Why is sentence similarity benchmark not predictive of application-oriented task performance?

Kaori Abe[1], **Sho Yokoi**[1,2], **Tomoyuki Kajiwara**[3] **and Kentaro Inui**[1,2]
[1]Tohoku University [2]RIKEN [3]Ehime University
{abe-k, yokoi, kentaro.inui}@tohoku.ac.jp, kajiwara@cs.ehime-u.ac.jp

Abstract

Computing the semantic similarity between two texts is crucial in various NLP tasks. For more than a decade, a framework, known as Semantic Textual Similarity (STS) has been used to test computational models of semantic similarity (Agirre et al., 2012). The STS evaluation framework assumes that a model that performs well for the general STS task should also perform well for specific application-oriented tasks. However, does this assumption indeed hold? This study empirically demonstrates that the answer is not always positive. We found a considerable gap between model performance in STS and each specific task. We identified three factors that contributed to the gap, namely, (i) sentence length distribution, (ii) vocabulary coverage, and (iii) granularity of gold-standard similarity scores. We believe that these findings will be considered in future research on semantic similarity.

1 Introduction

Computing the semantic similarity between two texts is crucial in various NLP tasks. One prominent cluster of application examples is the use of semantic similarity as a metric for evaluating automatically generated text (e.g., machine translation and text summarization) considering gold reference texts (Zhang et al., 2020a; Sellam et al., 2020; Rei et al., 2020). Such semantic similarity metrics are also reported effective as a loss function for training language generation models (Wieting et al., 2019; Yasui et al., 2019). Another common application of the semantic similarity can be seen in text/sentence retrieval, where estimating the relevance between a given query and retrieved texts is an essential component (Chen et al., 2017; Karpukhin et al., 2020; Gao et al., 2021a; Qu et al., 2021).

For more than a decade, a framework, known as Semantic Textual Similarity (STS) has been widely used to test computational models of semantic similarity (Agirre et al., 2012). Over the last decade, STS has emerged as the de-facto standard task for evaluating semantic similarity models, and numerous studies have been published to propose semantic similarity models over a decade (Severyn et al., 2013; Lan and Xu, 2018; Reimers and Gurevych, 2019; Li et al., 2020; Zhang et al., 2020b; Yan et al., 2021; Giorgi et al., 2021; Gao et al., 2021b; Chuang et al., 2022, etc.).

The STS evaluation framework assumes that a model that performs well for the general STS task should also perform well for specific application-oriented tasks. Based on this assumption, models proposed for and evaluated on STS have been applied to application-oriented tasks. For example, in machine translation (MT) evaluation, for the model incorporating several universal sentence encoders (USE) (Conneau et al., 2017; Logeswaran and Lee, 2018; Cer et al., 2018), which performed well on STS, had the highest performance in WMT18 (Shimanaka et al., 2018). In addition, for semantic retrieval, STS-based models such as USE have been developed and validated their effectiveness (Yang et al., 2020). These studies appear to provide empirical evidence supporting the assumption that STS performs well as a general proxy for specific application-oriented tasks.

However, in this study, we question this widely accepted assumption. Specifically, we empirically investigated whether semantic similarity models superior to the general STS task perform better on specific application-oriented tasks. In the experiments, we chose two representative application-oriented tasks, MT Metrics (MTM) and passage retrieval (PR), and investigated the correlation of the performance of numerous (> 20) sampled models between STS and each specific task. From the results, we gained several findings as follows:

- Semantic similarity models exhibited a non-negligible gap in performance on STS and each specific task (i.e., MTM or PR) (Fig. 1).

Proceedings of the 3rd Workshop on Evaluation and Comparison of NLP Systems, pages 70 - 87

- The discrepancies appeared to be caused by the discrepancies between the STS dataset and each application-specific dataset, including (i) sentence length distribution, (ii) vocabulary coverage, and (iii) granularity of gold-standard similarity scores.

The identified gap, which we refer to as **the evaluation gap**, indicates that the assumption in question does not necessarily hold and demonstrates the potential dangers of relying solely on the current STS-based evaluation alone in studying the semantic similarity. We believe that our findings will be considered in future research on the crucial components of NLP.

2 Related work

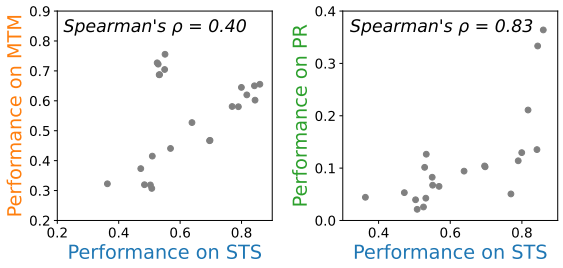
The necessity of the semantic similarity in application-oriented tasks. Semantic similarity is required in various NLP application tasks, and STS was motivated by being a surrogate task for such application-oriented tasks (Agirre et al., 2012; Cer et al., 2017). These tasks comparing similarity can be categorized into two types, namely, (1) reference-based evaluation and (2) semantic retrieval. For example, the reference-based evaluation is commonly used in the natural language generation (NLG) fields such as MT, summarization, and simplification. Semantic retrieval includes PR, dialog retrieval, as well as machine reading comprehension. Among these application-oriented tasks, we selected (1) MT evaluation and (2) PR as representatives, respectively.

In fact, MT evaluation and semantic retrieval have several examples that incorporate STS-based models. For example, Castillo and Estrella (2012); Shimanaka et al. (2018) applied STS model for MT evaluation and demonstrated the effectiveness of those models. For semantic retrieval, Yang et al. (2020) demonstrates the effectiveness of multilingual USE as a semantic retriever. Following this success, recent semantic similarity models have also reported performance as semantic retrievers (Gao et al., 2021b; Chuang et al., 2022). However, relying on the STS evaluation for semantic similarity models could be risky when there is no sufficient correlation between the evaluation of STS and application-oriented tasks. We investigates the evaluation gap between STS and two tasks, such as MT evaluation and PR, to identify vulnerabilities in the STS evaluation in the real world.

Validity of NLP evaluation protocol. Recently, the validity of evaluation protocols, such as benchmark datasets (Bowman and Dahl, 2021) or metrics (Mathur et al., 2020; Durmus et al., 2022) has been questioned on various NLP tasks. Many studies have identified the bias or lack of certain factors in the evaluation protocol. Søgaard et al. (2021); Varis and Bojar (2021) investigated the effects of differences in the sentence length distribution between train and test sets. Additionally, a difference in vocabulary distribution (domain mismatch) is also often mentioned as an important factor affecting the evaluation (Zhang et al., 2020b; Wang et al., 2022). In terms of an STS-specific factor, Reimers et al. (2016) highlighted the difference in the granularity of similarity between STS and downstream tasks. They focus on appropriate task-intrinsic evaluation metrics for STS-based models, considering different downstream tasks; however, their thought is also based on the assumption that the STS-based models are useful for the downstream tasks. In our study, we question this assumption. Based on these previous studies, we analyze the effects of three factors, **sentence length**, **vocabulary**, and **similarity granularity**, contributing to the evaluation gap between STS and the application-oriented tasks.

Discussion of the problems of STS benchmark. While many models have been proposed using the STS evaluation, some studies have also questioned the STS or conducted an additional evaluation for specific factors that are not captured by the STS evaluation. Wang et al. (2021) argue that previous studies rely on the STS evaluation and argues that STS lacks domain adaptability. Futhermore, Liu et al. (2021) did not adopt the STS evaluation because of the lack of domain coverage and lack of consideration for context, so they created a new contextual dialog domain STS dataset. In addition, Wieting et al. (2020) extracted a more difficult subset which contains the examples with low word overlap by focusing on a specific factor such as word overlap. Wang et al. (2022) focused on the discrepancy between the evaluation of STS and single-sentence downstream tasks in SentEval, highlighting the problems of domain mismatch and ambiguous annotations. In comparison, we investigated whether STS satisfies the original motivation for application-oriented tasks *practically using semantic similarity* (Agirre et al., 2012; Cer et al., 2017).

In summary, we shed the light on the specific



(a) STS × MT Metrics (b) STS × Passage Retrieval

Figure 1: Correlation between evaluation using STS and that using task-specific datasets, such as MT Metrics (MTM) and Passage Retrieval (PR).

factors such as sentence length, vocabulary, and similarity granularity to make the relationship to the evaluation gap explicit. We provided the first evidence that STS has a considerable evaluation gap even from two tasks, such as MT evaluation and PR that have been considered representative application tasks since the inception of STS.

3 Is there a gap between evaluation using STS and that using individual tasks?

STS dataset (Agirre et al., 2012; Cer et al., 2017) was proposed as a semantic similarity benchmark that can be directly applied to several NLP tasks and is currently the de-facto standard for evaluating semantic similarity models. In this study, to validate the STS benchmark, we conducted comprehensive experiments to examine whether there is a sufficient correlation between the evaluation results on STS and that on two specific application-oriented task datasets.

