as BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019), with a fine-tuning mechanism as well as Sentence-BERT (Reimers and Gurevych, 2019). Sentence-BERT is one of the state-of-the-art sentence embedding models, which is based on pre-trained language models that are fine-tuned on NLI datasets. Overviews of Sentence-BERT and DefSent are depicted on Figure 1.

2 Sentence Embedding Methods

In this section, we introduce BERT, RoBERTa, and Sentence-BERT, followed by a description of DefSent, our proposed sentence embedding method.

2.1 BERT and RoBERTa

BERT is a pre-trained language model based on the Transformer architecture (Vaswani et al., 2017). Utilizing masked language modeling and next sentence prediction, BERT acquires linguistic knowledge and outputs contextualized word embeddings. In masked language modeling, a specific proportion of input tokens is replaced with a special token [MASK], and the model is trained to predict these

masked tokens. Next sentence prediction is a task to predict whether two sentences connected by a sentence separator token [SEP] are consecutive sentences in the original text data. BERT uses the output embedding of the unique token [CLS] at the beginning of each such sentence for prediction.

RoBERTa has the same structure as BERT. It attempts to improve BERT by removing the next sentence prediction from pre-training objectives and increasing the data size and batch size. While both Sentence-BERT and DefSent are applicable to BERT and RoBERTa, we use BERT for the explanations in this paper.

2.2 Sentence-BERT

Conneau et al. (2017) proposed InferSent, a sentence encoder based on a Siamese network structure. InferSent trains the sentence encoder such that similar sentences are distributed close to each other in the semantic space. Reimers and Gurevych (2019) proposed Sentence-BERT, which also uses a Siamese network to create BERT-based sentence embeddings. An overview of Sentence-BERT is depicted on the left side of Figure 1. Sentence-BERT first inputs the sentences to BERT and then constructs a sentence embedding from the output contextualized word embeddings by pooling. They utilize the following three types of pooling strategy.

CLS Using the [CLS] token embedding.; When using RoBERTa, since the [CLS] token does not exist, the beginning-of-sentence token <s> is used as an alternative.

Mean Using the mean of the contextualized embeddings of all words in a sentence.

Max Using the max-over-time of the contextualized embeddings of all words in a sentence.

Let u and v be the sentence embeddings for each of the sentence pairs obtained by pooling. Then compose a vector $[u; v; |u - v|]$ and feed it to the label prediction layer, which has the same number of output dimensions as the number of classes. For fine-tuning, Reimers and Gurevych uses the SNLI dataset (Bowman et al., 2015) and the Multi-Genre NLI dataset (Williams et al., 2018), which together contain about one million sentences.

2.3 DefSent

Since they have the same meaning, we focus on the relationship between a definition sentence and the word it represents. To learn how to embed

sentences in the semantic vector space, we train the sentence embedding model by predicting the word from definitions. An overview of DefSent is depicted on the right side of Figure 1. We call the layer that predicts the original token from the [MASK] embeddings used in the masked language modeling during BERT pre-training a word prediction layer. Also, we use w_k to denote the word corresponding to a given definition sentence X_k .

DefSent inputs the definition sentence X_k to BERT and derives the sentence embedding u by pooling the output embeddings. As in Sentence-BERT, three types of pooling strategy are used: CLS, Mean, and Max. Then, the derived sentence embedding u is input to the word prediction layer to obtain the probability $P(w_k|X_k)$. We use cross-entropy loss as a loss function and fine-tune BERT to maximize $P(w_k|X_k)$.

In DefSent, the parameters of the word prediction layer are fixed. This setting allows us to fine-tune models without training an additional classifier, as is the case with both InferSent and Sentence-BERT. Additionally, since our method uses a word prediction layer that has been pre-trained in masked language modeling, the sentence embedding u is expected to be similar to the contextualized word embedding of w_k when w_k appears as the same meaning as X_k .

3 Word Prediction Experiment

To evaluate how well DefSent can predict words from sentence embeddings, we conducted an experiment to predict a word from its definition.

