Improving Sentence Embedding With Sentence Relationships From Word Analogies

Qixuan Zhang[1,*], Yves Lepage[1,*]

[1] *Waseda University, Japan*

Abstract

In this study, we introduce a novel approach to enhance sentence embedding by leveraging word analogy. Compared with past methods that use word analogy on sentence-level tasks, our method is less affected by sentence patterns and pays more attention to semantic relations. By fine-tuning pre-trained models as BERT, RoBERTa and Sentence-BERT and evaluating their performance on inter-sentence downstream tasks, we demonstrate the efficiency of our method. Our experimental results show that each model, following fine-tuning using our approach, exhibits improvements across all inter-sentence tasks. In the STS task, our method increases the average result from 18.63% to 62.52% on BERT. This outcome substantiates that sentence relationships derived from word analogy contain valuable knowledge that can enhance the performance of sentence embedding models.

Keywords

word analogy, sentence embedding, semantic relationship

1. Introduction

Generating meaningful representations for sentences has been a subject of great interest in the field of natural language processing (NLP). Accurate sentence embeddings are crucial for a wide range of downstream tasks, including sentiment analysis and translation. Previous research, as summarized by Li et al. [1], has shown that models trained on Natural Language Inference (NLI) datasets often outperform others in various evaluation tasks. NLI datasets provide valuable world knowledge that helps sentence embedding models understand the meaning of sentences. However, creating NLI datasets, such as the Stanford Natural Language Inference (SNLI) dataset [2], requires substantial human effort, with thousands of contributors involved.

Therefore, we propose a method to generate sentence relationship data almost automatically, thus can be applied to low-resource languages with low cost. The main idea is to map the semantic relationships in the word analogy dataset to the definition sentences corresponding to the words. This process results in organized clusters of sentence relationships, which we refer to as **D**efinition **S**entences from **BATS** (DSBATS). Each DSBATS cluster contains pairs of sentences that represent specific relationships, such as the relationship between an animal and its sound (e.g., *"feline mammal usually having thick soft fur and no ability to roar"* and *"the*

ICCBR ATA'23: Workshop on Analogies: From Theory to Applications – AR & CBR Tools for Metric and Representation Learning at ICCBR2023, July 17 – 20, 2023, Aberdeen, Scotland

[*] Corresponding author.

oakori@toki.waseda.jp (Q. Zhang); yves.lepage@waseda.jp (Y. Lepage)
0009-0006-3247-3146 (Q. Zhang); 0000-0002-3059-4271 (Y. Lepage)

CEUR Workshop Proceedings (CEUR-WS.org)

sound made by a cat"). In total, DSBATS based on semantic network (DSBATS-sn) consists of 20 clusters, each capturing a distinct relationship.

We employ contrastive learning and DSBATS-sn to fine-tune popular models, including BERT, RoBERTa, and Sentence-BERT. We do data augmentation on DSBATS-sn to get DSBATS for contrastive learning (DSBATS4CL). Through a series of experiments, we evaluate the performance of our approach in three inter-sentence downstream tasks. Firstly, we propose an intrinsic evaluation task called "sentence relationship similarity distinguishing", a task of identifying whether the two sentence relationships are the same. Fine-tuning the models with DSBATS4CL leads to performance improvements of 8.37%, 7.42%, and 7.87% on BERT, RoBERTa, and Sentence-BERT, respectively, compared to the performance of the original pre-trained models. Secondly, in the Semantic Textual Similarity (STS) task, our method achieves improvements of 43.6%, 21.46%, and 13.89% on the three pre-trained models, respectively. Additionally, our approach consistently produces modest improvements on the Microsoft Research Paraphrase Corpus (MRPC) dataset.

We prove that the language model can learn knowledge from sentence relationships generated from word analogy to improve the performance on semantic analysis tasks. Compared with the sentence relationships from NLI datasets, our method reduces the need for human annotation and increases the diversity of inter-sentence relations effort by using semantic network and word analogy data. Meanwhile, sentence relationship similarity distinguishing task proposed in this paper is also a challenging evaluation metric for sentence embedding method.