3.1 Tasks and datasets

General settings. We present the definitions of three tasks—STS and two application-oriented tasks—that must capture the semantic similarity addressed in this study. The main structure of all three tasks is comparing a sentence pair (s, s') and predicting the semantic similarity score between the two sentences. We selected two application-oriented tasks, MTM and PR, on which the STS motivation is focused. The two tasks are identical in that they require considering the semantic similarity, but they are very different in nature. MTM compares relatively similar sentence pairs and provides a gradation score as the gold standard. PR compares sentence pairs with large differences in sentence length and provides a binary label (related or not) as the gold standard. We examine the evaluation gap between these two different tasks and STS to test the adaptability of the STS evaluation to different tasks.

STS (STS-b). STS (Agirre et al., 2012) is a task that compares a sentence pair (s1, s2) and predicts a similarity score between the two sentences. The gold-standard similarity score is provided in the range of 0-5. Model prediction scores are evaluated using Pearson or Spearman correlations with the gold standard. In this study, we used Pearson correlation. We used the STS-b dataset (Cer et al., 2017) with image captions, news articles, and forum domain data over a 5-year pilot task (STS12-17).

MT Metrics (WMT17). MT Metrics (MTM) is a task that compares a (model hypothesis, reference) pair and predicts the adequacy scores of the model hypothesis relative to the reference. In this study, we use the segment-level Direct Assessment dataset (to-English) in WMT17 (Bojar et al., 2017).[1] We selected this because of the reliability of the manual scores (Zhang et al., 2020a; Sellam et al., 2020). The gold standard score is the normalized value of scores manually evaluated with 100 scales to the pair (model hypothesis, reference). The Pearson or Kendall correlation between the gold standard and the model prediction score is usually used in the evaluation. In this study, we used the Pearson correlation.

Passage Retrieval (MS-MARCO). Passage Retrieval (PR) is an important subtask of question-answering that is required to improve the performance of search systems used by many users. We use passage re-ranking data from MS-MARCO (Bajaj et al., 2018) as a dataset for PR. MS-MARCO is a highly competitive dataset that has been used as a PR benchmark in several studies (Gao et al., 2021a; Qu et al., 2021). Passage re-ranking must re-rank 1,000 candidate passages for a query in the order of their relevance to the query. Generally, the model predicts the relevance of each candidate sentence to the (query, passage) pair and extracts the sentence with the highest relevance score. Models are usually evaluated using Mean Reciprocal Rank (MRR), which determines whether passages with a gold-standard related labels appear at the top after re-ranking.

[1] We use cs–en, de–en, fi–en, lv–en, ru–en, tr–en and zh–en datasets, which are sourced from news domain texts. https://www.statmt.org/wmt17/results.html

3.2 Semantic similarity prediction model

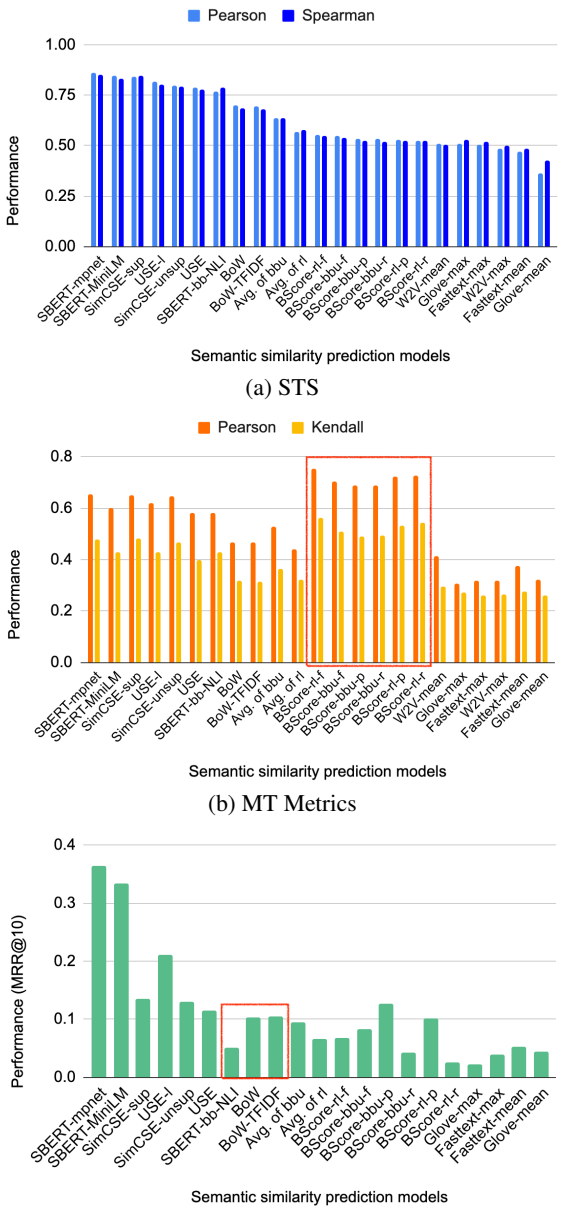
A semantic similarity prediction model usually involves the following two steps: (i) obtaining a sentence representation and (ii) calculating the similarity between two representations.

To determine whether there is an evaluation gap between various models, we measured the correlation between the evaluation results on STS and those on the two application tasks. In this study, we used the following 23 semantic similarity prediction models: **BoW**-{raw, TFIDF}-sum, **BoV**-{Word2vec*, Glove, Fasttext}-{mean, max}, **USE**-{normal, large}, **Avg. of BERT**-{BERT-base-uncased (bbu), RoBERTa-large (rl)}, **BERTScore (BScore)**-{BERT-base-uncased, RoBERTa-large}-{precision, recall, F1-score}, **Sentence-BERT (SBERT)**-{bertbase-NLI-mean, MiniLM, mpnet}, and **SimCSE**-{supervised, unsupervised}.[2]

3.3 Experimental procedure and results

Fig. 2 compares the evaluation for each semantic similarity prediction model on STS and the two application tasks, MTM and PR. The x-axis represents the semantic similarity prediction models, which are ordered by decreasing the performance on STS from left to right. Compared with STS, the performance of each model differs largely in both MTM and PR. For the STS evaluation, SBERT (mpnet: 0.86) outperforms BScore (RoBERTa-large, F1-score: 0.55); however, in the MT evaluation task (MTM), those performances are inverse as SBERT (0.66) < BScore (0.76). By comparing STS and PR, the performances of the SBERT-bb-NLI, the original model in (Reimers and Gurevych, 2019), and BoW models are much lower with PR than with STS. Both STS and MTM, both correlation measures have a similar trend for model ranking in each task (Fig. 2), thus we used the Pearson correlation in each task's evaluation. In addition, we calculated Spearman correlation coefficients between the performance on STS and that on each task to precisely visualize these performance gaps (Fig. 1). Here, we define these correlation coefficients as the value of the evaluation gap. A lower correlation value indicated a larger evaluation gap. In Sec. 4, we examine changes in the evaluation gap when the explanatory variables (e.g., sentence length, vocabulary coverage, similarity granularity) are changed.

[2]* We remove Word2vec models due to computational order in PR.



(a) STS

(b) MT Metrics

(c) Passage Retrieval

Figure 2: Performance of semantic similarity models on STS and task-specific datasets (MT Metrics and Passage Retrieval).

4 What factors cause the evaluation gap?

As mentioned in Sec. 3, there is a large gap between the specific application-oriented tasks and STS used as frameworks for evaluating the sentence similarity prediction models. In this section, we discuss three potential factors contributing to the gap between evaluation frameworks, as well as the dataset features that should considered to when using STS for evaluation.

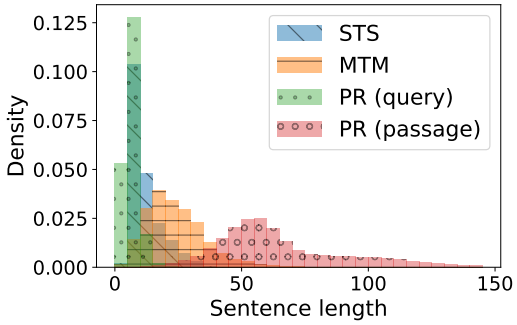


Figure 3: Histogram of sentence length in STS and two application-oriented datasets (MT Metrics: MTM and Passage Retrieval: PR).

4.1 Factor 1: difference in sentence length

In the following, we discuss the sentence length (i.e., the number of words in a sentence). Words are commonly used as the basic unit in NLP models. This is also true when making predictions of semantic similarity measures. We focused on the large variance in the number of words (i.e., sentence length) in the target text for similarity measurement. For example, in PR, the model should handle very short search snippets (queries) or very long documents (passages). Some studies reported that differences in the sentence length distributions produce different scores on different test sets (Søgaard et al., 2021; Varis and Bojar, 2021). Therefore, we hypothesize that differences in the distribution of sentence lengths by task may result in an evaluation gap.