3.1 Dataset

DefSent requires pairs of a word and its definition sentence. We extracted these from the Oxford Dictionary dataset used by Ishiwatari et al. (2019). Each entry in the dataset consists of a word and its definition sentence, and a word can have multiple definitions. We split this dataset into train, dev, and test sets in the ratio of 8:1:1 word by word to evaluate how well the model can embed unseen definitions of unseen words. It is worth noting that since DefSent utilizes the pre-trained word prediction layer of BERT and RoBERTa, it is impossible to obtain probabilities for out-of-vocabulary (OOV) words. Therefore, we cannot calculate losses of these OOV words in a straightforward way.¹ In our

¹Although we could substitute the mean of subwords as OOV word embeddings, we opted to filter out OOV words for

experiments, we only use words and their respective definitions in the dataset, as contained by the model vocabulary. The statistics of the datasets are listed in Table 1.

3.2 Settings

We used the following pre-trained models: BERT-base (bert-base-uncased), BERT-large (bert-large-uncased), RoBERTa-base (roberta-base), and RoBERTa-large (roberta-large) from Transformers (Wolf et al., 2020). The batch size was 16, a fine-tuning epoch size was 1, the optimizer was Adam (Kingma and Ba, 2015), and we set a linear learning rate warm-up over 10% of the training data. For each respective model and pooling strategy, the learning rate was chosen based on the highest recorded Mean Reciprocal Rank (MRR) for the dev set in the range of $2^x \times 10^{-6}$, $x \in \{0, 0.5, 1, \dots, 7\}$. We conducted experiments with ten different random seeds, and their mean was used as the evaluation score. Top- k accuracy (the percentage of correct answers within the first, third, and tenth positions) and MRR were calculated from the output word probabilities when a definition sentence was fed into the model. Also, we evaluated the performance of BERT-base without fine-tuning for comparison.

3.3 Results

Table 2 shows the experimental results.² Max was the best pooling strategy for BERT-base without fine-tuning, but its top-1 accuracy was extremely low at 0.0157. This indicates that it is not adequate for predicting words from definitions without fine-tuning. DefSent performed higher for larger models. In the case of BERT, CLS was the best pooling strategy for both base and large models. CLS was also the best pooling strategy for RoBERTa-base but Mean was the best for RoBERTa-large.

4 Extrinsic Evaluations

Next, to evaluate the general quality of the constructed sentence embedding, we conducted evaluations on semantic textual similarity (STS) tasks and SentEval tasks (Conneau and Kiela, 2018).

simplicity and intuitiveness.

²We report the fine-tuning time and computing infrastructure in Appendix A, and report the learning rate, means, and standard deviations on the word prediction experiment in Appendix B. We also show the actual predicted words when definition sentences and other sentences are given as inputs in Appendices C and D, respectively.

All	Words	Definitions	Avg. length
Train	29,413	97,759	9.921
Dev	3,677	12,127	9.874
Test	3,677	12,433	9.846
In BERT vocab.	Words	Definitions	Avg. length
Train	7,732	54,142	9.531
Dev	936	6,544	9.512
Test	979	6,930	9.551
In RoBERTa vocab.	Words	Definitions	Avg. length
Train	7,269	53,935	9.376
Dev	901	6,625	9.372
Test	925	6,945	9.410

Table 1: Statistics of datasets.

Model	Pooling	MRR	Top1	Top3	Top10
BERT-base (no fine-tuning)	CLS	.0009	.0000	.0000	.0000
	Mean	.0132	.0001	.0043	.0242
	Max	.0327	.0157	.0320	.0626
BERT-base	CLS	.3200	.2079	.3670	.5418
	Mean	.3091	.1972	.3524	.5356
	Max	.2939	.1840	.3350	.5207
BERT-large	CLS	.3587	.2388	.4139	.6011
	Mean	.3286	.2091	.3792	.5723
	Max	.2925	.1814	.3356	.5194
RoBERTa-base	CLS	.3436	.2241	.3983	.5836
	Mean	.3365	.2170	.3906	.5783
	Max	.3072	.1941	.3523	.5386
RoBERTa-large	CLS	.3863	.2611	.4460	.6364
	Mean	.3995	.2699	.4634	.6599
	Max	.3175	.2015	.3646	.5543

Table 2: Results of word prediction experiments.

