Just Rank: Rethinking Evaluation with Word and Sentence Similarities

Bin Wang[†], **C.-C. Jay Kuo**[§], **Haizhou Li**[‡,†,♯]
[†] National University of Singapore, Singapore
[§] University of Southern California, USA
[‡] The Chinese University of Hong Kong, Shenzhen, China [♯] Kriston AI, China
bwang28c@gmail.com

Abstract

Word and sentence embeddings are useful feature representations in natural language processing. However, intrinsic evaluation for embeddings lags far behind, and there has been no significant update since the past decade. Word and sentence similarity tasks have become the *de facto* evaluation method. It leads models to overfit to such evaluations, negatively impacting embedding models' development. This paper first points out the problems using semantic similarity as the gold standard for word and sentence embedding evaluations. Further, we propose a new intrinsic evaluation method called *EvalRank*, which shows a much stronger correlation with downstream tasks. Extensive experiments are conducted based on 60+ models and popular datasets to certify our judgments. Finally, the practical evaluation toolkit is released for future benchmarking purposes.[1]

1 Introduction

Distributed representation of words (Bengio et al., 2003; Mikolov et al., 2013a; Pennington et al., 2014; Bojanowski et al., 2017) and sentences (Kiros et al., 2015; Conneau et al., 2017; Reimers and Gurevych, 2019; Gao et al., 2021) have shown to be extremely useful in transfer learning to many NLP tasks. Therefore, it plays an essential role in how we evaluate the quality of embedding models. Among many evaluation methods, the word and sentence similarity task gradually becomes the de facto intrinsic evaluation method.

Figure 1 shows examples from word and sentence similarity datasets. In general, the datasets consist of pairs of words (w_1, w_2) (or sentences) and human-annotated similarity scores S_h. To evaluate an embedding model $\phi(\cdot)$, we first extract embeddings for (w_1,w_2): $(\mathbf{e}_1, \mathbf{e}_2) = (\phi(w_1), \phi(w_2))$.

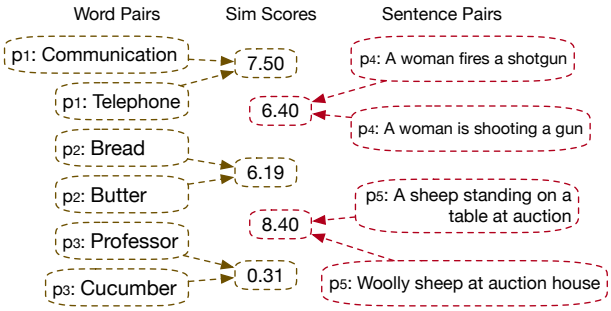

Figure 1: Word and sentence pairs with human-annotated similarity scores from WS-353 and STS-B datasets (scaled to range 0 (lowest) to 10 (highest)).

Then, a similarity measure is applied to compute an predicted score $S_p = sim(\mathbf{e}_1, \mathbf{e}_2)$, where cosine similarity is adopted as sim unquestionably in the majority of cases. Finally, the correlation between S_h and S_p is computed, and a higher correlation suggests good alignment with human annotations and a better embedding model.

Many studies, especially those targeting on information retrieval via semantic search and clustering (Reimers and Gurevych, 2019; Su et al., 2021), have used the similarity task as the only or main evaluate method (Tissier et al., 2017; Mu et al., 2018; Arora et al., 2017; Li et al., 2020; Gao et al., 2021). We observe a number of issues in word or sentence similarity tasks ranging from dataset collection to the evaluation paradigm, and consider that focusing too much on similarity tasks would negatively impact the development of future embedding models.

The significant concerns are summarized as follows, which generally apply to both word and sentence similarity tasks. First, the definition of similarity is too vague. There exist complicated relationships between sampled data pairs, and almost all relations contribute to the similarity score, which is challenging to non-expert annotators. Second, the similarity evaluation tasks are not directly

[1] Available at https://github.com/BinWang28/EvalRank-Embedding-Evaluation.

Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 6060 - 6077

relevant to the downstream tasks. We believe it is because of the data discrepancy between them, and the properties evaluated by similarity tasks are not the ones important to downstream applications. Third, the evaluation paradigm can be tricked with simple post-processing methods, making it unfair to benchmark different models.

Inspired by Spreading-Activation Theory (Collins and Loftus, 1975), we propose to evaluate embedding models as a retrieval task, and name it as *EvalRank* to address the above issues. While similarity tasks measure the distance between similarity pairs from all similarity levels, *EvalRank* only considers highly similar pairs from a local perspective.

Our main contributions can be summarized as follows:

1 We point out three significant problems for using word and sentence similarity tasks as the de facto evaluation method through analysis or experimental verification. The study provides valuable insights into embeddings evaluation methods.

2 We propose a new intrinsic evaluation method, *EvalRank*, that aligns better with the properties required by various downstream tasks.

3 We conduct extensive experiments with 60+ models and 10 downstream tasks to certify the effectiveness of our evaluation method. The practical evaluation toolkit is released for future benchmarking purposes.

2 Related Work

Word embedding has been studied extensively, and popular work (Mikolov et al., 2013a; Pennington et al., 2014; Bojanowski et al., 2017) are mainly built on the distributional hypothesis (Harris, 1954), where words that appear in the same context tend to share similar meanings. The early work on sentence embedding are either built upon word embedding (Arora et al., 2017; Rücklé et al., 2018; Almarwani et al., 2019) or follow the distributional hypothesis on a sentence level (Kiros et al., 2015; Hill et al., 2016; Logeswaran and Lee, 2018). Recent development of sentence embedding are incorporating quite different techniques including multi-task learning (Cer et al., 2018), supervised inference data (Conneau et al., 2017; Reimers and Gurevych, 2019), contrastive learning (Zhang et al., 2020; Carlsson et al., 2020; Yan et al., 2021; Gao et al., 2021) and pre-trained language models (Li et al., 2020; Wang and Kuo, 2020; Su et al., 2021). Nonetheless, even though different methods choose different evaluation tasks, similarity task is usually the shared task for benchmarking purposes.

Similarity task is originally proposed to mimic human's perception about the similarity level between word or sentence pairs. The first word similarity dataset was collected in 1965 (Rubenstein and Goodenough, 1965), which consists of 65 word pairs with human annotations. It has been a standard evaluation paradigm to use cosine similarity between vectors for computing the correlation with human judges (Agirre et al., 2009). Many studies raise concerns about such evaluation paradigm. Faruqui et al. (2016) and Wang et al. (2019b) points out some problems with word similarity tasks, including low correlation with downstream tasks and lack of task-specific similarity. Reimers et al. (2016), Eger et al. (2019) and Zhelezniak et al. (2019) states current evaluation paradigm for Semantic Textual Similarity (STS) tasks are not ideal. One most recent work (Abdalla et al., 2021) questions about the data collection process of STS datasets and creates a new semantic relatedness dataset (STR) by comparative annotations (Louviere and Woodworth, 1991).

There are also other intrinsic evaluation methods for word and sentence embedding evaluation, but eventually did not gain much popularity. Word analogy task is first proposed in (Mikolov et al., 2013a,c) to detect linguistic relations between pairs of word vectors. Zhu and de Melo (2020) recently expanded the analogy concept to sentence level. However, the analogy task is more heuristic and fragile as an evaluation method (Gladkova et al., 2016; Rogers et al., 2017). Recently, probing tasks have been proposed to measure intriguing properties of sentence embedding models without worrying much about practical applications (Zhu et al., 2018; Conneau et al., 2018; Barancíková and Bojar, 2019). Because of the lack of effective intrinsic evaluation methods, Reimers and Gurevych (2019) and Wang et al. (2021) seeks to include more domain-specific tasks for evaluation.