2. Pretrained models and sentence embedding

Pretrained models have played a significant role in the advancement of natural language processing (NLP) tasks. They are models that are pre-trained on large corpora of text data to learn language representations that capture semantic and syntactic properties of words and sentences. Transformer-based pre-trained model like BERT [3] are not only effective in word-level tasks, but also in sentence-level tasks, because of their ability to capture contextual information and because they can be simply transfered to different downstream tasks. BERT uses the [CLS] token specifically to capture a sentence-level semantics. By extracting the representation of the [CLS] token from the output, we can obtain a sentence embedding that reflects the contextual information of the entire sentence. BERT's pretraining procedure includes two specific tasks: masked language modeling (MLM) and next sentence prediction (NSP). RoBERTa [4] builds upon BERT. It introduces dynamic masking and removes next sentence prediction, leading to improved performance and robustness.

Many sentence embedding methods opt to fine-tune BERT or RoBERTa using sentence-level pre-training tasks. Sentence-BERT (SBERT) [5] is typically based on the BERT architecture. Its fine-tuning task focuses on natural language inference (NLI), aiming to train a sentence embedding space that effectively captures semantic relationships between sentences. In [1], it is noted that methods based on natural language inference (NLI) datasets exhibit excellent performance in various downstream tasks. The authors argue that the sentence relationships captured in NLI include world knowledge that can improve language models.

Table 1
Examples from BATS

Animal	Sounds
bee	buzz/hum
dog	bark/growl/howl/yelp/whine/arf/woof
cat	meow/meu/purr/caterwaul
duck	quack

3. Definition Sentences from BATS (DSBATS)

In this section, we introduce how we extract sentence relationship data, where the relationships and sentences are from, and how to connect them together. The extraction result is DSBATS based on semantic networks (DSBATS-sn). We give statistics on DSBATS-sn in Table 2 and examples in Tables 3 and 4. The dataset is available. [1]

3.1. Relationship resource: word analogy

We use word analogy as the relationship resource. The most common example of word analogy is $king : queen :: man : woman,$ it states that *"king is to queen as man is to woman"*. Phenomena like this are to be studied as an important process of human cognition, with the development of language models, computational analogy has attracted more and more attention [6]. Mikolov [7] proposed to use the word offset technique to calculate this phenomenon with vectors corresponding to the words. That means, in an ideal word embedding space, the result of $\vec{king} - \vec{man} + \vec{woman}$ should be equal to $\vec{queen}$. This method is widely used as a benchmark to evaluate the quality of word embedding technique, and several word analogy test datasets have been proposed, like the Google analogy test set [7] and the Bigger Analogy Test Set (BATS) [8]. We choose BATS because it has fewer homonymy problems and various categories. BATS includes 40 morphological and semantic categories, each category can be regarded as a word analogy cluster. Example of a small word analogy cluster from BATS is shown in Table 1. Any two lines of words in the same cluster can form an analogical quadruple, like $bee : buzz :: dog : bark.$

There have also been past studies on constructing sentence relationships through word analogy, in [9]. They create general-purpose templates and replace a word that matches the word in the word analogy dataset in the template. For the sentence templates *"They traveled to **Havana**"* and *"They took a trip to **Cuba**"*, by replacing ***Havana-Cuba*** with capital-country word pairs found in word analogy datasets, a cluster of sentence pairs with similar relationships can be generated. The sentences generated by this method have the same sentence patterns. In fact, similar sentence patterns are not necessary to express similar semantics. Here by contrast, we construct sentences with semantic relationships that are not affected by sentence patterns.

[1] https://drive.google.com/drive/folders/DSBATS

Table 2
The size of categories of DSBATS-sn. The categories come from BATS. The sizes are the number of pairs of sentences.