4.1 Settings

We compared the performance of DefSent with several existing sentence embedding methods including InferSent (Conneau et al., 2017), Universal Sentence Encoder (Cer et al., 2018), and SentenceBERT (Reimers and Gurevych, 2019). For the pooling strategies, we used the strategy that achieved the highest MRR in the word prediction task for each pre-trained model.³ The performance of the existing methods was taken from Reimers and Gurevych (2019).

4.2 Semantic textual similarity tasks

We evaluated DefSent on unsupervised STS tasks. In these tasks, we compute semantic similarities of given sentence pairs and calculate Spearman’s rank correlation ρ between similarities and gold scores of sentence similarities. In the unsupervised setting, none of the models are optimized on the STS datasets. Instead, the similarities of the given sentence embeddings are calculated using common similarity measures such as negative Manhattan distance, negative Euclidean distance, and cosine-similarity. In this study, we used cosine-similarity.

³We report the means and standard deviations on the unsupervised STS tasks and SentEval tasks for each respective model and pooling strategy in Appendices E and F.

Model	STS12	STS13	STS14	STS15	STS16	STS-B	SICK-R	Avg.
Avg. GloVe embeddings (Pennington et al., 2014)	55.14	70.66	59.73	68.25	63.66	58.02	53.76	61.32
Avg. BERT embeddings	38.78	57.98	57.98	63.15	61.06	46.35	58.40	54.81
BERT CLS-vector	20.16	30.01	20.09	36.88	38.08	16.50	42.63	29.19
InferSent - GloVe (Conneau et al., 2017)	52.86	66.75	62.15	72.77	66.87	68.03	65.65	65.01
Universal Sentence Encoder (Cer et al., 2018)	64.49	67.80	64.61	76.83	73.18	74.92	76.69	71.22
Sentence-BERT-base (Mean)	70.97	76.53	73.19	79.09	74.30	77.03	72.91	74.89
Sentence-BERT-large (Mean)	72.27	78.46	74.90	80.99	76.25	79.23	73.75	76.55
Sentence-RoBERTa-base (Mean)	71.54	72.49	70.80	78.74	73.69	77.77	74.46	74.21
Sentence-RoBERTa-large (Mean)	74.53	77.00	73.18	81.85	76.82	79.10	74.29	76.68
DefSent-BERT-base (CLS)	67.56	79.86	69.52	76.83	76.61	75.57	73.05	74.14
DefSent-BERT-large (CLS)	66.22	82.07	71.48	79.34	75.38	73.46	74.30	74.61
DefSent-RoBERTa-base (CLS)	65.55	80.84	71.87	78.77	79.29	78.13	74.92	75.62
DefSent-RoBERTa-large (Mean)	58.36	76.24	69.55	73.15	76.90	78.53	73.81	72.36

Table 3: Spearman’s rank correlation $\rho \times 100$ between cosine similarities of sentence embeddings and human ratings. STS-B denotes STS Benchmark, and SICK-R denotes SICK-Relatedness.

Model	MR	CR	SUBJ	MPQA	SST-2	TREC	MRPC	Avg.
Avg. GloVe embeddings	77.25	78.30	91.17	87.85	80.18	83.00	72.87	81.52
Avg. BERT embeddings	78.66	86.25	94.37	88.66	84.40	92.80	69.45	84.94
BERT CLS-vector	78.68	84.85	94.21	88.23	84.13	91.40	71.13	84.66
InferSent - GloVe	81.57	86.54	92.50	90.38	84.18	88.20	75.77	85.59
Universal Sentence Encoder	80.09	85.19	93.98	86.70	86.38	93.20	70.14	85.10
Sentence-BERT-base (Mean)	83.64	89.43	94.39	89.86	88.96	89.60	76.00	87.41
Sentence-BERT-large (Mean)	84.88	90.07	94.52	90.33	90.66	87.40	75.94	87.69
DefSent-BERT-base (CLS)	80.94	87.57	94.59	89.98	85.78	89.73	73.82	86.06
DefSent-BERT-large (CLS)	85.79	90.54	95.58	90.15	91.17	90.47	73.74	88.20
DefSent-RoBERTa-base (CLS)	83.94	90.44	94.05	90.70	89.16	90.80	75.52	87.80
DefSent-RoBERTa-large (Mean)	86.47	91.53	95.02	91.15	90.77	92.33	73.91	88.74

Table 4: Accuracy (%) for each task in SentEval.