3 Problems with Similarity Tasks

In this work, we discuss the problems of similarity tasks both on word and sentence levels. They are highly similar from data collection to evaluation

paradigm and are troubled by the same problems.

3.1 Multifaceted Relationships

First, the concept of similarity and relatedness are not well-defined. Similar pairs are related but not vise versa. Taking synonym, hypernym, and antonym relations as examples, the similarity rank should be "synonym > hypernym > antonym" while the relatedness rank should be "synonym > hypernym ≈ antonym". This was not taken into consideration when constructing datasets. Agirre et al. (2009) intentionally split one word similarity dataset into similarity and relatedness subsets. However, we find that obtained subsets are erroneous towards polysemy, and the relatedness between pair ('stock', 'egg', 1.81) is much lower than pair ('stock', 'oil', 6.34). It is because only the 'financial stock market' is compared but not the 'stock of supermarkets'. Furthermore, relationships between samples are far more complicated than currently considered, which is a challenge to all current datasets.

Second, the annotation process is not intuitive to humans. The initial goal of the similarity task is to let the model mimic human perception. However, we found that the instructions on similarity levels are not well defined. For example, on STS 13∼16 datasets, annotators must label sentences that 'share some details' with a score of 2 and 'on the same topic' with a score of 1. According to priming effect theory, (Meyer and Schvaneveldt, 1971; Weingarten et al., 2016), humans are more familiar with ranking several candidate samples based on one pivot sample (priming stimulus). Therefore, a more ideal way of annotation is to give one pivot sample (e.g. 'cup') and rank candidates with different similarity levels (e.g. 'trophy', 'tableware', 'food', 'article', 'cucumber'). In other words, it is more intuitive for human to compare (a,b) > (a,c) than (a,b) > (c,d) as far as similarity is concerned. However, in practice, it is hard to collect a set of candidates for each pivot sample, especially for sentences.

3.2 Weak Correlation with Downstream Tasks

In previous studies, it was found that the performance of similarity tasks shows little or negative correlation with the performance of downstream tasks (Faruqui et al., 2016; Wang et al., 2019b, 2021). An illustration is shown in Table 1a. We think there are two reasons behind 1) low testing corpus overlap and 2) mismatch of tested properties.

Score (rank)	STS-B	SST2	MR
GloVe	47.95 (4)	79.52 (6↓)	77.54 (5↓)
InferSent	70.94 (3)	83.91 (3)	77.61 (4↓)
BERT-cls	20.29 (6)	86.99 (1↑)	80.99 (1↑)
BERT-avg	47.29 (5)	85.17 (2↑)	80.05 (2↑)
BERT-flow	71.76 (2)	80.67 (4↓)	77.01 (6↓)
BERT-whitening	71.79 (1)	80.23 (5↓)	77.96 (3↓)

(a)

Rank	cos	l_2
SBERT	1	2↓
SimCSE	2	1↑
BERT-avg	5	3↑
BERT-flow	4	4
BERT-whitening	3	5↓

(b)

Table 1: (a) Performance scores and rank of embedding models on STS-B, SST2, and MR tasks. (b) Performance rank of models on STS-B testset with cos and l_2 similarity metrics.

First, similarity datasets have their data source and are not necessarily close to the corpus of downstream tasks. For example, Baker et al. (2014) collect word pairs for verbs only while Luong et al. (2013) intentionally test on rare words. Also, for STS datasets, (Agirre et al., 2012) annotates on sentence pairs from paraphrases, video captions, and machine translations, which has limited overlap on downstream tasks like sentiment classification.

Second, the original goal for the similarity task is to mimic human perceptions. For example, STS datasets are originally proposed as a competition to find the most effective STS systems instead of a gold standard for generic sentence embedding evaluation. Some properties evaluated by similarity tasks are trivial to downstream tasks, and it is more important to test on mutually important ones. As examples in Figure 1, the similarity tasks inherently require the model to predict sim(p_1)>sim(p_2) and sim(p_5)>sim(p_4), which we believe are unnecessary for most downstream applications. Instead, similar pairs are more important than less similar pairs for downstream applications (Kekäläinen, 2005; Reimers et al., 2016). Therefore, it is enough for good embedding models to focus on gathering similar pairs together while keeping dissimilar ones far away to a certain threshold.

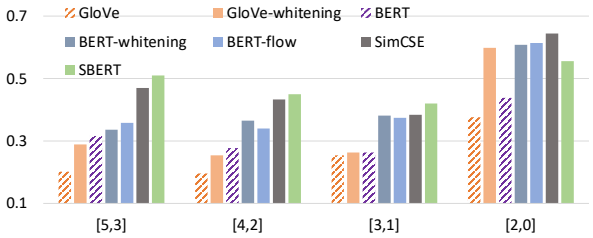

Figure 2: Performance of embedding models on the whole STS-Benchmark dataset w.r.t different similarity levels.

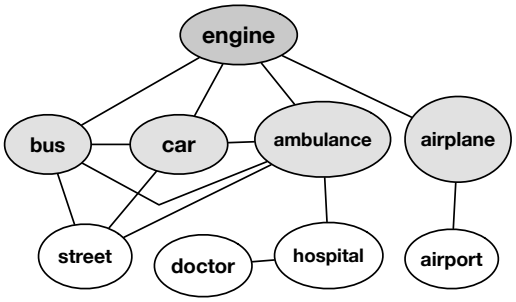

Figure 3: Example of Concept Network in SAT

3.3 Overfitting

As similarity tasks become one de facto evaluation method for embedding models, recent work tend to overfit the current evaluation paradigm, including the choice of similarity measure and the post-processing step.

Similarity Metrics. Cosine similarity is the default choice for similarity tasks. However, simply changing the similarity metric to other commonly used ones can lead to contradictory results.

In Table 1b, we compare recent five BERT-based sentence embedding models including BERT (Devlin et al., 2019), BERT-whitening (Su et al., 2021), BERT-flow (Li et al., 2020), SBERT (Reimers and Gurevych, 2019) and SimCSE (Gao et al., 2021).[2] The results on standard STS-Benchmark testset are reported under both cosine and l_2 similarity. As we can see, the performance rank differs under different similarity metrics. This is especially true for BERT-flow and BERT-whitening, which do not even outperform their baseline models when evaluating with l_2 metric. Therefore, we can infer that some models overfit to the default cosine metric for similarity tasks.

Whitening Tricks. A number of studies attempted the post-processing of word embeddings (Mu et al., 2018; Wang et al., 2019a; Liu et al., 2019b) and sentence embeddings (Arora et al., 2017; Liu et al., 2019a; Li et al., 2020; Su et al., 2021). The shared concept is to obtain a more isotropic embedding space (samples evenly distributed across directions) and can be summarized as a space whitening process. Even though the whitening tricks help a lot with similarity tasks, we found it is usually not applicable to downstream tasks or even hurt the model performance.[3] We think the whitening methods are overfitted to similarity tasks and would like to find the reasons behind.

First, we take the whole STS-Benchmark dataset and create subsets of sentence pairs from certain similarity levels. We test on two baseline sentence embedding models: GloVe, BERT; three whitening tricks: ABTT on GloVe (Mu et al., 2018), BERT-whitening, BERT-flow; two strong sentence embedding models that perform well on both STS and downstream tasks: SBERT, SimCSE. Figure 2 shows the result, and we can see that the whitening-based methods are boosting the baseline performance mainly for less similar pairs (e.g., pairs with a similarity score within [2,0]). In contrast, the models that perform well on downstream tasks show consistent improvement on all subsets with different similarity scores. As discussed in Section 3.2, highly similar pairs are more critical than less similar pairs for downstream tasks. Since the post-processing methods mainly help with less similar pairs, they do not help much on downstream tasks.