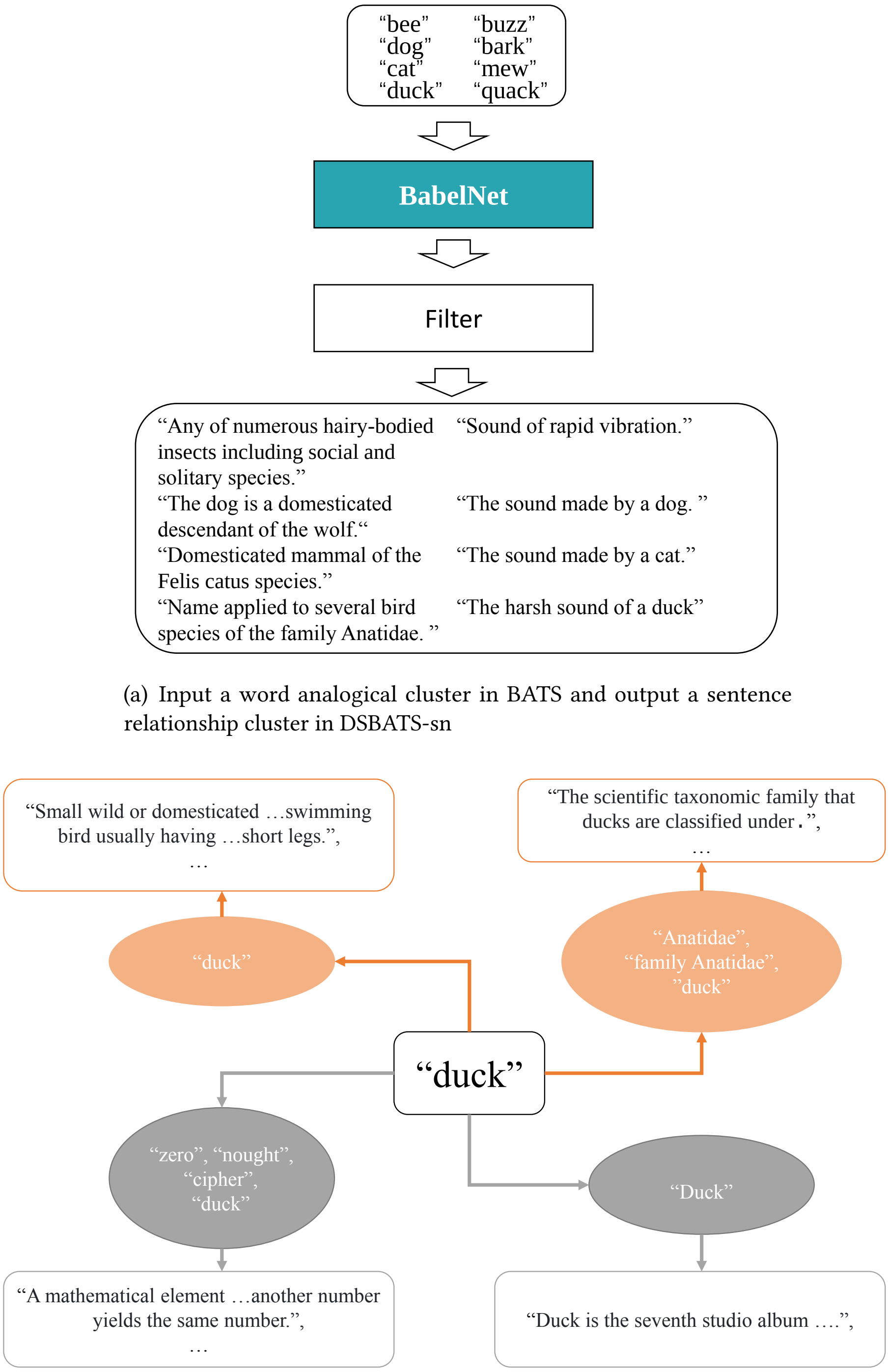
Encyclopedic	Size	Lexicographic	Size
E01 country - capital	447	L01 hypernyms - animals	4318
E02 country - language	669	L02 hypernyms - misc	5005
E03 UK city - county	426	L03 hyponyms - misc	6768
E04 name - nationality	570	L04 meronyms - substance	1312
E05 name - occupation	912	L05 meronyms -part	854
E06 animal - young	566	L06 meronyms - part	4036
E07 animal - sound	633	L07 synonyms - intensity	1645
E08 animal - shelter	877	L08 synonyms - exact	1307
E09 things - color	934	L09 antonyms - gradable	5560
E10 male - female	384	L10 antonyms - binary	1453

3.2. Sentences resource: semantic networks

We use semantic networks as the sentences resource. Semantic network is a kind of resource in a graphical form that shows the relationships between concepts or entities. In a semantic network, concepts are represented by nodes, and the relationships between those concepts are represented by edges that connect the nodes. BabelNet is a multilingual semantic network and ontology that provides a wide range of information about words and concepts in multiple languages [10]. BabelNet integrates information from a variety of sources, including WordNet [11], Wikipedia, and other lexical and semantic resources, it currently supports over 300 languages. A synset node in BabelNet includes its synonyms set, the part of speech, the domain category, the definition sentences, and other related information. The most important information for us is the definition sentences.