We performed experiments on unsupervised STS tasks using the STS12-16 (Agirre et al., 2012, 2013, 2014, 2015, 2016), STS Benchmark (Cer et al., 2017), and SICK-Relatedness (Marelli et al., 2014) datasets. These datasets contain sentence pairs and their similarity scores, which is a real number from 0 to 5 assigned by human evaluations. Experiments were conducted with ten different random seeds, and the mean was used as the evaluation score.

Table 3 shows the experimental results. Although the training data size used in DefSent was only about 5% that of Sentence-BERT, DefSent-BERT-base and DefSent-RoBERTa-base performed comparably to Sentence-BERT-base and Sentence-RoBERTa-base. In particular, DefSent-RoBERTa models showed high performance in the STS Benchmark.

4.3 SentEval

SentEval (Conneau and Kiela, 2018) is a popular toolkit for evaluating the quality of universal sentence embeddings that aggregates various tasks, including binary and multi-class classification, natural language inference, and sentence similarity. For the SentEval evaluations, we trained a logistic regression classifier using sentence embeddings as

input features to evaluate the extent to which each sentence embedding contained the important information for each task. We used the same tasks and settings as Reimers and Gurevych (2019) and performed a 10-fold cross-validation. We conducted experiments with three different random seeds, and the mean was used as the evaluation score.

Table 4 shows the results.⁴ DefSent-RoBERTa-large achieved the best average score among all models. Also, increasing the model size improved the performance consistently. The performances of DefSent-BERT-large, DefSent-RoBERTa-base, and DefSent-RoBERTa-large were better than the performances of Sentence-BERT-based methods. These results indicate that DefSent embeds useful information that can be applied to various tasks.

5 Conclusion

In this paper, we proposed DefSent, a new sentence embedding method using a dictionary, and demonstrated its effectiveness through a series of experiments. Its performance was comparable to or even slightly better than existing methods using

⁴Reimers and Gurevych (2019) reported that there were minor difference from Sentence-BERT, so we omitted the results of Sentence-RoBERTa.

large NLI datasets. DefSent is based on dictionaries developed for many languages, so it does not require new language resources when applied to other languages. Since the model is trained with the same word prediction process as the masked language modeling, sentence embeddings derived by DefSent are expected to be similar to contextualized word embeddings of a word when it appears with the same meaning as the definition.

In future work, we will evaluate the performance of DefSent when it is applied to languages other than English and when it is applied to a broader range of downstream tasks, such as document classification tasks. We will also analyze the relationship between the sentence embeddings by DefSent and the contextualized word embeddings in the semantic vector space and investigate how model architecture and size influence the embeddings.

Acknowledgements

This work was supported by JSPS KAKENHI Grant Number 21H04901.