4.1.1 Short sentence length in STS benchmark

Here, we demonstrate that the *STS dataset has shorter sentence lengths than the datasets for other specific tasks*, such as MTM and PR. Histograms of the sentence length distribution for each dataset are presented in Fig. 3. Note that the PR queries contain many short nonsentences, such as "define preventive." Compared with the sentence length distribution of the application-oriented task, STS has a biased sentence length distribution consisting of short sentences.

4.1.2 Does the sentence length gap cause the evaluation gap?

There is a difference in the sentence length distribution between STS and the application-oriented task datasets. Here, we investigate whether eliminating the difference in sentence length between the STS and the application tasks alleviates the evaluation gap.

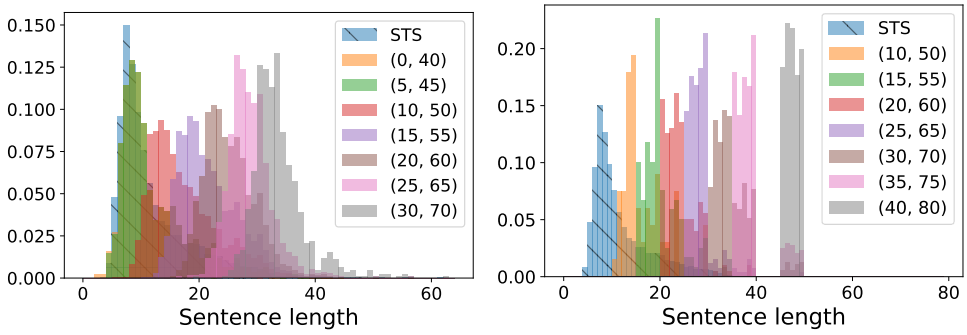
Settings. We created subsets of the application-oriented datasets (MTM and PR) to match or differ the STS sentence length distribution, and then, compared the correlations between the STS evaluation result and each subset's result for the different models. The subset [x, y) was drawn from a range of sentence lengths [x, y) according to the STS distribution. In MTM, the subsets were split based on the average sentence length of the sentence pairs. In PR, the split was based on the length of the passage because of a large-sentence length difference between the query and passage. Histograms of the created subsets according to sentence length distribution are shown in Fig. 4. We created MTM subsets from [0, 40) to [30, 70) and PR subsets from [10, 50) to [40, 80). The shorter MTM subsets, such as [0, 40) and [5, 45), had nearly the same distribution as the STS set. Note that we could not create a subset of PR with the same distribution as STS because the original sentence length distributions were very different. We investigated whether correlations were lower in the task-specific datasets (i.e., the evaluation gap was amplified) when their sentence length distribution was more different from that of STS.

Results. Figs. 5(a) and (b) present the Spearman correlations between the performance of the models on STS and those on the MTM and PR subsets with adjusted sentence length distributions, respectively. For MTM, the greater the difference in the sentence length distribution, the lower the correlation (i.e., the larger the evaluation gap). In comparison, no trend was observed for PR. This result indicates that the difference in the sentence length distribution contributes to the evaluation gap between STS and MTM.

Analysis: In-domain vs. Out-of-domain. The STS dataset is sourced from three different domains (news, image captions, and forum), and the sentence length distribution actually differs for each domain. We conducted additional experiments for three sub-domain sets following the same procedure using subsets, and found that the similar trends that the evaluation gap increases with the larger sentence length subset (See Appendix for details).

4.2 Factor 2: difference in vocabulary coverage

Beyond sentence length, there are still other factors that may contribute to the evaluation gap between STS and the application-oriented tasks. Here, we



(a) MT Metrics (b) Passage Retrieval

Figure 4: Histogram of subsets extracted from two application tasks (MT Metrics and Passage Retrieval) according to sentence length.

	STS		STS
MTM-[0, 40)	0.351	PR-[10, 50)	0.613
MTM-[5, 45)	0.351	PR-[15, 55)	0.792
MTM-[10, 50)	0.390	PR-[20, 60)	0.752
MTM-[15, 55)	0.407	PR-[25, 65)	0.777
MTM-[20, 60)	0.317	PR-[30, 70)	0.779
MTM-[25, 65)	0.284	PR-[35, 75)	0.770
MTM-[30, 70)	0.313	PR-[40, 80)	0.814

(a) MT Metrics (b) Passage Retrieval

Figure 5: Spearman correlations between performance with STS and that with the subsets split according to sentence length with specific tasks (MT Metrics: MTM and Passage Retrieval: PR). The darker color represents the lower correlation (= the larger evaluation gap). [x, y) means that the subsets consist of the examples of the sentence length from x to y.

discuss the vocabulary coverage of the application-oriented tasks using STS. One reason for focusing on this factor is that the text domains represented in the datasets are distinct. Some studies have highlighted the strong dependence of the STS-based models on domains (Zhang et al., 2020b), as well as mismatch with a dialog domain (Liu et al., 2021). Therefore, we hypothesize that differences in vocabulary coverage due to domain differences may influence the evaluation gap.

4.2.1 Low vocabulary coverage with STS for vocabulary in the applications

Here, we demonstrate that *the STS vocabulary does not adequately cover task vocabulary (MTM, PR).*

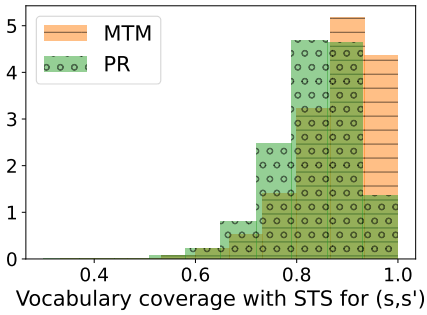


Figure 6: Histogram of the ratio of the vocabulary covered with the vocabulary of STS in the application tasks (MT Metrics: MTM and Passage Retrieval: PR) for each sentence pair.

For each sentence pair, we calculate the vocabulary coverage, which is the recall of vocabulary in STS (V_{sts}) to the vocabulary in the sentences in the specific task (s, s'), as follows:

$$\text{Recall}(s, s') = \frac{|(s \cup s') \cap \mathcal{V}_{\text{STS}}|}{|s \cup s'|} \quad (1)$$

Fig. 6 shows the histograms of $\text{Recall}(s, s')$ for each sentence pair in MTM and PR. In both tasks, most sentence pairs have a vocabulary coverage of less than 1, i.e., they contain vocabulary not covered by STS. Thus, STS vocabulary does not sufficiently cover the vocabulary of the other tasks.

4.2.2 Does the vocabulary distribution gap cause an evaluation gap?

We investigate whether the low vocabulary coverage with STS examined in Sec. 4.2.1 is indeed a factor contributing to the evaluation gap.

Settings. For the MTM and PR datasets, we extract the top and bottom 100 pairs as the $\text{Recall}(s, s')$-*High* and $\text{Recall}(s, s')$-*Low* subsets,

respectively. The MTM $\mathrm{Recall}(s, s')$-*High* subset contains all sentence pairs composed of STS vocabulary. Furthermore, the average of the PR *High* subset is 0.988 ± 0.011, which is almost all the pairs composed of STS vocabulary. In this experiment, we examine whether higher lexical coverage with the STS vocabulary for the subsets resulted in a higher correlation.

Results. Table 1 presents the Spearman correlation between the performance on STS and those on the $\mathrm{Recall}(s, s')$-*High* and *Low* subsets in MTM and PR, respectively. The PR *High* subset correlated better than the *Low* subset, as hypothesized. However, no such trend was observed in MTM. A reason for the MTM result is that STS is a mix of three different domains (news, image captions, and forum). In contrast, MTM is a single news domain dataset, which might have caused a divergence in the evaluation of sentence pairs from the same or different domains.

Analysis: In-domain vs. Out-of-domain. To confirm the influence of STS inner domains, we performed an additional analysis. We created vocabulary coverage subsets for the three STS sub-domain sets (news, image captions, and forum) in the same way as for the entire STS, and calculated the correlation between the three STS sub-domain sets and MTM *High/Low* subsets. For an in-domain setting, the MTM subset with *High* vocabulary coverage using STS-news correlated better than that with *Low* vocabulary coverage ($0.438 > 0.373$), as hypothesized. For out-of-domain settings, the STS-forum set also showed that the *High* subset has a better correlation than the *Low* subset ($0.779 > 0.458$); however, in the image caption set, the correlation of the *Low* subset (0.177) is better than that of *High* subset (0.046). For the image caption domain, the correlation values are extremely low for both the subsets, indicating that the STS image caption set did not play a good role in the evaluation of application-oriented tasks such as MTM. In summary, these results indicate that the vocabulary coverage contributes to evaluating gap between STS and the two application-oriented tasks, such as MTM and PR.

4.3 Factor 3: difference in granularity of gold-standard scores

Below, we consider the granularity gap of the gold-standard similarity scores between STS and MTM. [3]

	$\mathrm{Recall}(s, s')$-*Low*		$\mathrm{Recall}(s, s')$-*High*
MTM	0.276	>	0.272
PR	**0.673**	<	0.851

Table 1: Spearman correlations between the performance with STS and that with the subsets split according to higher vocabulary coverage ($\mathrm{Recall}(s, s')$-*High*) and lower one ($\mathrm{Recall}(s, s')$-*Low*) with STS of specific tasks (MT Metrics: MTM and Passage Retrieval: PR).