4 Evaluation by Ranking

4.1 Theory and Motivations

In cognitive psychology, Spreading-Activation Theory (SAT) (Collins and Loftus, 1975; Anderson, 1983) is to explain how concepts store and interact within the human brain. Figure 3 shows one example about the concept network. In the network, only highly related concepts are connected. To find the relatedness between concepts like *engine* and *street*, the activation is spreading through mediating concepts like *car* and *ambulance* with decaying factors. Under this theory, the similarity task is measuring the association between any two concepts in the network, which requires complicated long-distance activation propagation. Instead, to test the soundness of the concept network, it is enough to ensure the local connectivity between concepts. Moreover, the long-distance relationships can be inferred thereby with various spreading activation

[2] Experimental details in Appendix B.
[3] Analysis in Appendix C.1.

	Type	# pos pairs	# background samples	Source
EvalRank	Word	5,514	22,207	Word Similarity Datasets & Wiki
	Sent	6,989	24,957	STS-Benchmark & STR

Table 2: Statistics of *EvalRank* Datasets

algorithms (Cohen and Kjeldsen, 1987).

Therefore, we propose *EvalRank* to test only on highly related pairs and make sure they are topologically close in the embedding space. It also alleviates the problems of similarity tasks. First, instead of distinguishing multifaceted relationships, we only focus on highly related pairs, which are intuitive to human annotators. Second, it shows a much stronger correlation with downstream tasks as desired properties are measured. Third, as we treat the embedding space from a local perspective, it is less affected by the whitening methods.

4.2 Methodology

We frame the evaluation of embeddings as a retrieval task. To this purpose, the dataset of *EvalRank* contains two sets: 1) the positive pair set $P = \{p_1, p_2, ..., p_m\}$ and 2) the background sample set $C = \{c_1, c_2, ..., c_n\}$. Each positive pair $p_i = (c_x, c_y)$ in P consists of two samples in C that are semantically similar.

For each sample (c_x) and its positive correspondence (c_y), a good embedding model should has their embeddings ($\mathbf{e_x}$, $\mathbf{e_y}$) close in the embedding space. Meantime, the other background samples should locate farther away from the sample c_x. Some samples in the background may also be positive samples. We assume it barely happens and is negligible if good datasets are constructed.

Formally, given an embedding model $\phi(\cdot)$, the embeddings for all samples in C are computed as $\{\mathbf{e_1}, \mathbf{e_2}, ..., \mathbf{e_n}\} = \{\phi(c_1), \phi(c_2), ..., \phi(c_n)\}$. The cos similarity and l_2 similarity between two samples (c_x, c_y) are defined as:

$$S_{cos}(c_x, c_y) = \frac{\mathbf{e}_x^T \mathbf{e}_y}{||\mathbf{e}_x|| \cdot ||\mathbf{e}_y||}$$

$$S_{l_2}(c_x, c_y) = \frac{1}{1 + ||\mathbf{e}_x - \mathbf{e}_y||}$$

Further, the similarity score is used to sort all background samples in descending order and the performance at each positive pair p_i is measured by the rank of c_x's positive correspondence c_y w.r.t all background samples:

$$rank_i = rank(S(c_x, c_y), [||_{j=1, j\neq x}^{n} S(c_x, c_j)])$$

where $||$ refers to the concatenation operation. To measure the overall performance of model $\phi(\cdot)$ on all positive pairs in P, the mean reciprocal rank (MRR) and Hits@k scores are reported and a higher score indicates a better embedding model:

$$MRR = \frac{1}{m} \sum_{i=1}^{m} \frac{1}{rank_i}$$

$$Hits@k = \frac{1}{m} \sum_{i=1}^{m} \mathbb{1}[rank_i \leq k]$$

Note that there are two similarity metrics, and we found that S_{cos} shows a better correlation with downstream tasks while S_{l_2} is more robust to whitening methods. We use S_{cos} in the experiments unless otherwise specified.

4.3 Dataset Collection

Word-Level. We collect the positive pairs from 13 word similarity datasets (Wang et al., 2019b). For each dataset, the pairs with the highest 25% similarity score are gathered as positive pairs. Background word samples contain all words that appear in the similarity datasets. Further, we augment the background word samples using the most frequent 20,000 words from Wikipedia corpus.

Sentence-Level. Similarly, the pairs with top 25% similarity/relatedness score from STS-Benchmark dataset (Cer et al., 2017) and STR dataset (Abdalla et al., 2021) are collected as positive pairs. All sentences that appear at least once are used as the background sentence samples.

In both cases, if positive pair (c_x, c_y) exists, the reversed pair (c_y, c_x) is also added as positive pairs. Detailed statistics of *EvalRank* datasets are listed in Table 2.

4.4 Alignment and Uniformity

Recently, Wang and Isola (2020) identifies the alignment and uniformity properties as an explanation to the success of contrastive loss. It shares many similarities with our method and can also shed light on why *EvalRank* works. First, the alignment property requires similar samples to have similar features, which aligns with the objective of

		SCICITE	MR	CR	MPQA	SUBJ	SST2	SST5	TREC	MRPC	SICK-E
WS-353-All		62.87	43.68	40.94	37.50	15.57	41.65	45.03	34.70	8.98	57.96
WS-353-Rel		66.13	47.92	45.15	41.77	11.65	47.25	48.18	26.36	20.56	61.83
WS-353-Sim		67.86	45.94	43.97	38.68	17.41	44.03	50.32	34.85	10.67	56.13
RW-STANFORD		75.56	74.65	55.35	66.08	46.82	81.50	68.25	45.91	13.08	43.29
MEN-TR-3K		66.91	44.15	45.37	39.14	1.70	38.51	42.11	22.82	28.63	**71.26**
MTURK-287		68.48	65.95	48.01	52.36	31.94	71.96	58.01	29.22	7.54	36.23
MTURK-771		79.93	60.87	49.45	57.92	24.04	62.75	62.03	29.14	17.44	60.23
SIMLEX-999		68.20	48.02	40.90	46.43	19.03	47.30	50.95	38.14	15.32	60.26
SIMVERB-3500		65.13	45.60	36.95	47.04	21.57	45.16	48.56	41.74	10.70	58.08
EvalRank	MRR	89.96	87.91	68.23	78.03	51.35	91.54	83.36	48.15	25.70	61.34
	Hits@1	85.91	83.69	66.93	81.43	**55.95**	89.74	79.46	43.53	28.82	53.86
	Hits@3	**90.11**	**88.82**	**69.92**	**82.05**	54.52	**93.32**	**84.41**	**48.44**	**30.87**	62.77

Table 3: Spearman's rank correlation ($\rho \times 100$) between performance scores of word-level intrinsic evaluation and downstream tasks, where the best is marked with **bold** and second best with underline.

EvalRank. Second, the uniformity property is measured by the average Gaussian distance between any two samples. In contrast, *EvalRank* focuses on the distance between points from a local perspective and would require the pivot sample to have longer distances to any background samples than its positive candidate. Measuring the distance from a local perspective has unique advantages because the learned embedding space will likely form a manifold and can only approximates euclidean space locally. Therefore, simple similarity metrics like cos or l_2 are not suitable to model long-distance relationships.

4.5 Good Intrinsic Evaluator

A good intrinsic evaluator can test the properties that semantically similar samples are close in vector space (Reimers and Gurevych, 2019; Gao et al., 2021) and serve as prompt information to real-world applications. As *EvalRank* directly test on the first property, we design experiments to show the correlation with various downstream tasks as a comparison of intrinsic evaluators. To be comprehensive, we first collect as many embedding models as possible and test them on the intrinsic evaluator and downstream task. The Spearman's rank correlation is computed between the results, and a higher score indicates better correlation with downstream tasks and better intrinsic evaluator.