References

- Eneko Agirre, Carmen Banea, Claire Cardie, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Weiwei Guo, Iñigo 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)*, pages 252–263.
- 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)*, pages 81–91.
- 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 (SemEval)*, pages 497–511.
- Eneko Agirre, Daniel Cer, Mona Diab, and Aitor Gonzalez-Agirre. 2012. [SemEval-2012 Task 6: A Pilot on Semantic Textual Similarity](#). In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Semantic Evaluation (SemEval)*, pages 385–393.
- 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)*, pages 32–43.
- Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. [A large annotated corpus for learning natural language inference](#). In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 632–642.
- Daniel Cer, Mona Diab, Eneko Agirre, Iñigo 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)*, pages 1–14.
- Daniel Matthew Cer, Yinfei Yang, Sheng yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, C. Tar, Yun-Hsuan Sung, B. Strope, and R. Kurzweil. 2018. [Universal Sentence Encoder](#). *arXiv:1803.11175*.
- 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)*, pages 1699–1704.
- 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 (EMNLP)*, pages 670–680.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. [BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)*, pages 4171–4186.
- Felix Hill, Kyunghyun Cho, Anna Korhonen, and Yoshua Bengio. 2016. [Learning to Understand Phrases by Embedding the Dictionary](#). In *Transactions of the Association for Computational Linguistics (ACL)*, pages 17–30.
- Shonosuke Ishiwatari, Hiroaki Hayashi, Naoki Yoshinaga, Graham Neubig, Shoetsu Sato, Masashi Toyoda, and Masaru Kitsuregawa. 2019. [Learning to Describe Unknown Phrases with Local and Global Contexts](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)*, pages 3467–3476.
- Diederik P. Kingma and Jimmy Ba. 2015. [Adam: A Method for Stochastic Optimization](#). In *3rd International Conference on Learning Representations (ICLR)*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, M. Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. [RoBERTa: A Robustly Optimized BERT Pretraining Approach](#). *arXiv:1907.11692*.

Marco Marelli, Stefano Menini, Marco Baroni, Luisa Bentivogli, Raffaella Bernardi, and Roberto Zamparelli. 2014. [A SICK cure for the evaluation of compositional distributional semantic models](#). In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC)*, pages 216–223.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. [GloVe: Global Vectors for Word Representation](#). In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Nils Reimers and Iryna Gurevych. 2019. [Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. [Attention is All you Need](#). In *Advances in Neural Information Processing Systems (NIPS)*, pages 5998–6008.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. [A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference](#). In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)*, pages 1112–1122.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. [Transformers: State-of-the-Art Natural Language Processing](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP): System Demonstrations*, pages 38–45.

A Average Runtime and Computing Infrastructure

Fine-tuning for DefSent-BERT-base and DefSent-RoBERTa-base took about 5 minutes on a single NVIDIA GeForce GTX 1080 Ti. Fine-tuning for DefSent-BERT-large and DefSent-RoBERTa-large took about 15 minutes on a single Quadro GV100.

B Full Results of the Word Prediction Experiment

Table 5 shows the experimental results on the word prediction experiment for each model and pooling strategy with learning rate.

C Predictions for definition sentences

Table 6 shows the predicted words when the embeddings of definition sentences are input. We used BERT-large as a model and CLS as a pooling strategy for the experiment. For prediction, sentences were first input into the model to obtain sentence embeddings. Then the sentence embeddings were input into the pre-trained word prediction layer to obtain word probabilities. We show the top five words with the highest probability.

D Predictions for sentences other than definition sentences

Table 7 shows the predicted words when the embeddings of sentences other than definition sentences are input. We used BERT-large as a model and CLS as a pooling strategy for the experiment. The evaluation procedure is the same as for Appendix C.

E Full Results of the STS Evaluation

Table 8 shows the experimental results on STS tasks for each model and pooling strategy.

F Full Results of the SentEval Evaluation

Table 9 shows the experimental results on SentEval tasks for each model and pooling strategy.