3.3. Extraction process

With BATS as our relationship resource and BabelNet as our sentence resource, we build the dataset **D**efinition **S**entences from **BATS** based on sementic network (DSBATS-sn). The extraction process is as the Figure 1(a). We input word analogy clusters into BabelNet, and BabelNet will deliver several synsets for each word in the clusters. In Figure 1(b), we have "duck" and "quack" as a pair in the cluster of animal:sound relationship, the search for"duck" in BabelNet delivers 35 different synsets, including synsets that do not conform to the animal:sound relationship like the synset with number 0 in math area. We use a filter to select the valid synsets that refer to the concept corresponding to the relationship. The filter takes advantage of the information contained in BabelNet, like the domain category or the part of speech, to select the synsets that match the relationships. In Figure 1(b), orange synsets are chosen, and gray synsets are discarded. The definition sentences in the valid synsets will be organized as sentence relationship clusters as the output part in Figure 1(a). In DSBATS-sn, there are 20 clusters corresponding to 20 different relationships, the size of different clusters (categories) are shown in Table 2. Some additional examples are shown in Tables 3 and 4.


"bee" "buzz"
"dog" "bark"
"cat" "mew"
"duck" "quack"
BabelNet
Filter
"Any of numerous hairy-bodied insects including social and solitary species."
"Sound of rapid vibration."
"The dog is a domesticated descendant of the wolf."
"The sound made by a dog. "
"Domesticated mammal of the Felis catus species."
"The sound made by a cat."
"Name applied to several bird species of the family Anatidae. "
"The harsh sound of a duck"
"Small wild or domesticated …swimming bird usually having …short legs.", …
"The scientific taxonomic family that ducks are classified under.", …
"duck"
"Anatidae", "family Anatidae", "duck"
"duck"
"zero", "nought", "cipher", "duck"
"Duck"
"A mathematical element …another number yields the same number.", …
"Duck is the seventh studio album ….",

(a) Input a word analogical cluster in BATS and output a sentence relationship cluster in DSBATS-sn

(b) Filter selection

Figure 1: The process for building DSBATS-sn

Table 3
Examples extracted from word pairs in L04 category in BATS, describe the relationship of things and their substance.

Word 1	Sentence 1	Word 2	Sentence 2
atmosphere	The gases surrounding the Earth or any astronomical body.	oxygen	Chemical element.
chocolate	Chocolate is a food product made from roasted and ground cacao seed kernels, that is available as a liquid, solid or paste, on its own or as a flavoring agent in other foods.	cocoa	Good, condiment, flavor, food ingredient or product solid derived from Theobroma cacao; precursor of commercial chocolates.
cocktail	Alcoholic mixed drink	water	Chemical compound; main constituent of the fluids of most living organisms.

4. Fine-tuning with DSBATS-sn

When using DSBATS-sn for fine-tuning, we aim to have similar relationships close to each other and different relations far away from each other in the embedding space. For example, the relationship between *"Domesticated mammal of the felis catus species"* and *"The sound made by a cat"*, and the relationship between *"The dog is a domesticated descendant of the wolf"* and *"The sound made by a dog"* are both the relationship of animal:sound, they are positive examples that should be close, and we can generate sentence pairs in another relationship sound:animal as negative examples by exchange the position in the pair. This requirement conforms to the basic idea of contrastive learning, which is to narrow the distance of relevant samples and push the distance of irrelevant samples in a certain feature space. Contrastive learning does not require very large-scale labeled data, and it can make the samples more uniformly distributed in the feature space [12]. We basically follow the contrastive learning framework and configuration of [13]. Following the idea of contrastive learning, we create negative examples through the operation above and get DSBATS for contrastive learning (DSBATS4CL). One example in DSBATS4CL includes 3 relationships, that is 6 sentences, as Table 5, the red one is is the negative pair. We have 2,244,530 such examples in DSBATS4CL for training.