Meantime, we do not think similarity evaluations should be discarded, even though it fails to correlate well with downstream applications. It has its advantages as aiming to mimic human perception about semantic-related pairs.

5 Word-Level Experiments

5.1 Experimental Setup

Word Embedding Models. We collect 19 word embedding models from GloVe (Pennington et al., 2014), word2vec (Mikolov et al., 2013b), fastText (Bojanowski et al., 2017), Dict2vec (Tissier et al., 2017) and PSL (Wieting et al., 2015). Meantime, we apply ABTT (Mu et al., 2018) post-processing to all models to double the total number of embedding models. When testing on downstream tasks, the simplest bag-of-words feature is used as sentence representations in order to focus on measuring the quality of word embeddings.

Word Similarity Tasks. 9 word similarity datasets are compared as the baseline methods including WS-353-All (Finkelstein et al., 2001), WS-353-Rel (Agirre et al., 2009), WS-353-Sim (Agirre et al., 2009), RW-STANFORD (Luong et al., 2013), MEN-TR-3K (Bruni et al., 2014), MTURK-287 (Radinsky et al., 2011), MTURK-771 (Halawi et al., 2012), SIMLEX-999 (Hill et al., 2015), SIMVERB-3500 (Gerz et al., 2016). The word similarity datasets with less than 200 pairs are not selected to avoid evaluation occasionality. Cosine similarity and Spearman's rank correlation are deployed for all similarity tasks.

Downstream Tasks. SentEval (Conneau and Kiela, 2018) is a popular toolkit in evaluating sentence embeddings. We use 9 downstream tasks from SentEval including MR (Pang and Lee, 2005), CR (Hu and Liu, 2004), MPQA (Wiebe et al., 2005), SUBJ (Pang and Lee, 2004), SST2 (Socher et al., 2013), SST5 (Socher et al., 2013), TREC (Li and Roth, 2002), MRPC (Dolan et al., 2004), SICK-E (Marelli et al., 2014). Previous work spot

	SCICITE	MR	SST2
EvalRank	89.96	87.91	91.54
w/o wiki vocabs	88.55	83.99	88.26
w/ WN synonym	90.56	86.56	91.12
w/ l_2 metric	77.47	78.34	81.51

Table 4: Ablation study on variants of *EvalRank*. Spearman's rank correlation ($\rho \times 100$) between MRR scores and downstream task scores are reported.

EvalRank	MRR	Hits@1	Hits@3
GloVe	13.15	4.66	15.72
word2vec	12.88	4.57	14.35
fastText	**17.22**	**5.77**	**19.99**
Dict2vec	12.71	4.03	13.04

(a) Word-Level

EvalRank	MRR	Hits@1	Hits@3
GloVe	61.00	44.94	74.66
InferSentv1	60.72	41.92	77.21
InferSentv2	63.89	45.59	80.47
BERT-first-last-avg	68.01	51.70	81.91
BERT-whitening	66.58	46.54	84.22
SBERT	64.12	47.07	79.05
SimCSE	**69.50**	**52.34**	**84.43**

(b) Sentence-Level

Table 5: Benchmarking results on *EvalRank*. Performance is reported as % ($\times$100).

that SentEval tasks are biased towards sentiment analysis (Wang et al., 2018). Therefore, we add one extra domain-specific classification task SCICITE (Cohan et al., 2019) which assigns intent labels (background information, method, result comparison) to sentences collected from scientific papers that cite other papers. For all tasks, a logistic regression classifier is used with cross-validation to predict the class labels.

5.2 Results and Analysis

Table 3 shows the word-level results. In short, *EvalRank* outperforms all word similarity datasets with a clear margin. For evaluation metrics, we can see that Hits@3 score shows a higher correlation than MRR and Hits@1 scores. However, the gap between the evaluation metrics is not big, which makes them all good measures. Among all 10 downstream tasks, *EvalRank* shows a strong correlation (ρ>0.6) with 7 tasks and a very strong correlation (ρ>0.8) with 5 tasks. While, among all word similarity datasets, only one dataset (RW-STANFORD) shows a strong correlation with one downstream task (SST2).

For word similarity datasets, RW-STANFORD dataset shows the best correlation with downstream tasks. It confirms the finding in Wang et al. (2019b) that this dataset contains more high-quality and low-frequency word pairs.

Ablation Study. We experiment with several variants of our *EvalRank* method and the result is shown in Table 4. First, if we do not augment the background word samples with the most frequent 20,000 words from the Wikipedia corpus, it leads to certain performance downgrading. Without sufficient background samples, positive pairs are not challenging enough to test each model's capability. Second, we tried to add more positive samples (e.g. 5k samples) using synonym relations from WordNet (WN) database (Miller, 1998). However, no obvious improvement is witnessed because the synonym pairs in WN contain too many noisy pairs. Last, for similarity measures, we notice that *cos* similarity is consistently better than l_2 similarity while both outperform word similarity baselines.

Benchmarking Results. In Table 5a, we compared four popular word embedding models, including GloVe, word2vec, fastText, and Dict2vec, where fastText achieves the best performance.

6 Sentence-Level Experiments

6.1 Experimental Setup

Sentence Embedding Models. We collect 67 embedding models, where 38 of them are built upon word embeddings with bag-of-words features and 29 of them are neural-network-based models. For neural-network-based models, we collect variants from InferSent (Conneau et al., 2017), BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2020), BERT-flow (Li et al., 2020), BERT-whitening (Su et al., 2021), SBERT (Reimers and Gurevych, 2019) and SimCSE (Gao et al., 2021).

Sentence Similarity Tasks. We evaluate on 7 standard semantic textual similarity datasets including STS12$\sim$16 (Agirre et al., 2012, 2013, 2014, 2015, 2016), STS-Benchmark (Cer et al., 2017) and SICK-Relatedness (Marelli et al., 2014). Recently, Abdalla et al. (2021) questioned the labeling process of STS datasets and released a new semantic textual relatedness (STR) dataset, which is also included in our experiments.

Downstream Tasks. We use 7 classification tasks

		SCICITE	MR	CR	MPQA	SUBJ	SST2	SST5	TREC
STS12		32.96	38.62	44.77	31.52	21.76	33.79	35.68	30.79
STS13		22.04	32.62	41.23	12.39	7.64	26.45	22.98	12.16
STS14		25.91	34.77	41.89	19.23	10.13	29.20	26.82	17.70
STS15		31.84	40.64	48.11	25.12	16.48	35.50	33.30	24.70
STS16		29.56	40.14	51.66	14.35	16.53	33.61	29.44	21.43
STS-Benchmark		32.99	46.03	52.78	21.09	26.47	40.41	36.75	34.64
SICK-Relatedness		40.38	38.51	50.68	29.87	18.87	34.54	36.73	25.25
STR		-14.48	-8.38	-7.79	-29.57	-23.91	-16.33	-22.77	-14.30
EvalRank	MRR	65.95	83.43	87.08	43.93	72.72	80.97	74.16	76.74
	Hits@1	**69.01**	**85.39**	**89.36**	**45.81**	**74.93**	**82.65**	**76.65**	**78.72**
	Hits@3	63.35	83.92	85.43	41.24	70.98	80.36	72.05	74.70

Table 6: Spearman's rank correlation ($\rho \times 100$) between performance scores of sentence-level intrinsic evaluation and downstream tasks, where the best is marked with **bold** and second best with underline.

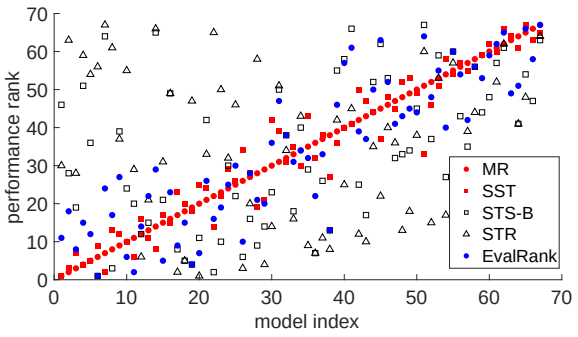

Figure 4: Visualization of models' performance rank on 2 downstream tasks and 3 intrinsic evaluation methods.