We suspect that the granularity of the similarity that was considered in each task varies. The distinction between better or worse hypotheses for high-similarity sentence pairs is an arresting challenge in MTM (Ma et al., 2019). More concretely, the current semantic evaluation model for MTM is unable to finely discriminate the better outputs in highly competitive language pairs such as to-English because of high quality of recent MT output for highly competitive language pairs. Considering this application, we hypothesize that the similarity granularity of STS is insufficient to evaluate such MTM problems.

4.3.1 The discrepancy of the similarity granularity between STS and MTM

The difference in the similarity score between STS and MTM can be seen in some real examples. The actual examples in STS and MTM are illustrated in Table 2. STS provides give relatively high scores for the difference between the past and present progressive tenses, and the difference in including proper nouns such as *cholera*, as long as they generally share some elements. However, in MTM, the first example is given a relatively higher score (0.49) for the different actions between *continues to take* and *is already given*, whereas the second example (*Fresh fruit ...*) is assigned a lower score (-0.83), sharing almost similar elements but the hypothesis is somewhat difficult to understand. Can this similarity granularity gap cause the evaluation gap?

4.3.2 Does the gap in the granularity of similarity cause an evaluation gap?

Here, we investigate whether the difference in the similarity granularity mentioned in Sec. 4.3.1 results in the evaluation gap.

[3] In this section, we omit considering PR because the property of PR is different from the other tasks in terms of binary labels.

source		s1 (ref)	s2 (hyp)	gold	BScore	SimCSE
STS	(i)	A man **is riding** a mechanical bull.	A man **rode** a mechanical bull.	4	0.98	0.96
	(ii)	A total of 17 cases have been confirmed in the southern city of Basra, the Organization said.	A total of 17 confirmed cases of **cholera** were reported yesterday by the **World Health** Organisation in the southern **Iraqi** city of Basra.	3.6	0.93	0.74
MTM	(i)	This drug **continues to take** 12 months after a heart attack, which can reduce the risk of a stroke or heart attack.	The drug **is already given for** 12 months after a heart attack, reducing the risk of a stroke or another attack.	0.49	0.94	0.90
	(ii)	Fresh fruit **was replaced with** cheaper dried fruit.	Fresh fruit **is** cheap dried fruit **instead**.	**-0.83**	0.94	0.82

Table 2: Actual examples of STS and MT Metrics (MTM). The gold scores of MTM are normalized in the range (-1.81, 1.44) from with manually evaluated 100-scale scores. "BScore" and "SimCSE" mean prediction scores with BERTScore (RoBERTa-large, F1-score) and SimCSE (supervised), respectively.

Settings. For the STS and MTM datasets, we create subsets according to the similarity scores for a sentence pair. We divide the STS dataset into five subsets by considering six labels from 0 to 5. For the MTM dataset, we separated four subsets (*Sim-{Low, MidLow, MidHigh and High}*) by quartiles for human-rated golden scores. We determined the gap between the evaluations using STS and MTM subsets to confirm which range of the similarity granularity impacts the gap in the evaluation. Specifically, the correlation might be higher between the narrower range of the similarity band of STS and the wider range of that of MTM. We anticipate that the higher similarity band in STS only correlates with the MTM dataset, to consider the demand of the MTM that must distinguish higher similarity pairs.

Results. Fig. 7 shows the Spearman correlations between the similarity granularity subsets of STS and that of the MTM. As hypothesized, only the high-similarity subsets of STS, *STS-(3,4]* and *STS-(4,5]*, were highly correlated with all the MTM subsets. These results significantly show that STS is unable to evaluate discrimination performance in the fine-grained higher similarity bands.

In Fig. 8, we describe one interpretation of the above result. We suspect that STS cannot capture fine-grained granularity at higher similarity bands, as discussed (Sec 4.3.1). Not only is the evaluation of the high-similarity band of STS is higher correlated with that of MTM, but the low-similarity band of STS and MTM are nearly uncorrelated or inversely correlated (Fig. 7). We should consider introducing finer granularity in high similarity bands for STS, while also considering exclusion examples in ineffective low similarity bands as a widely applicable benchmark.

	STS-[0, 1]	STS-(1,2]	STS-(2,3]	STS-(3,4]	STS-(4,5]
MTM-Sim-Low	0.101	-0.001	-0.008	0.627	0.643
MTM-Sim-MidLow	0.065	-0.046	-0.172	0.708	0.690
MTM-Sim-MidHigh	-0.097	-0.214	-0.330	0.639	0.592
MTM-Sim-High	-0.088	-0.267	-0.387	0.533	0.529

Figure 7: Spearman correlations between performance on subsets according to gold-standard similarity scores of STS and MT Metrics (MTM). The darker color represents the lower correlation (= the larger evaluation gap).

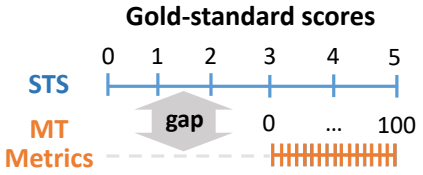


Figure 8: The relationship of the granularity of similarity scores between STS and MT Metrics.

Analysis: Tendency for each domain. As in the previous analyses, we investigated the difference in the tendencies for each domain. The correlations between subsets and MTM similarity subsets in each STS sub-domain sets are shown in Fig. 11. For the in-domain setting (STS-news ↔ MTM), only the middle similarity band showed a strong negative correlation with the MTM evaluation. For the two domains in the out-of-domain setting, the image caption set showed no correlation with MTM at lower similarity levels, whereas the forum domain set showed correlation only at very high or low similarity levels. One of the possible reasons for this strange phenomenon is the ambiguity of STS annotations due to label definition and amateur annotator discussed in (Wang et al., 2022).

	STS-news-[0, 1]	STS-news-(1,2]	STS-news-(2,3]	STS-news-(3,4]	STS-news-(4,5]
MTM-Low	0.341	0.154	-0.217	0.479	0.632
MTM-MidLow	0.379	0.566	-0.537	0.664	0.716
MTM-MidHigh	0.249	0.515	-0.632	0.595	0.650
MTM-High	0.260	0.466	-0.728	0.529	0.588

(a) news

	STS-image-[0, 1]	STS-image-(1,2]	STS-image-(2,3]	STS-image-(3,4]	STS-image-(4,5]
MTM-Low	-0.019	0.070	0.405	0.409	0.514
MTM-MidLow	-0.086	0.167	0.399	0.490	0.569
MTM-MidHigh	-0.215	0.029	0.319	0.384	0.437
MTM-High	-0.271	-0.006	0.238	0.215	0.352

(b) image captions

	STS-forum-[0, 1]	STS-forum-(1,2]	STS-forum-(2,3]	STS-forum-(3,4]	STS-forum-(4,5]
MTM-Low	0.475	0.112	-0.359	0.170	0.452
MTM-MidLow	0.587	0.165	-0.426	0.059	0.555
MTM-MidHigh	0.554	0.136	-0.239	-0.006	0.658
MTM-High	0.548	0.183	-0.079	0.016	0.688

(c) forum

Figure 9: Spearman correlations between performance on subsets divided according to gold-standard similarity scores of each STS domain (news, forum, image captions) and MT Metrics (MTM). The darker color represents the lower correlation (= the larger evaluation gap).

Particularly, there is a large gap between the definitions of 2 (*not equivalent but share some details*) and 3 (*roughly equivalent*) in terms of semantic equivalence, which can be attributed to this result.

5 Discussion and conclusions

We have investigated the gap between evaluation scores on the STS benchmark dataset and those on the evaluation datasets for MT evaluation (MTM) and Passage Retrieval (PR). We identified three factors contributing to this evaluation gap; namely, (i) sentence length distribution, (ii) vocabulary coverage ratio, and (iii) similarity granularity. These factors actually contributed to the evaluation gap, indicating that STS is not currently a directly applicable benchmark for evaluating semantic similarity at present. Future work could include checking for causal effects and controlling for covariates to rigorously identify factors, as well as investigate evaluation gaps in other tasks and domains.

Therefore, what should we do? The evaluation of semantic similarity alone must continue to be studied because of the significant demand for predicting semantic similarity (Sec. 1). One feasible approach is to evaluate and validate the model performance on multiple datasets that engage real-world tasks, rather than just STS. Wang et al. (2021) argued that the evaluation of existing semantic similarity models is biased toward STS and reported evaluation results on several datasets, including STS. Additionally, there have also been attempts to create a union of evaluation datasets from multiple task data and use it as a basis for evaluation in neighboring fields, such as PASCAL-RTE (Dagan et al., 2006) or SentEval (Conneau and Kiela, 2018). While these attempts have been achieved, there is an assumption that there are substantial costs are involved in regularly maintaining the infrastructure in each of these areas. To proceed with this approach, including STS, we should address the problem of STS shown in this study, and pursue what it should be as a benchmark for semantic similarity evaluation. Whatever approach we take, we must consider each of these factors contributing to the evaluation gap described in this study and refine them stably.