Our loss is basiclly InfoNCE [14], in a batch of size S, the InfoNCE loss of the ith example x_i is:

$$\text{loss}_i = -\log\left(\frac{e^{\text{sim}(x_i, {x_i}^+)/\tau}}{\sum_{j=1}^{S} e^{\text{sim}(x_i, {x_j}^+)/\tau}}\right) \tag{1}$$

But there is a little difference, we only use the third pairs in the batch as negative examples, instead of using all the samples in the same batch except for x_i as negative examples for x_i, so

Table 4
Examples extracted from word pairs in E09 category in BATS, describe the relationship of things and their color.

Word 1	Sentence 1	Word 2	Sentence 2
tomato	The tomato is the edible berry of the plant Solanum lycopersicum, commonly known as the tomato plant.	red	Red color or pigment; the chromatic color resembling the hue of blood
potato	Annual native to South America having underground stolons bearing edible starchy tubers; widely cultivated as a garden vegetable; vines are poisonous.	brown	Brown can be considered a composite color but is mainly a darker shade of red.
grass	A very large and widespread family of Monocotyledoneae, with more than 10.000 species, most of which are herbaceous, but a few are woody. The stems are jointed, the long, narrow leaves originating at the nodes. The flowers are inconspicuous, with a much reduced perianth, and are wind-pollinated or cleistogamous.	green	A colour sometimes referred to as Luggage or Luggage Green

Table 5
DSBATS4CL example

Sentences	Relationship
Any of numerous hairy-bodied insects including social and solitary species. Sound of rapid vibration.	animal:sound
The dog is a domesticated descendant of the wolf. The sound made by a dog.	animal:sound
Sound of rapid vibration. Any of numerous hairy-bodied insects including social and solitary species.	sound:animal

our loss of x_i is:

$$\text{loss}_i = -\log\left(\frac{e^{\text{sim}(x_i, {x_i}^{+})/\tau}}{\sum_{j=1}^{S} e^{\text{sim}(x_i, {x_j}^{-})/\tau}}\right) \tag{2}$$

${x_j}^{-}$ corresponding to the red example in Table 5.

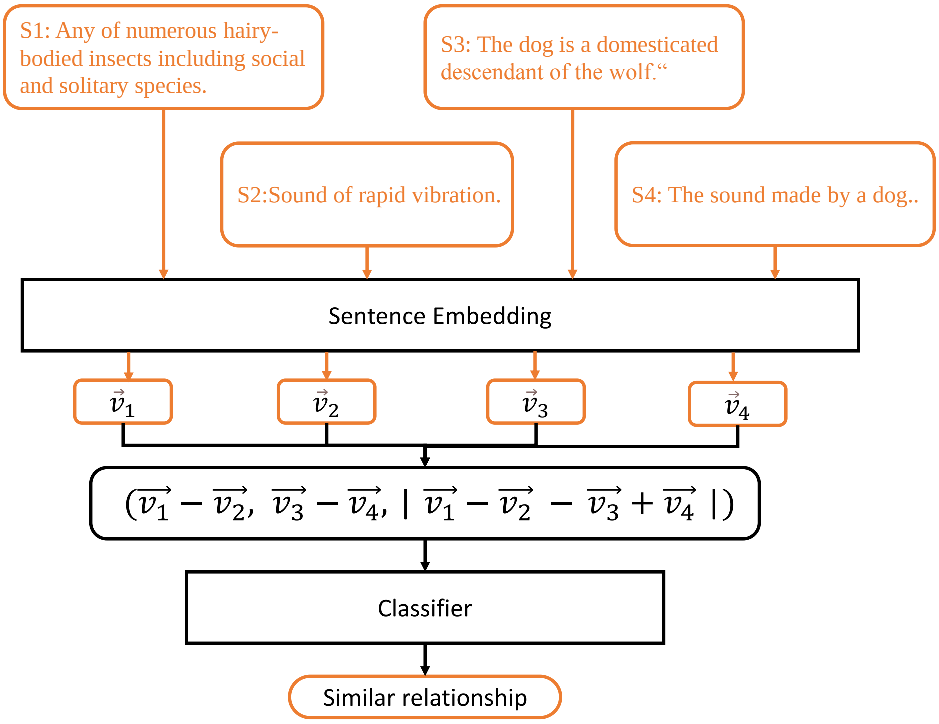

S1: Any of numerous hairy-bodied insects including social and solitary species.
S3: The dog is a domesticated descendant of the wolf.“
S2:Sound of rapid vibration.
S4: The sound made by a dog..
Sentence Embedding
$\vec{v}_1$
$\vec{v}_2$
$\vec{v}_3$
$\vec{v}_4$
$(\vec{v_1} - \vec{v_2},\ \vec{v_3} - \vec{v_4},\ |\ \vec{v_1} - \vec{v_2}\ -\ \vec{v_3} + \vec{v_4}\ |)$
Classifier
Similar relationship