		SCICITE	MR	SST2
EvalRank (STS-B + STR)	MRR	65.95	83.43	80.97
	Hits@1	69.01	85.39	82.65
	Hits@3	63.35	83.92	80.36
EvalRank (STS-B)	MRR	63.05	75.85	72.87
	Hits@1	66.22	77.94	75.20
	Hits@3	61.23	75.49	72.92
EvalRank (STR)	MRR	63.51	83.28	80.20
	Hits@1	66.59	84.53	82.14
	Hits@3	60.68	82.55	79.42

Table 7: Performance under different data sources. Spearman's rank correlation ($\rho \times 100$) is reported.

from SentEval evaluation toolkit, including MR, CR, MPQA, SUBJ, SST2, SST5, TREC, as well as the domain-specific classification task SCICITE. We exclude the MRPC and SICK-E because they are highly similar with STS tasks (Conneau and Kiela, 2018).

6.2 Results and Analysis

Table 6 shows the sentence-level results. *EvalRank* outperform all sentence similarity datasets with a clear margin. For evaluation metric, Hits@1 shows a higher correlation comparing with MRR and Hits@3. Among all 7 downstream tasks, *EvalRank* shows strong correlation ($\rho > 0.6$) with 6 tasks.

For sentence similarity datasets, no one clearly outperforms others. Additionally, we found that STR dataset shows the worst correlation with downstream tasks. Even though STR adopts a better data annotation schema than STS datasets, it still follows the previous standard evaluation paradigm and is exposed to the same problems. It further verifies our discussion about problems with sentence similarity evaluation.

Correlation Visualization. Figure 4 shows the performance rank of 67 sentence embedding models on five tasks, including 2 downstream tasks (MR, SST2) and 3 intrinsic evaluations (STS-B, STR, *EvalRank*). The models' performance rank on the MR task is used as the pivot.

As MR and SST2 datasets are both related to sentiment analysis, they correlate well with each other. Among the three intrinsic evaluation tasks, *EvalRank* shows a higher correlation with downstream tasks as the blue dots roughly follow the trend of red dots. In contrast, the dots of STS-B and STR are dispersed in different regions. This shows that the performance of STS-B and STR is not a good indicator of the performance on downstream tasks.

Ablation Study. In Table 7, we show the perfor-

mance of *EvalRank* with different data sources. By combining the positive pairs collected from both STS-B and STR datasets, *EvalRank* leads to the best performance. Interestingly, according to our results, even though STR evaluation does not correlate well with downstream tasks, the positive pairs collected from STR have better quality than STS-B. It also confirms the argument that STR improves the dataset collection process (Abdalla et al., 2021).

Benchmarking Results. Table 5b benchmarked seven popular sentence embedding models. As the widely accepted SOTA model, SimCSE outperforms others with a clear margin.

7 Conclusion

In this work, we first discuss the problems with current word and sentence similarity evaluations and proposed *EvalRank*, an effective intrinsic evaluation method for word and sentence embedding models. It shows a higher correlation with downstream tasks. We believe that our evaluation method can have a broader impact in developing future embedding evaluation methods, including but not limited to its multilingual and task-specific extensions.

Acknowledgement

This research is supported by the Agency for Science, Technology and Research (A*STAR) under its AME Programmatic Funding Scheme (Project No. A18A2b0046) and Science and Engineering Research Council, Agency of Science, Technology and Research (A*STAR), Singapore, through the National Robotics Program under Human-Robot Interaction Phase 1 (Grant No. 192 25 00054).

References

Mohamed Abdalla, Krishnapriya Vishnubhotla, and Saif M Mohammad. 2021. What makes sentences semantically related: A textual relatedness dataset and empirical study. *arXiv preprint arXiv:2110.04845*.

Eneko Agirre, Enrique Alfonseca, Keith Hall, Jana Kravalova, Marius Paşca, and Aitor Soroa. 2009. A study on similarity and relatedness using distributional and WordNet-based approaches. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 19–27, Boulder, Colorado. Association for Computational Linguistics.

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 2015)*, pages 252–263, Denver, Colorado. Association for Computational Linguistics.

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)*, pages 81–91, Dublin, Ireland. Association for Computational Linguistics.

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-2016)*, pages 497–511, San Diego, California. Association for Computational Linguistics.

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 – 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)*, pages 385–393, Montréal, Canada. Association for Computational Linguistics.

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*, pages 32–43, Atlanta, Georgia, USA. Association for Computational Linguistics.

Nada Almarwani, Hanan Aldarmaki, and Mona Diab. 2019. Efficient sentence embedding using discrete cosine transform. 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 3672–3678, Hong Kong, China. Association for Computational Linguistics.

John R Anderson. 1983. A spreading activation theory of memory. *Journal of Verbal Learning and Verbal Behavior*, 22(3):261–295.

Sanjeev Arora, Yingyu Liang, and Tengyu Ma. 2017. A simple but tough-to-beat baseline for sentence embeddings. In *International Conference on Learning Representations (ICLR)*.

Simon Baker, Roi Reichart, and Anna Korhonen. 2014. An unsupervised model for instance level subcategorization acquisition. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 278–289, Doha, Qatar. Association for Computational Linguistics.

Petra Barancíková and Ondrej Bojar. 2019. In search for linear relations in sentence embedding spaces. In *Information Technologies – Applications and Theory (ITAT)*.

Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Janvin. 2003. A neural probabilistic language model. *The Journal of Machine Learning Research*, 3:1137–1155.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Elia Bruni, Nam-Khanh Tran, and Marco Baroni. 2014. Multimodal distributional semantics. *Journal of Artificial Intelligence Research*, 49:1–47.

Fredrik Carlsson, Amaru Cuba Gyllensten, Evangelia Gogoulou, Erik Ylipää Hellqvist, and Magnus Sahlgren. 2020. Semantic re-tuning with contrastive tension. In *International Conference on Learning Representations (ICLR)*.

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-2017)*, pages 1–14, Vancouver, Canada. Association for Computational Linguistics.

Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Brian Strope, and Ray Kurzweil. 2018. Universal sentence encoder for English. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 169–174, Brussels, Belgium. Association for Computational Linguistics.

Arman Cohan, Waleed Ammar, Madeleine van Zuylen, and Field Cady. 2019. Structural scaffolds for citation intent classification in scientific publications. 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)*, pages 3586–3596, Minneapolis, Minnesota. Association for Computational Linguistics.

Paul R Cohen and Rick Kjeldsen. 1987. Information retrieval by constrained spreading activation in semantic networks. *Information Processing & Management*, 23(4):255–268.

Allan M Collins and Elizabeth F Loftus. 1975. A spreading-activation theory of semantic processing. *Psychological Review*, 82(6):407–428.

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)*, Miyazaki, Japan. European Language Resources Association (ELRA).

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. Association for Computational Linguistics.

Alexis Conneau, German Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. 2018. What you can cram into a single $&!#* vector: Probing sentence embeddings for linguistic properties. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2126–2136, Melbourne, Australia. Association for Computational Linguistics.

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, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Bill Dolan, Chris Quirk, and Chris Brockett. 2004. Unsupervised construction of large paraphrase corpora: Exploiting massively parallel news sources. In *COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics*, pages 350–356, Geneva, Switzerland. COLING.

Steffen Eger, Andreas Rücklé, and Iryna Gurevych. 2019. Pitfalls in the evaluation of sentence embeddings. In *Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019)*, pages 55–60, Florence, Italy. Association for Computational Linguistics.