Acknowledgments This work was supported by JSPS KAKENHI Grant Number JP20J21694 and JST, ACT-X Grant Number JPMJAX200S, Japan.

References

Eneko Agirre, Daniel Cer, Mona Diab, and Aitor Gonzalez-Agirre. 2012. SemEval-2012 task 6: A pilot on semantic textual similarity. In **SEM*, pages 385–393.

Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, Mir Rosenberg, Xia Song, Alina Stoica, Saurabh Tiwary, and Tong Wang. 2018. Ms marco: A human generated machine reading comprehension dataset. *ArXiv*, pages 1–16.

Ondřej Bojar, Yvette Graham, and Amir Kamran. 2017. Results of the wmt17 metrics shared task. In *WMT*, pages 489–513.

Samuel R. Bowman and George Dahl. 2021. What will it take to fix benchmarking in natural language understanding? In *NAACL*, pages 4843–4855.

Julio Castillo and Paula Estrella. 2012. Semantic textual similarity for MT evaluation. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 52–58.

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 *SemEval*, pages 1–14.

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 *EMNLP*, pages 169–174.

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading Wikipedia to answer open-domain questions. In *ACL*, pages 1870–1879.

Yung-Sung Chuang, Rumen Dangovski, Hongyin Luo, Yang Zhang, Shiyu Chang, Marin Soljacic, Shang-Wen Li, Wen-tau Yih, Yoon Kim, and James Glass. 2022. DiffCSE: Difference-based contrastive learning for sentence embeddings. In *NAACL*, pages 1–12.

Alexis Conneau and Douwe Kiela. 2018. SentEval: An evaluation toolkit for universal sentence representations. In *LREC*, pages 1699–1704.

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. 2017. Supervised learning of universal sentence representations from natural language inference data. In *EMNLP*, pages 670–680.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2006. The pascal recognising textual entailment challenge. In *Machine Learning Challenges. Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Tectual Entailment*, pages 177–190, Berlin, Heidelberg. Springer Berlin Heidelberg.

Esin Durmus, Faisal Ladhak, and Tatsunori Hashimoto. 2022. Spurious correlations in reference-free evaluation of text generation. In *ACL*, pages 1443–1454.

Luyu Gao, Zhuyun Dai, and Jamie Callan. 2021a. COIL: Revisit exact lexical match in information retrieval with contextualized inverted list. In *NAACL*, pages 3030–3042.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021b. SimCSE: Simple contrastive learning of sentence embeddings. In *EMNLP*, pages 6894–6910.

John Giorgi, Osvald Nitski, Bo Wang, and Gary Bader. 2021. DeCLUTR: Deep contrastive learning for unsupervised textual representations. In *ACL-IJCNLP*, pages 879–895.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. XTREME: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In *ICML*, pages 4411–4421.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *EMNLP*, pages 6769–6781.

J. Peter Kincaid, Robert P. Jr. Fishburne, Richard L. Rogers, , and Brad S. Chissom. 1975. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel.

Matt Kusner, Yu Sun, Nicholas Kolkin, and Kilian Weinberger. 2015. From word embeddings to document distances. In *ICML*, pages 957–966.

Wuwei Lan and Wei Xu. 2018. Neural network models for paraphrase identification, semantic textual similarity, natural language inference, and question answering. In *COLING*, pages 3890–3902.

Hang Le, Loïc Vial, Jibril Frej, Vincent Segonne, Maximin Coavoux, Benjamin Lecouteux, Alexandre Allauzen, Benoit Crabbé, Laurent Besacier, and Didier Schwab. 2020. FlauBERT: Unsupervised language model pre-training for French. In *LREC*, pages 2479–2490.

Bohan Li, Hao Zhou, Junxian He, Mingxuan Wang, Yiming Yang, and Lei Li. 2020. On the sentence embeddings from pre-trained language models. In *EMNLP*, pages 9119–9130.

Yaobo Liang, Nan Duan, Yeyun Gong, Ning Wu, Fenfei Guo, Weizhen Qi, Ming Gong, Linjun Shou, Daxin Jiang, Guihong Cao, Xiaodong Fan, Ruofei Zhang, Rahul Agrawal, Edward Cui, Sining Wei, Taroon Bharti, Ying Qiao, Jiun-Hung Chen, Winnie Wu, Shuguang Liu, Fan Yang, Daniel Campos, Rangan Majumder, and Ming Zhou. 2020. XGLUE: A new benchmark datasetfor cross-lingual pre-training, understanding and generation. In *EMNLP*, pages 6008–6018.

Che Liu, Rui Wang, Jinghua Liu, Jian Sun, Fei Huang, and Luo Si. 2021. DialogueCSE: Dialogue-based contrastive learning of sentence embeddings. In *EMNLP*, pages 2396–2406.

Lajanugen Logeswaran and Honglak Lee. 2018. An efficient framework for learning sentence representations. In *ICLR*, pages 1–16.

Qingsong Ma, Johnny Wei, Ondřej Bojar, and Yvette Graham. 2019. Results of the WMT19 metrics shared task: Segment-level and strong MT systems pose big challenges. In *WMT*, pages 62–90.

Nitika Mathur, Timothy Baldwin, and Trevor Cohn. 2020. Tangled up in BLEU: Reevaluating the evaluation of automatic machine translation evaluation metrics. In *ACL*, pages 4984–4997.

Sungjoon Park, Jihyung Moon, Sungdong Kim, Won Ik Cho, Jiyoon Han, Jangwon Park, Chisung Song, Junseong Kim, Yongsook Song, Taehwan Oh, Joohong Lee, Juhyun Oh, Sungwon Lyu, Younghoon Jeong, Inkwon Lee, Sangwoo Seo, Dongjun Lee, Hyunwoo Kim, Myeonghwa Lee, Seongbo Jang, Seungwon Do, Sunkyoung Kim, Kyungtae Lim, Jongwon Lee, Kyumin Park, Jamin Shin, Seonghyun Kim, Lucy Park, Alice Oh, Jung-Woo Ha, and Kyunghyun Cho. 2021. Klue: Korean language understanding evaluation. *ArXiv*, pages 1–76.

Yingqi Qu, Yuchen Ding, Jing Liu, Kai Liu, Ruiyang Ren, Wayne Xin Zhao, Daxiang Dong, Hua Wu, and Haifeng Wang. 2021. RocketQA: An optimized training approach to dense passage retrieval for open-domain question answering. In *NAACL*, pages 5835–5847.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *EMNLP*, pages 2685–2702.

Nils Reimers, Philip Beyer, and Iryna Gurevych. 2016. Task-oriented intrinsic evaluation of semantic textual similarity. In *COLING*, pages 87–96.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *EMNLP-IJCNLP*, pages 3982–3992.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *ACL*, pages 7881–7892.

Aliaksei Severyn, Massimo Nicosia, and Alessandro Moschitti. 2013. Learning semantic textual similarity with structural representations. In *ACL*, pages 714–718.

Hiroki Shimanaka, Tomoyuki Kajiwara, and Mamoru Komachi. 2018. RUSE: Regressor using sentence embeddings for automatic machine translation evaluation. In *WMT*, pages 751–758.

Anders Søgaard, Sebastian Ebert, Jasmijn Bastings, and Katja Filippova. 2021. We need to talk about random splits. In *EACL*, pages 1823–1832.

Robyn Speer, Joshua Chin, Andrew Lin, Sara Jewett, and Lance Nathan. 2018. Luminosoinsight/wordfreq: v2.2.

Dusan Varis and Ondřej Bojar. 2021. Sequence length is a domain: Length-based overfitting in transformer models. In *EMNLP*, pages 8246–8257.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *ICLR*, pages 1–20.

Bin Wang, C.-c. Kuo, and Haizhou Li. 2022. Just rank: Rethinking evaluation with word and sentence similarities. In *ACL*, pages 6060–6077.

Kexin Wang, Nils Reimers, and Iryna Gurevych. 2021. TSDAE: Using transformer-based sequential denoising auto-encoderfor unsupervised sentence embedding learning. In *EMNLP Findings*, pages 671–688.

John Wieting, Taylor Berg-Kirkpatrick, Kevin Gimpel, and Graham Neubig. 2019. Beyond BLEU:training neural machine translation with semantic similarity. In *ACL*, pages 4344–4355.

John Wieting, Graham Neubig, and Taylor Berg-Kirkpatrick. 2020. A bilingual generative transformer for semantic sentence embedding. In *EMNLP*, pages 1581–1594.

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 *ACL*, pages 5065–5075.

Yinfei Yang, Daniel Cer, Amin Ahmad, Mandy Guo, Jax Law, Noah Constant, Gustavo Hernandez Abrego, Steve Yuan, Chris Tar, Yun-hsuan Sung, Brian Strope, and Ray Kurzweil. 2020. Multilingual universal sentence encoder for semantic retrieval. In *ACL (System Demonstrations)*, pages 87–94.