Figure 2: Sentence relationship similarity distinguishing (SRSD) task

5. Experiment and evaluation

5.1. Intrinsic evaluation

We designed the **S**entence **R**elationship **S**imilarity **D**istinguishing (SRSD) task as intrinsic evaluation. It inputs a pair of relationships at a time, which is 4 sentences, and predicts if they are two similar relationships. Figure 2 shows the process with two sentence relationships in the same category. The test set is a manually annotated DSBATS dataset different from the automatically extracted DSBATS-sn in Section 3. The manual version uses the same relationship resource BATS but different sentence resources like Oxford Dictionary, Webster's Dictionary, and Collins Dictionary, so we call the manually annotated DSBATS as DSBATS-dic. The number of pairs of sentences in DSBATS-dic is shown in Table 6. After learning with DSBATS4CL, all three models improved accuracy on this task. The improvement is basically about 6%. The best performance is from Sentence-Bert with DSBATS4CL, which reaches 69.55%. Table 7 shows the result of SRSD task.

Table 6
Statistics on DSBATS-dic

Category	Size
L01 hypernyms - animals	251
L02 hypernyms - misc	225
L04 meronyms - substance	127

Table 7
Evaluation result of all the tasks.

		Intrinsic eval.	Extrinsic eval.	
Model	DSBATS4CL	SRSD	STS avg.	MRPC
BERT	w/o	58.18	18.63	68.81
	w/	**64.27**	**62.53**	**70.14**
RoBERTa	w/o	58.47	43.65	71.42
	w/	**65.83**	**65.11**	**71.83**
SBERT	w/o	61.68	62.84	73.51
	w/	**69.55**	**77.56**	**74.20**

5.2. Extrinsic evaluation

We conducted extrinsic evaluations using the Semantic Textual Similarity (STS) and Microsoft Research Paraphrase Corpus (MRPC) datasets as extrinsic evaluations. We use SentEval [15] to do the evaluation and follow the default configurations. The STS evaluation involves inputting two sentences and predicting a score between 0 and 5 that represents the similarity between the two sentences. Higher scores indicate better performance, as they align more closely with human-labeled similarity. The results, as summarized in Tables 7 and 8. They demonstrate the impact of fine-tuning with DSBATS4CL on the performance of the three pre-trained models. After fine-tuning with DSBATS4CL, the performance of all three pre-trained models improves. Notably, BERT and RoBERTa, which had not previously learned the relationship between sentences, improve by 43.89% and 20.02% in average, respectively. MRPC input two sentences and predict if they are similar or not. Higher scores correspond to higher accuracy. In comparison to the STS task, the improvements on MRPC are relatively small. The best performance is achieved by Sentence-BERT with DSBATS4CL, attaining accuracy of 66.06% on STS and 74.20% on MRPC, respectively. The results indicate that knowledge captured from sentence relationships derived from word analogy is valuable, fine-tuning with DSBATS4CL enhances the models' ability to understand the semantic relation between sentences.

6. Conclusion

In this work, we introduced a method to enhance sentence embedding using word analogy. We map the relationships between words to relationships between sentences by using definition

Table 8
Result on STS. STSB stands for STSBenchmark, SICK-R stands for SICK Relatedness. The last column is the same as in Table 7.