Manaal Faruqui, Yulia Tsvetkov, Pushpendre Rastogi, and Chris Dyer. 2016. Problems with evaluation of word embeddings using word similarity tasks. In *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*, pages 30–35, Berlin, Germany. Association for Computational Linguistics.

Lev Finkelstein, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppin. 2001. Placing search in context: The

concept revisited. In *Proceedings of the 10th international conference on World Wide Web*, pages 406–414.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Daniela Gerz, Ivan Vulić, Felix Hill, Roi Reichart, and Anna Korhonen. 2016. SimVerb-3500: A large-scale evaluation set of verb similarity. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2173–2182, Austin, Texas. Association for Computational Linguistics.

Anna Gladkova, Aleksandr Drozd, and Satoshi Matsuoka. 2016. 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*, pages 8–15, San Diego, California. Association for Computational Linguistics.

Guy Halawi, Gideon Dror, Evgeniy Gabrilovich, and Yehuda Koren. 2012. Large-scale learning of word relatedness with constraints. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1406–1414.

Zellig S Harris. 1954. Distributional structure. *Word*, 10(2-3):146–162.

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*, pages 1367–1377, San Diego, California. Association for Computational Linguistics.

Felix Hill, Roi Reichart, and Anna Korhonen. 2015. SimLex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*, 41(4):665–695.

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*, KDD '04, page 168–177, New York, NY, USA. Association for Computing Machinery.

Jaana Kekäläinen. 2005. Binary and graded relevance in IR evaluations: Comparison of the effects on ranking of IR systems. *Information Processing and Management*, 41(5):1019–1033.

Ryan Kiros, Yukun Zhu, Russ R Salakhutdinov, Richard Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Skip-thought vectors. In *Advances in Neural Information Processing Systems*, pages 3294–3302.

Bohan Li, Hao Zhou, Junxian He, Mingxuan Wang, Yiming Yang, and Lei Li. 2020. On the sentence embeddings from pre-trained language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9119–9130, Online. Association for Computational Linguistics.

Xin Li and Dan Roth. 2002. Learning question classifiers. In *COLING 2002: The 19th International Conference on Computational Linguistics*.

Tianlin Liu, Lyle Ungar, and João Sedoc. 2019a. Continual learning for sentence representations using conceptors. 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)*, pages 3274–3279, Minneapolis, Minnesota. Association for Computational Linguistics.

Tianlin Liu, Lyle Ungar, and Joao Sedoc. 2019b. Unsupervised post-processing of word vectors via conceptor negation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6778–6785.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2020. RoBERTa: A robustly optimized BERT pretraining approach. *International Conference on Learning Representations (ICLR)*.

Lajanugen Logeswaran and Honglak Lee. 2018. An efficient framework for learning sentence representations. *International Conference on Learning Representations (ICLR)*.

Jordan J Louviere and George G Woodworth. 1991. Best-worst scaling: A model for the largest difference judgments. Technical report, Working paper.

Thang Luong, Richard Socher, and Christopher Manning. 2013. Better word representations with recursive neural networks for morphology. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, pages 104–113, Sofia, Bulgaria. Association for Computational Linguistics.

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'14)*, pages 216–223, Reykjavik, Iceland. European Language Resources Association (ELRA).

David E Meyer and Roger W Schvaneveldt. 1971. Facilitation in recognizing pairs of words: evidence of a dependence between retrieval operations. *Journal of Experimental Psychology*, 90(2):227.

Tomas Mikolov, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. In *International Conference on Learning Representations (ICLR)*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013b. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, pages 3111–3119.

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013c. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 746–751, Atlanta, Georgia. Association for Computational Linguistics.

George A Miller. 1998. *WordNet: An electronic lexical database*. MIT press.

Jiaqi Mu, Suma Bhat, and Pramod Viswanath. 2018. All-but-the-top: Simple and effective postprocessing for word representations. *International Conference on Learning Representations (ICLR)*.

Bo Pang and Lillian Lee. 2004. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, pages 271–278, Barcelona, Spain.

Bo Pang and Lillian Lee. 2005. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 115–124, Ann Arbor, Michigan. Association for Computational Linguistics.

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, Doha, Qatar. Association for Computational Linguistics.

Kira Radinsky, Eugene Agichtein, Evgeniy Gabrilovich, and Shaul Markovitch. 2011. A word at a time: Computing word relatedness using temporal semantic analysis. In *Proceedings of the 20th International Conference on World Wide Web*, WWW '11, page 337–346, New York, NY, USA. Association for Computing Machinery.

Nils Reimers, Philip Beyer, and Iryna Gurevych. 2016. Task-oriented intrinsic evaluation of semantic textual similarity. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 87–96, Osaka, Japan. The COLING 2016 Organizing Committee.

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, Hong Kong, China. Association for Computational Linguistics.

Anna Rogers, Aleksandr Drozd, and Bofang Li. 2017. The (too many) problems of analogical reasoning with word vectors. In *Proceedings of the 6th Joint Conference on Lexical and Computational Semantics (*SEM 2017)*, pages 135–148, Vancouver, Canada. Association for Computational Linguistics.

Herbert Rubenstein and John B. Goodenough. 1965. Contextual correlates of synonymy. *Communications of the ACM*, 8(10):627–633.

Andreas Rücklé, Steffen Eger, Maxime Peyrard, and Iryna Gurevych. 2018. Concatenated power mean word embeddings as universal cross-lingual sentence representations. *arXiv preprint arXiv:1803.01400*.

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 *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA. Association for Computational Linguistics.

Jianlin Su, Jiarun Cao, Weijie Liu, and Yangyiwen Ou. 2021. Whitening sentence representations for better semantics and faster retrieval. *arXiv preprint arXiv:2103.15316*.

Julien Tissier, Christophe Gravier, and Amaury Habrard. 2017. Dict2vec : Learning word embeddings using lexical dictionaries. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 254–263, Copenhagen, Denmark. Association for Computational Linguistics.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Bin Wang, Fenxiao Chen, Angela Wang, and C.-C. Jay Kuo. 2019a. Post-processing of word representations via variance normalization and dynamic embedding. In *2019 IEEE International Conference on Multimedia and Expo (ICME)*, pages 718–723.

Bin Wang and C.-C. Jay Kuo. 2020. SBERT-WK: A sentence embedding method by dissecting BERT-based word models. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:2146–2157.

Bin Wang, Angela Wang, Fenxiao Chen, Yuncheng Wang, and C-C Jay Kuo. 2019b. Evaluating word embedding models: Methods and experimental results. *APSIPA transactions on signal and information processing*, 8.

Kexin Wang, Nils Reimers, and Iryna Gurevych. 2021. TSDAE: Using transformer-based sequential denoising auto-encoderfor unsupervised sentence embedding learning. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 671–688, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Tongzhou Wang and Phillip Isola. 2020. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International Conference on Machine Learning (ICML)*, pages 9929–9939. PMLR.

Evan Weingarten, Qijia Chen, Maxwell McAdams, Jessica Yi, Justin Hepler, and Dolores Albarracín. 2016. From primed concepts to action: A meta-analysis of the behavioral effects of incidentally presented words. *Psychological Bulletin*, 142(5):472.

Janyce Wiebe, Theresa Wilson, and Claire Cardie. 2005. Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation*, 39(2):165–210.

John Wieting, Mohit Bansal, Kevin Gimpel, and Karen Livescu. 2015. From paraphrase database to compositional paraphrase model and back. *Transactions of the Association for Computational Linguistics*, 3:345–358.

Yuanmeng Yan, Rumei Li, Sirui Wang, Fuzheng Zhang, Wei Wu, and Weiran Xu. 2021. ConSERT: A contrastive framework for self-supervised sentence representation transfer. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5065–5075, Online. Association for Computational Linguistics.