Go Yasui, Yoshimasa Tsuruoka, and Masaaki Nagata. 2019. Using semantic similarity as reward for reinforcement learning in sentence generation. In *ACL-SRW*, pages 400–406.

Seid Muhie Yimam, Chris Biemann, Shervin Malmasi, Gustavo Paetzold, Lucia Specia, Sanja Štajner, Anaïs Tack, and Marcos Zampieri. 2018. A report on the complex word identification shared task 2018. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 66–78.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020a. Bertscore: Evaluating text generation with bert. In *ICLR*, pages 1–43.

Yan Zhang, Ruidan He, Zuozhu Liu, Kwan Hui Lim, and Lidong Bing. 2020b. An unsupervised sentence embedding method by mutual information maximization. In *EMNLP*, pages 1601–1610.

A Appendix

A.1 Limitation: Experiments on only English STS

We would like to investigate other languages in this paper, but we are only concerned with the original English STS. Other languages than English also have benchmark datasets of the semantic similarity but are generally based on the STS framework. Since the GLUE (Wang et al., 2019), including STS, is facilitating model development for each task, a language-specific GLUE-like benchmark set (Le et al., 2020; Park et al., 2021) or cross-lingual benchmark set (Liang et al., 2020; Hu et al., 2020) are constructed. The benchmarks of the semantic similarity for each language are created in two methods: re-constriction by automatic translation or new construction by each language's expert following the original method. Crucially, the former method is likely to fundamentally face the same biases such as vocabulary distribution as those in the English benchmarks, albeit including the issue of translation quality. Regarding the latter, dataset creators may improve the original dataset creation method. For example, in the Korean GLUE (KLUE; Park et al., 2021), they added more detailed instructions on label definition when annotating the similarity by non-expert. Thus, it is necessary to re-consider the requirements for an appropriate benchmarks before straightforwardly following the original method when creating datasets.

A.2 Statistics of datasets and subsets in the experiments

Satistics of datasets. Table 3 shows statistics of three datasets (STS, MTM and PR) employed in this paper. The dataset size of STS is larger than that of MTM, whereas the total word counts are comparable between STS and MTM. The sentence length distribution (the number of of words / {s,s'}) shows that STS has very few words per sentence compared to the application-oriented tasks. As for the STS sub-domain sets, the three sets have different sentence length distributions. We additionally describe the histograms of the sentence length distributions for the three STS sub-domain sets in Fig. 10. As illustrated here, the average sentence length of the image-caption domain is particularly highly biased for shorter sentence lengths.

Statistics of subsets used in the experiment. Statistics of the subset of sentence length, vocabulary coverage, and the granularity of similarity are shown in Table 4, 6, and 7, respectively.

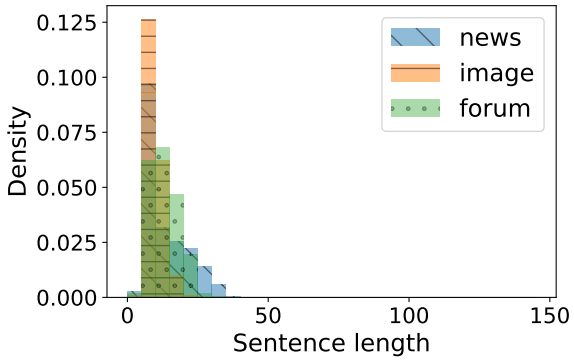


Figure 10: Histograms of sentence length in the STS sub-domain (news, image captions, forum) sets.

A.3 In-domain vs. Out-of-domain analysis in sentence length factor

Settings. We create subsets from the MTM dataset to match the sentence length distribution for each of three STS sub-domain sets. Notably, the forum and image caption domains have relatively small sentence length distributions (in Fig. 5, we thus reduced the range of the subsets from [0, 40) to [20, 60). Statistics of the subset of sentence length are shown in Table 5.

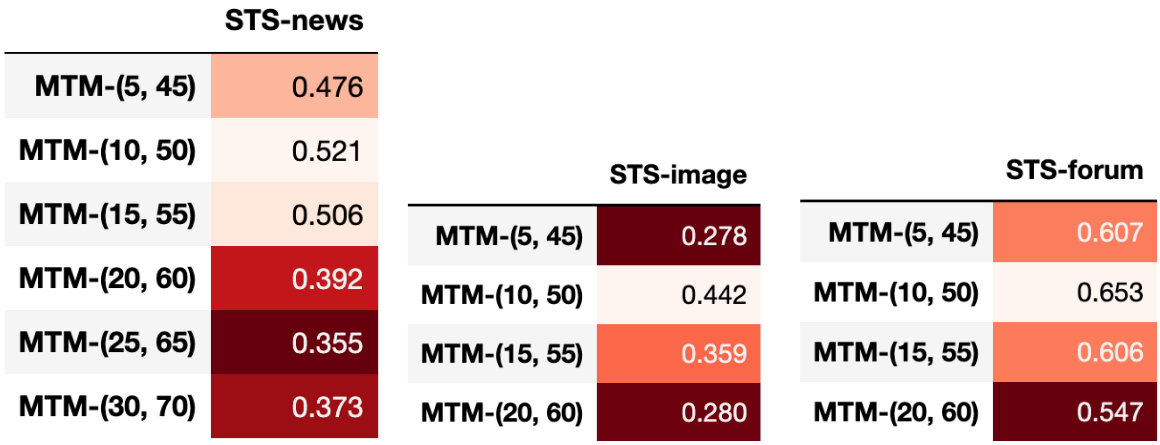
Results. Fig. 11 shows the correlation with MTM when sentence length subsets are created separately for each domain. We observed a similar tendency for all sub-domain sets that the evaluation gap increases for subsets of longer sentence lengths. This suggests that the evaluation results differ due to different sentence length distribution even within the same domain, which is consistent with a previous study's report in a different benchmark (Søgaard et al., 2021).

A.4 Extended Vocabulary analysis

STS has easier vocabulary STS contains more familiar words than that appear in the application tasks. As quantitative indicators of word familiarity, word frequency (Yimam et al., 2018) and word length (Kincaid et al., 1975) are often used mainly in the text simplification task. Intuitively, the higher the word frequency or the shorter the word length, the more familiar the word. In this case, we use "word frequency (wordfreq)" and "zipf frequency (zipffreq)" scale in `wordfreq` mod-

	STS (s1, s2)	MTM (hyp, ref)	PR (query, passage)
#sentence pairs	8,628	3,793	6,668,967
#sentences ({s, s'})	15,487	4,261	13,337,934
#words	186,134	170,565	472,778,794
#words / {s, s'}	**11.443±6.143**	23.381±11.215	35.908±35.266
#words / s	**11.450±6.188**	23.296±11.290	6.176± 2.642
#words / s'	**11.437±6.099**	23.467±11.138	65.640±26.692
	STS-news (s1, s2)	STS-forum (s1, s2)	STS-image-captions (s1, s2)
#sentence pairs	4,299	1,079	3,250
#sentences	8,268	1,913	5,306
#words	107,957	25,456	52,721
#words / {s, s'}	12.927±7.506	12.642±4.978	9.0823±2.910
#words / s	12.949±7.564	12.677±5.007	9.0585±2.906
#words / s'	12.905±7.448	12.608±4.949	9.1062±2.914

Table 3: Stats. of sentences and words and average of sentence length for STS (all and sub-domain sets) and application datasets (MT Metrics: MTM, Passage Retrieval: PR).



(a) news (b) image captions (c) forum

Figure 11: Spearman correlations between performance on sentence length subsets of STS-news, image captions, forum and MT Metrics (MTM) . The darker color indicates the lower correlation (= the larger evaluation gap).

	MTM		PR	
	size	avg. sent len	size	avg. sent len
[0, 40)	481	11.610±5.794	-	-
[5, 45)	481	11.790±5.979	-	-
[10, 50)	1225	16.841±5.747	67	16.045±4.420
[15, 55)	1484	21.086±5.015	119	19.849±3.759
[20, 60)	1112	24.722±4.286	199	23.704±3.285
[25, 65)	715	28.260±3.733	262	28.000±2.980
[30, 70)	465	33.184±4.462	561	34.526±3.855
[35, 75)	-	-	690	38.323±3.549
[40, 80)	-	-	932	46.987±1.390

Table 4: Stats. of sentence length subsets for MTM and PR. The "size" means the number of sentence pairs and the "avg. sent len" means the average of sentence length for each subset.

ule (Speer et al., 2018).[4] Wordfreq is the normalized frequency in the corpora, and zipffreq is the logarithmically scale of wordfreq. The word length is the number of characters in each word. We use `nltk.word_tokenize()` as word split and filtered out URLs and those with more than 50 characters.

Table 8 shows the average word frequency with the wordfreq module and word length for each dataset. In zipffreq, the average of STS is shorter than that of both the application tasks. Also in word length, we could observe that the average of STS is higher than that of MTM and PR. Thus, in both the indicators, word familiarity distribution in STS is higher than in the two application tasks.