Model	DSBATS4CL	STS12	STS13	STS14	STS15	STS16	STSB	SICK-R	STS avg.
BERT	w/o	7.19	29.06	12.55	16.16	28.92	6.43	30.11	18.63
	w/	**54.29**	**68.37**	**58.54**	**67.23**	**69.21**	**60.80**	**59.25**	**62.53**
RoBERTa	w/o	16.73	45.56	30.24	55.27	56.87	39.14	61.76	43.65
	w/	**53.22**	**67.60**	**60.96**	**69.50**	**71.17**	**68.52**	**64.81**	**65.11**
SBERT	w/o	64.92	65.56	65.79	63.66	60.92	62.49	56.51	62.84
	w/	**71.85**	**82.21**	**79.85**	**82.44**	**77.67**	**77.54**	**71.35**	**77.56**

sentences in semantic network. Compared with the past methods that use word analogy in sentence-level tasks by replacing words in sentences, our method is less limited by morphology and pays more attention to semantics. The improvements on downstream tasks like STS and MRPC prove that the sentence relationships from word analogy include the knowledge that can enhance the semantic understanding of sentence embedding models. Sentence relationship similarity distinguishing task proposed as an intrinsic evaluation in our work can also be a challenging evaluation task for other sentence embedding methods. We believe that it is worth further exploring ways to combine analogy with contrastive learning, as analogy relation has many equivalent forms suitable for contrastive learning to construct positive and negative examples.

Acknowledgments

This work has been supported in part by a research grant from JSPS Kakenhi Kiban C n° 21K12038 entitled "Theoretically founded algorithms for the automatic production of analogy test sets in NLP."

References

[1] R. Li, X. Zhao, M.-F. Moens, A brief overview of universal sentence representation methods: A linguistic view, ACM Computing Surveys (CSUR) 55 (2022) 1–42.

[2] S. R. Bowman, G. Angeli, C. Potts, C. D. Manning, A large annotated corpus for learning natural language inference, in: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Lisbon, Portugal, 2015, pp. 632–642. URL: https://aclanthology.org/D15-1075. doi:10.18653/v1/D15-1075.

[3] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, 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, Volume 1 (Long and Short Papers), Association for Computational

Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186. URL: https://aclanthology.org/N19-1423. doi:10.18653/v1/N19-1423.

[4] L. Zhuang, L. Wayne, S. Ya, Z. Jun, A robustly optimized BERT pre-training approach with post-training (2021) 1218–1227. URL: https://aclanthology.org/2021.ccl-1.108.

[5] N. Reimers, I. Gurevych, 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), 2019, pp. 3982–3992.

[6] H. Prade, G. Richard, Computational Approaches to Analogical Reasoning: Current Trends, volume 548, Springer Publishing Company, Incorporated, 2014.

[7] T. Mikolov, K. Chen, G. S. Corrado, J. Dean, Efficient estimation of word representations in vector space, in: International Conference on Learning Representations, 2013.

[8] A. Gladkova, A. Drozd, S. Matsuoka, Analogy-based detection of morphological and semantic relations with word embeddings: what works and what doesn't., in: Proceedings of the NAACL Student Research Workshop, Association for Computational Linguistics, San Diego, California, 2016, pp. 8–15. URL: https://aclanthology.org/N16-2002. doi:10.18653/v1/N16-2002.

[9] X. Zhu, G. de Melo, Sentence analogies: Linguistic regularities in sentence embeddings, in: Proceedings of the 28th International Conference on Computational Linguistics, 2020, pp. 3389–3400.

[10] R. Navigli, S. P. Ponzetto, Babelnet: Building a very large multilingual semantic network, in: Proceedings of the 48th annual meeting of the association for computational linguistics, 2010, pp. 216–225.

[11] G. A. Miller, Wordnet: a lexical database for english, Communications of the ACM 38 (1995) 39–41.

[12] T. Wang, P. Isola, Understanding contrastive representation learning through alignment and uniformity on the hypersphere, in: International Conference on Machine Learning, PMLR, 2020, pp. 9929–9939.

[13] T. Gao, X. Yao, D. Chen, SimCSE: Simple contrastive learning of sentence embeddings, in: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 2021, pp. 6894–6910. URL: https://aclanthology.org/2021.emnlp-main.552. doi:10.18653/v1/2021.emnlp-main.552.

[14] A. van den Oord, Y. Li, O. Vinyals, Representation learning with contrastive predictive coding, CoRR abs/1807.03748 (2018). URL: http://arxiv.org/abs/1807.03748. arXiv:1807.03748.

[15] A. Conneau, D. Kiela, SentEval: An evaluation toolkit for universal sentence representations, in: Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), European Language Resources Association (ELRA), Miyazaki, Japan, 2018. URL: https://aclanthology.org/L18-1269.