Yan Zhang, Ruidan He, Zuozhu Liu, Kwan Hui Lim, and Lidong Bing. 2020. An unsupervised sentence embedding method by mutual information maximization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1601–1610, Online. Association for Computational Linguistics.

Vitalii Zhelezniak, Aleksandar Savkov, April Shen, and Nils Hammerla. 2019. Correlation coefficients and semantic textual similarity. 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)*, pages 951–962, Minneapolis, Minnesota. Association for Computational Linguistics.

Xunjie Zhu and Gerard de Melo. 2020. Sentence analogies: Linguistic regularities in sentence embeddings. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3389–3400, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Xunjie Zhu, Tingfeng Li, and Gerard de Melo. 2018. Exploring semantic properties of sentence embeddings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 632–637, Melbourne, Australia. Association for Computational Linguistics.

A Embedding Models

A good intrinsic evaluator should be a good indicator for downstream tasks. We want to compute the correlation between the results from intrinsic evaluators and downstream tasks to measure the quality of intrinsic evaluators. For this purpose, we collect as many models as possible and finally involved 38 word embedding models and 67 sentence embedding models in our experiments. We give a detailed introduction to the collected embedding models in this section.

A complete set of selected word embedding models is shown in Table 8. We collect pre-trained word embeddings with different dimensions and training corpus from GloVe, word2vec, fastText, Dict2vec, and PSL. ABTT (Mu et al., 2018) post-processing is further applied to each model to double the total number of word embedding models.

A complete set of selected sentence embedding models is shown in Table 9. Besides the models obtained using bag-of-words features from word embeddings, we also include popular neural-network-based models including InferSent, BERT, RoBERTa, SBERT, BERT-whitening, BERT-flow, and SimCSE. Different variants of these models are considered in order to be more comprehensive.

B More Experimental Details

In Section 3, we conduct several experiments to certify our judgments, and we would like to elaborate on the detailed experiment settings here.

In Table 1b, the performance rank of five BERT-based sentence embedding models are shown under both cos and l_2 distance measure. Detailed model settings are shown below:

- **SBERT**: BERT-base model trained on Natural Language Inference data with mean token embeddings.
  https://huggingface.co/sentence-transformers/bert-base-nli-mean-tokens

- **SimCSE**: Unsupervised SimCSE trained upon BERT-based-uncased.
  https://huggingface.co/princeton-nlp/unsup-simcse-bert-base-uncased

- **BERT**: BERT-based uncased model
  https://huggingface.co/bert-base-uncased

- **BERT-flow**: We use the BERT-base-uncased and average the word representations from the first and last layers as the sentence representation. The Gaussian mapping is trained on the target corpus, which is the STS-B testset in our case.

- **BERT-whitening**: Similar to BERT-flow, the averaging of word representation from first and last layers are used as sentence representations, and the BERT-base-uncased model is used. The whitening objective is computed using the target corpus.

In Table 1a, 6 models are selected, and their performance on one similarity task: STS-B and two downstream tasks: MR and SST2 are reported. The setting of the models follows the experiments in Table 1b. For GloVe and InferSent, the following settings are used:

- **GloVe**: 300-dimensional vector trained on Common Crawl corpus (840B tokens).

- **InferSent**: Version 1 of InferSent is used where the GloVe model is served as input.

Figure 2 shows a detailed analysis on different similarity levels. For the experiment, we first collect all sentence pairs from STS-B dataset. Then, we split the pairs into four subsets based on their similarity levels ([5,3],[4,2],[3,1],[2,0]). Further, we randomly sampled 3,000 samples for each subset as the final dataset splits to keep the number of samples even.

C More Discussions

C.1 Effect of Whitening on Downstream Tasks

A lot whitening methods been proposed targeting on improving the quality of word embeddings (Mu et al., 2018; Wang et al., 2019a; Liu et al., 2019b) and sentence embeddings (Arora et al., 2017; Liu et al., 2019a; Li et al., 2020; Su et al., 2021). However, in previous studies, the whitening methods are only proven to be effective with similarity tasks. The performance comparison on downstream tasks is either missing or limited.

Therefore, we conduct extensive experiments on two popular post-processing methods. For word embedding, the ABTT (Mu et al., 2018) post-processing technique is examined. For sentence embedding, the Principal Component Removal (Arora et al., 2017) method is applied for word-embedding-based models and BERT-whitening (Su et al., 2021) or BERT-flow (Li et al., 2020) is applied to BERT-based models. Arora et al. (2017) propose a weighting schema and post-processing step for sentence embeddings. Here, we solely test the effectiveness of the post-processing step.

Table 10 shows the performance comparison between the original model and the post-processed model. From both word-level and sentence-level experiments, we conclude that the post-processing methods play no obvious role or even hurt the performance in downstream tasks. In contrast, the results on similarity tasks improve a lot.

C.2 Alignment and Uniformity

Wang and Isola (2020) discussed alignment and uniformity property as an explanation to the success of contrastive learning. *EvalRank* can be viewed as a variant of these two measures and focus more on the local perspective. Therefore, the success of *EvalRank* also can be explained under the same umbrella. Meantime, measuring from a local perspective is more suitable for word and sentence embedding models because they are likely to form a manifold and can only approximate euclidean space locally.

Alignment: In Wang and Isola (2020), the alignment loss is defined with the average distance between positive samples:

$$L_{align}(f;\alpha) = \mathbb{E}_{(x,y)\sim p_{pos}}[||f(x) - f(y)||_2^{\alpha}]$$

It measures the total distance between positive pairs, and the smaller, the better. The alignment

measure does not consider the local properties of the embedding space. In contrast, *EvalRank* requires the positive pairs to be close in the embedding space while considering the density of the local embedding regions. If the density of embedding space around positive pairs is high, *EvalRank* method requires the embeddings of positive pairs to be more tightly closed. If the density of embedding space around positive pairs is low, *EvalRank* has a looser distance requirement for the positive pairs.

Uniformity: In Wang and Isola (2020), the uniformity loss is designed as the logarithm of the average pairwise Gaussian potential:

$$L_{uniform}(f;t) = \log \ \mathbb{E}_{(x,y\tilde{p}_{data})}[e^{-t||f(x)-f(y)||_2^2}]$$

Intuitive, the uniformity loss asks features to be far away from each other. In contrast, *EvalRank* score focus on a local perspective. It requires the negative samples to have larger embedding distances than positive samples concerning the pivot sample. For the negative samples that are far away from the pivot sample in the embedding space, they are less likely to be confusing with positive samples and, therefore, not considered as important.

C.3 Correlation Results without Post-Processing Models

In previous experiments, we select as many models as possible in order to be more comprehensive. However, the side effect is that a reasonable portion of the models is built with post-processing techniques. It may lead to some concern that our selected embedding models might be biased on post-processed models. Therefore, we re-do the experiments on sentence embedding evaluations without considering post-processed models.

We filter out all models related to post-processing techniques, and as a result, 34 sentence embedding models are kept. We further conduct correlation analysis between the performance on intrinsic evaluation methods and downstream tasks.

The result is shown in Table 11. As we can see, *EvalRank* still outperforms sentence similarity tasks in 7 of the tasks. And we can witness a higher correlation between *EvalRank* and the downstream tasks comparing with the results in Table 6. *EvalRank* shows strong correlation ($\rho > 0.6$) on all 8 tasks and very strong correlation ($\rho > 0.8$) on 7 of the tasks. The result again proves the effectiveness of *EvalRank*.