Additionally, by comparing between "general" word frequencies (wordfreq) in the `wordfreq` module and actual word frequencies in the corpus (corpus-freq), we can identify words that appear particular high-frequently in the corpus. The words

[4] A tool to obtain word frequencies from 7 different corpora (Wikipedia, Subtitles, News, Books, Web text, Twitter, Reddit). https://pypi.org/project/wordfreq/

MTM						
	(STS-news-based)		(STS-forum-based)		(STS-image-captions-based)	
	size	avg. sent len.	size	avg. sent len.	size	avg. sent len.
[0, 40)	503	12.898±6.971	400	9.491±3.183	816	12.348±4.347
[5, 45)	506	13.238±7.259	398	9.521±3.162	867	13.106±4.855
[10, 50)	2150	19.356±6.201	676	13.024±2.620	1229	15.444±3.879
[15, 55)	1902	22.082±5.192	778	17.648±2.457	911	18.337±3.013
[20, 60)	1185	24.935±4.332	650	22.185±2.548	658	22.251±2.620
[25, 65)	715	28.260±3.733	-	-	-	-
[30, 70)	465	33.184±4.462	-	-	-	-

Table 5: Stats. of sentence length subsets for MTM according the sentence length distribution of STS sub-domain sets. The "size" means the number of sentence pairs and the "avg. sent len" means the average of sentence length (the average of {s, s'}) for each subset.

MTM								
	(STS-all-based)		(STS-news-based)		(STS-forum-based)		(STS-image-captions-based)	
	size	avg. Recall	size	avg. Recall	size	avg. Recall	size	avg. Recall
(all)	3,793	0.882±0.084	4,299	0.854±0.093	1,079	0.715±0.120	3,250	0.523±0.112
High	100	1.000±0.000	100	1.000±0.000	100	0.980±0.024	100	0.787±0.042
Low	100	0.631±0.060	100	0.588±0.058	100	0.418±0.063	100	0.252±0.062

PR		
	size	avg. Recall
all	6,614	0.835±0.079
High	100	0.988±0.011
Low	100	0.572±0.051

Table 6: Stats. of vocabulary subsets for MTM and PR.

STS								
	(all)		(news)		(forum)		(image captions)	
	size	avg. similarity	size	avg. similarity	size	avg. similarity	size	avg. similarity
[0, 1]	1182	0.655±0.280	594	0.522±0.393	275	0.472±0.420	931	0.360±0.353
(1, 2]	1348	1.631±0.285	640	1.631±0.283	248	1.687±0.286	460	1.601±0.283
(2, 3]	1672	2.653±0.291	876	2.678±0.291	232	2.656±0.292	564	2.615±0.286
(3, 4]	2317	3.614±0.287	1378	3.599±0.280	189	3.692±0.303	750	3.622±0.292
(4, 5]	1491	4.619±0.304	811	4.613±0.301	135	4.686±0.311	545	4.612±0.306

MTM		
	size	avg. similarity
Sim-Low: [-2, -0.47]	950	-0.820±0.266
Sim-MidLow: (-0.47, -0.03]	948	-0.240±0.126
Sim-MidHigh: (-0.03, 0.42]	943	0.193±0.127
Sim-High: (0.42, 1.5]	952	0.683±0.183

Table 7: Dataset size (#sentence pairs) and average & standard derivation of gold-standard similarity scores on STS and MTM subsets.

belongs to "corpus-freq − wordfreq > 0.001" for STS, MTM, and PR were 43, 18, and 26 words, respectively (if excluding stopwords and punctuation, 28, 3, and 6 words, respectively). Examples of higher frequent words in each dataset are shown in Table 9. As shown in this, some domain-specific words (STS: image captions, MTM: news, PR: question answering) are particularly frequent in each corpus. STS seems to be biased toward certain words (e.g., colors, present progressive forms, relatively abstract nouns such as *man* and *dog*). The results indicate that the STS has a relatively "easier" vocabulary (particularly sourced from the image-caption domain) than the application-oriented task.

Gap of proper noun in word representation distribution In actual semantic similarity predic-

	STS	MTM	PR
zipffreq (↑)	**3.59±1.24**	3.45±1.54	1.29±1.74
length (↓)	**6.97±2.76**	7.34±2.83	10.1±4.83

Table 8: Average of word frequency and word length in STS, MT Metrics: MTM, Passage Retrieval: PR. The higher (↑) the average for zipffreq (zipf scale of normalized word frequency) or the lower (↓) the average for word length, the higher the word familiarity can be considered.

tion models, words are embedded into a multi-dimensional space and treated as a soft distributed representation. Does the STS vocabulary still diverge from the vocabulary of the application-oriented tasks in the soft representations? To obtain an intuition for this, we plot word distribution in each dataset by t-SNE using the fasttext model. In the t-SNE setting, we use random initialization and set learning rate to 200 (scikit-learn), random state to 0. Fig. 12 shows the results of t-SNE plotting the top-frequency 5,000 words in each dataset. The areas surrounded with red lines are non-overlapping clusters between STS (blue) and the application tasks (MTM: orange, PR: green). Additionally, we enlarge some non-overlapping clusters in Fig. 13. These clusters mostly includes several proper nouns such as *Columbus*, *Carolina*, and *Robin* in all the datasets. In addition, to capture the quantitative distance between word distributions, we measured the Word Mover's Distance (WMD) (Kusner et al., 2015) with the above t-SNE representations. We use uniform distribution as the WMD weight and sqeuqlidian distance as the distance metric. The larger the value, the less STS covers the vocabulary of each application task. The distance between STS and MTM was 189.44 and the distance between STS and PR was 89.893. Thus, The vocabulary distribution gap between STS and the application-oriented tasks is caused by mainly the distribution of proper nouns.

A.5 NLI analysis

Various studies have found that pre-trained models of NLI dataset lead to improved performance on STS (Conneau et al., 2017; Reimers and Gurevych, 2019; Gao et al., 2021b). Gao et al. (2021b) tried several NLI and paraphrase identification datasets for model pre-training, indicating that NLI examples with the lowest lexical overlap have been the most effective. In this section, we show that the **sentence length** and **soft lexical** distribution of the NLI dataset are nearly STS-like. We suspect that the coincidence of these distributions is responsible for the improved performance of the NLI-supervised model on STS.

Sentence length analysis. Fig. 15 shows histograms of sentence length distribution for each dataset including NLI. As shown in this, NLI datasets have a relatively shorter sentence length distribution, similar to that of STS. Although MNLI contains relatively longer sentences than SNLI, there are still fewer examples of longer sentences compared to the application-oriented datasets such as MTM and PR.

Vocaburaly coverage analysis. In following, we see the vocabulary distribution on the NLI datasets. The statistics on NLI's vocabulary distribution are shown in Table 10. The Herdan's C of NLI is lower than that of STS; however, TTR of NLI close to that of MT Metrics. As the word familiarity distribution of NLI, the average of zipffreq shows that more high-frequency words appear in both SNLI and MNLI than in STS. However, the average of word length of NLI is close to that of MT Metrics. These results indicate that the words which appear in NLI are a fairly high frequent but those lengths are longer compared to STS. The visualization of the soft word distribution including NLI is shown in Fig. 14. As illustrated in this, the word distribution of NLI is similar for STS compared to the other datasets. This trend might contribute to the improvement of performances of NLI-supervised models such as SentenceBERT on STS.

A.6 Model description

Table 11 shows the descriptions of the models used in this paper.

STS	man, woman, playing, running, sitting, standing, guitar, white, black, red, dog, cat, horse, grass ⋯
MTM	said, police, olympic(, was, will, which, who, ⋯)
PR	name, definition, meaning, number, average(, what, your, ⋯)

Table 9: Examples of higher frequency words for STS, MT Metrics: MTM, Passage Retrieval: PR (stopwords in parentheses).