Model	Pooling	Learning rate	MRR	Top1	Top3	Top10
BERT-base	CLS	$2^{2.5} \times 10^{-6}$.3200 \pm .0020	.2079 \pm .0021	.3670 \pm .0029	.5418 \pm .0022
	Mean	$2^{3.5} \times 10^{-6}$.3091 \pm .0021	.1972 \pm .0030	.3524 \pm .0038	.5356 \pm .0029
	Max	$2^{3.5} \times 10^{-6}$.2939 \pm .0021	.1840 \pm .0026	.3350 \pm .0023	.5207 \pm .0045
BERT-large	CLS	$2^{2.5} \times 10^{-6}$.3587 \pm .0043	.2388 \pm .0047	.4139 \pm .0059	.6011 \pm .0054
	Mean	$2^{3.5} \times 10^{-6}$.3286 \pm .0044	.2091 \pm .0045	.3792 \pm .0055	.5723 \pm .0072
	Max	$2^{3.0} \times 10^{-6}$.2925 \pm .0138	.1814 \pm .0113	.3356 \pm .0172	.5194 \pm .0181
RoBERTa-base	CLS	$2^{2.5} \times 10^{-6}$.3436 \pm .0016	.2241 \pm .0016	.3983 \pm .0027	.5836 \pm .0017
	Mean	$2^{3.0} \times 10^{-6}$.3365 \pm .0017	.2170 \pm .0014	.3906 \pm .0029	.5783 \pm .0022
	Max	$2^{2.0} \times 10^{-6}$.3072 \pm .0037	.1941 \pm .0039	.3523 \pm .0050	.5386 \pm .0064
RoBERTa-large	CLS	$2^{2.0} \times 10^{-6}$.3863 \pm .0040	.2611 \pm .0045	.4460 \pm .0044	.6364 \pm .0041
	Mean	$2^{2.0} \times 10^{-6}$.3995 \pm .0041	.2699 \pm .0053	.4634 \pm .0042	.6599 \pm .0036
	Max	$2^{2.5} \times 10^{-6}$.3175 \pm .0069	.2015 \pm .0054	.3646 \pm .0087	.5543 \pm .0092

Table 5: MRR, top-1, top-3, and top-10 accuracy on the word prediction experiment. The scores are the mean and standard deviation of 10 evaluations with different random seeds.

Word	Definition	Predictions (1st, 2nd, 3rd)		
cost	be expensive for (someone)	cost	charge	pay
preserve	prevent (food) from rotting	preserve	keep	spoil
good	that which is pleasing or valuable or useful	good	pleasing	pleasure
linux	an open-source operating system modelled on unix.	linux	unix	gnu
pile	place or lay as if in a pile	pile	stack	heap
weird	very strange; bizarre	weird	strange	bizarre
sale	the general activity of selling	selling	sale	retail
satellite	a celestial body orbiting the earth or another planet.	planet	satellite	orbit
logic	the quality of being justifiable by reason	reason	justice	certainty
custom	a thing that one does habitually	habit	routine	ritual
chief	a person who is in charge	leader	boss	master
nirvana	an ideal or idyllic state or place	paradise	dream	ideal

Table 6: Predicted words when the embeddings of definition sentences are input. The first two columns represent words and their defining sentences, and the third to fifth columns represent the top three predicted words. Correctly predicted words shown in **bold**.

Input	Predictions (1st, 2nd, 3rd, 4th, 5th)				
royal man	king	royal	prince	noble	knight
royal woman	queen	princess	royal	regal	sovereign
royal boy	boy	prince	royal	king	baby
royal girl	princess	queen	lady	royal	belle
good	fine	good	great	right	solid
bad	bad	dirty	awful	ugly	nasty
not good	bad	poor	wrong	awful	terrible
not bad	okay	fair	good	fine	ok
Star wars	jedi	star	trek	galaxy	saga
Star wars in America	jedi	western	fan	hollywood	movie
Star wars in Europe	trek	space	adventure	cinema	fantas
Star wars in Japan	godzilla	anime	gundam	jedi	manga
captain america	marvel	hero	thor	superhero	hulk

Table 7: Predicted words when the embeddings of sentences other than definition sentences are input.