Model #	Model Name	Details	Post-process
1 / 2	GloVe (Pennington et al., 2014)	glove.6B.50d	no / yes
3 / 4	GloVe	glove.6B.100d	no / yes
5 / 6	GloVe	glove.6B.200d	no / yes
7 / 8	GloVe	glove.6B.300d	no / yes
9 / 10	GloVe	glove.42B.300d	no / yes
11 / 12	GloVe	glove.840B.300d	no / yes
13 / 14	GloVe	glove.twitter.27B.25d	no / yes
15 / 16	GloVe	glove.twitter.27B.50d	no / yes
17 / 18	GloVe	glove.twitter.27B.100d	no / yes
19 / 20	GloVe	glove.twitter.27B.200d	no / yes
21 / 22	word2vec (Mikolov et al., 2013b)	GoogleNews-vectors-negative300	no / yes
23 / 24	fastText (Bojanowski et al., 2017)	crawl-300d-2M	no / yes
25 / 26	fastText	crawl-300d-2M-subword	no / yes
27 / 28	fastText	wiki-news-300d-1M	no / yes
29 / 30	fastText	wiki-news-300d-1M-subword	no / yes
31 / 32	Dict2vec (Tissier et al., 2017)	dict2vec-100d	no / yes
33 / 34	Dict2vec	dict2vec-200d	no / yes
35 / 36	Dict2vec	dict2vec-300d	no / yes
37 / 38	PSL (Wieting et al., 2015)	paragram_300_sl999	no / yes

Table 8: Word embedding models used in our evaluation. We use ABTT as the post-processing method (Mu et al., 2018).

Model #	Model Type	Model Name	Details	Post-Process
1 / 2	we-bow	GloVe	glove.6B.50d	no / yes
3 / 4	we-bow	GloVe	glove.6B.100d	no / yes
5 / 6	we-bow	GloVe	glove.6B.200d	no / yes
7 / 8	we-bow	GloVe	glove.6B.300d	no / yes
9 / 10	we-bow	GloVe	glove.42B.300d	no / yes
11 / 12	we-bow	GloVe	glove.840B.300d	no / yes
13 / 14	we-bow	GloVe	glove.twitter.27B.25d	no / yes
15 / 16	we-bow	GloVe	glove.twitter.27B.50d	no / yes
17 / 18	we-bow	GloVe	glove.twitter.27B.100d	no / yes
19 / 20	we-bow	GloVe	glove.twitter.27B.200d	no / yes
21 / 22	we-bow	word2vec	GoogleNews-vectors-negative300	no / yes
23 / 24	we-bow	fasttext	crawl-300d-2M	no / yes
25 / 26	we-bow	fasttext	crawl-300d-2M-subword	no / yes
27 / 28	we-bow	fasttext	wiki-news-300d-1M	no / yes
29 / 30	we-bow	fasttext	wiki-news-300d-1M-subword	no / yes
31 / 32	we-bow	dict2vec	dict2vec-100d	no / yes
33 / 34	we-bow	dict2vec	dict2vec-200d	no / yes
35 / 36	we-bow	dict2vec	dict2vec-300d	no / yes
37 / 38	we-bow	PSL	paragram_300_sl999	no / yes
39 / 40	neural net	InferSent	v_1 / v_2	no
41 / 42	neural net	BERT	bert-base-uncased & cls	no / whitening
43 / 44	neural net	BERT	bert-base-uncased & last-avg	no / whitening
45 / 46	neural net	BERT	bert-base-uncased & first-last-avg	no / whitening
47 / 48	neural net	BERT	bert-large-uncased & cls	no / whitening
49 / 50	neural net	BERT	bert-large-uncased & last-avg	no / whitening
51 / 52	neural net	BERT	bert-large-uncased & first-last-avg	no / whitening
53 / 54	neural net	RoBERTa	roberta-base & last-avg	no / whitening
55 / 56	neural net	RoBERTa	roberta-base & first-last-avg	no / whitening
57 / 58	neural net	RoBERTa	roberta-large & last-avg	no / whitening
59 / 60	neural net	RoBERTa	roberta-large & first-last-avg	no / whitening
61	neural net	BERT-flow	bert-base-uncased & cls	N/A
62	neural net	BERT-flow	bert-base-uncased & last-avg	N/A
63	neural net	BERT-flow	bert-base-uncased & first-last-avg	N/A
64 / 65	neural net	SBERT	sbert-base-nli-mean-tokens	no / whitening
66	neural net	SimCSE	unsup-simcse-bert-base-uncased	no
67	neural net	SimCSE	sup-simcse-bert-base-uncased	no

Table 9: Sentence embedding models used in our evaluation. For word-embedding-based models, the bag-of-words feature is used, and the principal component removal algorithm is used as the post-processing of sentence embeddings (Arora et al., 2017). For post-processing for BERT-based model, the BERT-whitening model is applied (Su et al., 2021)

	Word-Level	Sentence-Level
Post-Process	No v.s. Yes	No v.s. Yes
SCICITE	45.0% < 55.0%	67.6% > 32.4%
MR	42.1% < 57.9%	63.6% > 36.4%
CR	26.3% < 73.7%	54.5% > 45.5%
MPQA	94.7% > 5.3%	97.0% > 3.0%
SUBJ	78.9% > 21.1%	87.9% > 12.1%
SST2	80.0% > 20.0%	88.2% > 11.8%
SST5	80.0% > 20.0%	88.2% > 11.8%
TREC	89.5% > 10.5%	81.8% > 18.2%
MRPC	50.0% = 50.0%	64.7% > 35.3%
SICK-E	15.0% < 85.0%	51.5% > 48.5%
WS-353-All	21.1% < 78.9%	NA
WS-353-Rel	26.3% < 73.7%	NA
WS-353-Sim	21.1% < 78.9%	NA
RW-STANFORD	47.4% < 52.6%	NA
MEN-TR-3K	15.8% < 84.2%	NA
MTURK-287	26.3% < 73.7%	NA
MTURK-771	21.1% < 78.9%	NA
SIMLEX-999	21.1% < 78.9%	NA
SIMVERB-3500	36.8% < 63.2%	NA
STS12	NA	3.0% < 97.0%
STS13	NA	0.0% < 100.0%
STS14	NA	0.0% < 100.0%
STS15	NA	0.0% < 100.0%
STS16	NA	3.0% < 97.0%
STS-Benchmark	NA	0.0% < 100.0%
SICK-Relatedness	NA	15.2% < 84.8%
STR	NA	6.0% < 94.0%

Table 10: Performance of models with and without post-processing step.

		SCICITE	MR	CR	MPQA	SUBJ	SST2	SST5	TREC
STS12		35.45	39.07	39.65	63.36	28.60	30.87	42.92	42.58
STS13		42.51	42.88	46.71	**72.68**	32.44	34.10	47.70	42.15
STS14		37.99	38.05	41.73	68.05	27.50	30.18	43.38	43.16
STS15		44.19	46.07	47.41	68.78	35.08	38.11	51.14	50.13
STS16		63.62	64.30	66.33	71.81	56.87	57.09	68.38	66.03
STS-Benchmark		47.10	48.82	51.05	62.93	38.26	40.98	53.76	54.22
SICK-Relatedness		49.98	51.65	54.90	68.44	41.18	42.57	57.04	54.82
STR		-2.57	-1.53	-7.81	27.70	-1.47	-5.42	-3.12	9.37
EvalRank	MRR	83.37	85.40	85.45	66.38	81.29	82.62	85.77	85.69
	Hits@1	83.69	**86.15**	**85.82**	65.80	**82.28**	**83.21**	**85.78**	**86.67**
	Hits@3	**83.92**	85.19	84.97	66.38	81.37	82.36	85.74	86.51

Table 11: Spearman's rank correlation ($\rho \times 100$) between performance scores of sentence-level intrinsic evaluation and downstream tasks, where the best is marked with **bold** and second best with underline. The models with post-processing are filtered out, resulting in a total number of 34 sentence embedding models.