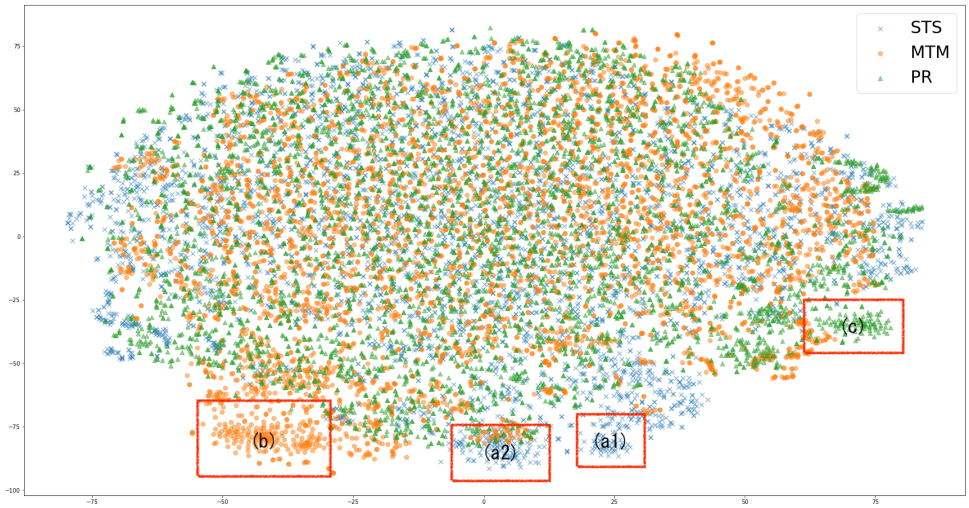
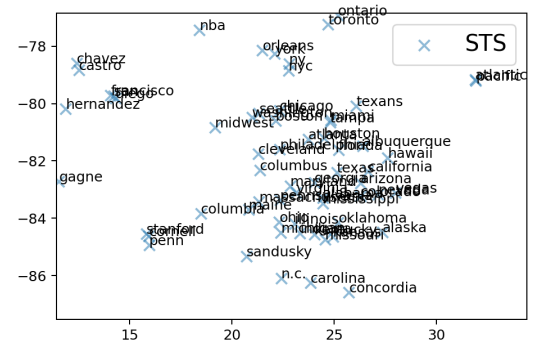
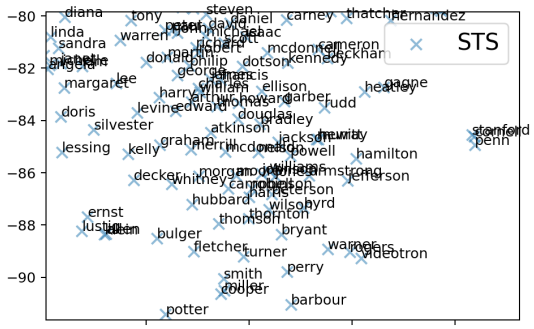


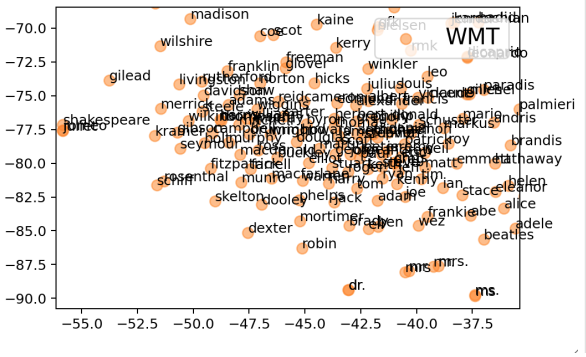
Figure 12: Word distribution of fasttext model in three datasets, STS (blue), MT Metrics (orange) and Passage Retrieval (green).



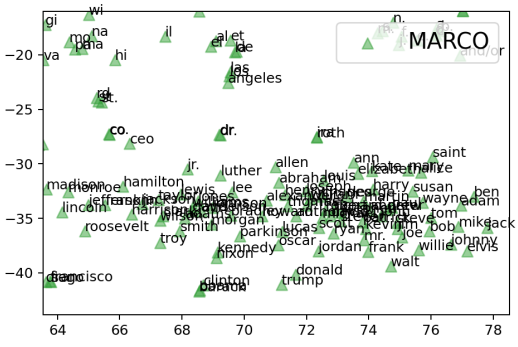
(a1) STS



(a2) STS



(b) Metrics



(c) Passage Retrieval

Figure 13: Expanded areas in the visualization of word distribution (Fig. 12).

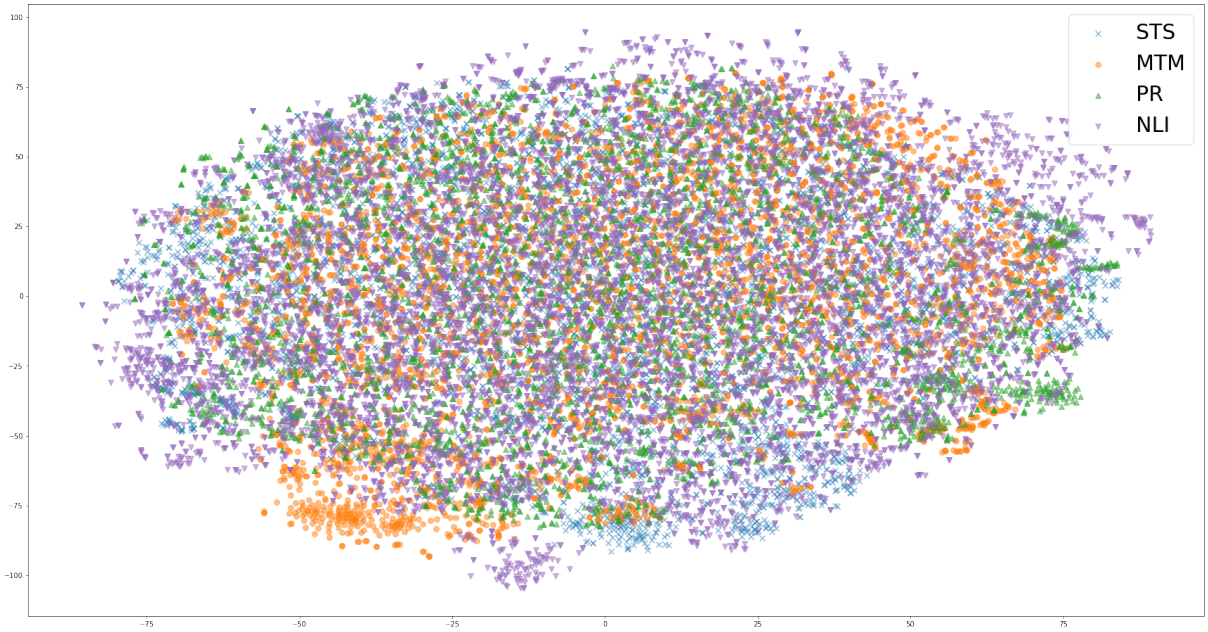


Figure 14: Word distribution of fasttext model in three datasets, STS (blue), MT Metrics (MTM: orange), Passage Retrieval (PR: green) and NLI (purple).

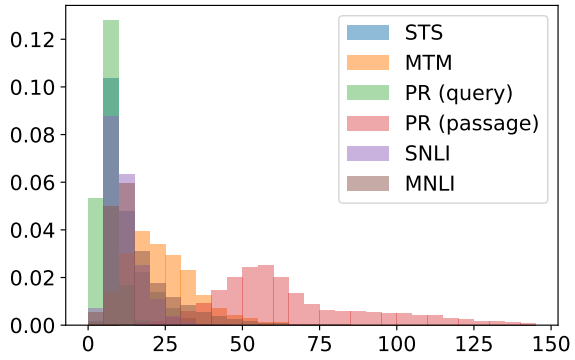


Figure 15: Histgrams of sentence length in the datasets includes NLI.

	SNLI	MNLI
#sentence pairs	570,152	402,703
#words	11,731,474	12,864,145
#types of words	37,179	85,789
TTR	0.0032	0.0067
Herdan's C	0.6465	0.6939
avg. zipffreq	2.871±1.488	2.685±1.448
avg. word len	7.544±2.613	8.206±3.313

Table 10: Statistics of vocabulary distribution on NLI datasets.

	model	dim	similarity function	pooling	others
SimCSE-sup	princeton-nlp/sup-simcse-bert-base-uncased	default	cos		
SimCSE-unsup	princeton-nlp/unsup-simcse-bert-base-uncased	default	cos		
SBERT-bb-NLI-mean	bert-base-nli-mean-tokens		cos	mean	
SBERT-MiniLM	all-MiniLM-L6-v2	384	cos	mean	
SBERT-mpnet	all-mpnet-base-v2	768	cos	mean	
BERTScore-rl-p	roberta-large	default	precision		
BERTScore-rl-r	roberta-large	default	recall		
BERTScore-rl-f	roberta-large	default	f1-score		
BERTScore-bbu-p	bert-base-uncased	default	precision		
BERTScore-bbu-r	bert-base-uncased	default	recall		
BERTScore-bbu-f	bert-base-uncased	default	f1-score		
avg. of BERT-bbl	bert-base-uncased	768	cos	mean	
avg. of BERT-rl	roberta-large	768	cos	mean	
BoV-Word2Vec (mean)	GoogleNews-vectors-negative300.magnitude	300	cos	mean	
BoV-Word2Vec (max)	GoogleNews-vectors-negative300.magnitude	300	cos	max	
BoV-Glove (mean)	glove.840B.300d.magnitude	300	cos	mean	
BoV-Glove (max)	glove.840B.300d.magnitude	300	cos	max	
BoV-fasttext (mean)	crawl-300d-2M.magnitude	300	cos	mean	
BoV-fasttext (max)	crawl-300d-2M.magnitude	300	cos	max	
BoW (sum)	CountVectorizer (sklearn, use smooth idf, stopwords)	vocab size	cos	sum	
BoW-TFIDF (sum)	TfidfVectorizer (sklearn, stopwords)	vocab size	cos	sum	norm=L2
USE	universal-sentence-encoder	512	cos		norm=L2
USE-l	universal-sentence-encoder-large	512	cos		norm=L2

Table 11: Semantic similarity model descriptions.