Model	Pooling	STS12	STS13	STS14	STS15	STS16	STS-B	SICK-R	Avg.
BERT-base	CLS	67.56±0.26	79.86±0.25	69.52±0.39	76.83±0.32	76.61±0.33	75.57±0.37	73.05±0.32	74.14±0.25
	Mean	67.30±0.44	81.96±0.24	71.92±0.28	77.68±0.47	76.71±0.48	76.90±0.40	73.28±0.30	75.11±0.21
	Max	64.61±0.87	82.06±0.21	72.43±0.31	76.56±0.74	75.61±0.43	76.61±0.52	72.15±0.46	74.29±0.33
BERT-large	CLS	66.22±0.79	82.07±0.39	71.48±0.33	79.34±0.44	75.38±0.60	73.46±0.45	74.30±0.50	74.61±0.41
	Mean	64.18±0.96	82.76±0.42	73.14±0.32	79.66±0.92	77.93±0.78	77.89±0.89	73.98±0.46	75.65±0.53
	Max	58.94±1.06	81.03±0.66	71.34±0.88	76.23±1.83	76.07±0.56	75.75±0.70	71.69±0.74	73.01±0.74
RoBERTa-base	CLS	65.55±0.89	80.84±0.26	71.87±0.39	78.77±0.70	79.29±0.27	78.13±0.61	74.92±0.18	75.62±0.38
	Mean	60.78±1.41	77.17±0.60	69.71±0.73	75.13±1.00	77.75±0.38	76.52±0.63	74.10±0.45	73.02±0.63
	Max	63.85±0.86	78.55±0.90	71.19±0.86	76.55±1.12	77.86±0.59	78.02±0.77	73.97±0.46	74.28±0.62
RoBERTa-large	CLS	63.84±1.34	77.33±2.53	68.64±1.34	72.86±1.96	77.13±1.32	78.32±1.08	74.14±1.31	73.18±1.20
	Mean	58.36±1.16	76.24±0.87	69.55±0.85	73.15±1.32	76.90±0.94	78.53±0.54	73.81±0.88	72.36±0.73
	Max	62.89±1.42	77.99±1.88	69.83±1.66	75.60±1.51	79.63±0.60	79.34±0.48	74.04±0.84	74.19±0.88

Table 8: Spearman’s rank correlation $\rho \times 100$ between the cosine similarities of the sentence embeddings and the human ratings for each model and pooling strategy. The scores are the mean and standard deviation of 10 evaluations with different random seeds.

Model	Pooling	MR	CR	SUBJ	MPQA	SST-2	TREC	MRPC	Avg.
BERT-base	CLS	80.94±0.08	87.57±0.12	94.59±0.09	89.98±0.04	85.78±1.14	89.73±0.76	73.82±0.19	86.06±0.28
	Mean	81.84±0.17	88.20±0.04	94.82±0.12	89.94±0.12	86.49±0.20	89.73±0.31	75.32±0.78	86.62±0.18
	Max	80.74±0.16	88.00±0.09	94.32±0.07	89.92±0.25	85.03±0.09	89.13±0.50	74.11±0.49	85.89±0.02
BERT-large	CLS	85.79±0.19	90.54±0.26	95.58±0.14	90.15±0.04	91.17±0.06	90.47±0.95	73.74±0.61	88.20±0.07
	Mean	84.05±0.25	89.50±0.24	95.21±0.12	90.19±0.36	89.44±0.14	88.60±0.87	73.99±0.90	87.28±0.05
	Max	83.48±0.30	89.04±0.37	94.55±0.09	89.88±0.17	87.50±0.26	90.87±1.30	74.28±1.27	87.09±0.27
RoBERTa-base	CLS	83.94±0.30	90.44±0.49	94.05±0.06	90.70±0.17	89.16±0.22	90.80±0.35	75.52±0.42	87.80±0.20
	Mean	84.88±0.21	91.09±0.01	94.60±0.10	90.69±0.07	89.73±0.54	93.13±0.12	77.22±0.46	88.76±0.08
	Max	83.98±0.03	90.78±0.24	93.96±0.07	90.63±0.11	90.05±0.06	93.60±0.72	77.80±0.32	88.69±0.12
RoBERTa-large	CLS	85.63±0.27	90.74±0.15	94.53±0.14	91.20±0.11	90.08±0.59	93.53±0.76	72.66±1.73	88.34±0.28
	Mean	86.47±0.29	91.53±0.06	95.02±0.08	91.15±0.07	90.77±0.34	92.33±0.64	73.91±0.96	88.74±0.12
	Max	85.60±0.26	90.73±0.70	94.21±0.65	91.09±0.32	90.65±0.37	91.53±1.70	76.15±0.33	88.56±0.57

Table 9: The percentage of correct answers (%) for each task of SentEval. The scores are the mean and standard deviation of three evaluations with different random seeds.