A Appendix

A.1 Limitation

Experiments on only English STS Although we would like to investigate other languages as well, we have only dealt with the original English STS (and PAWS in Appendix) in this paper. There are semantic similarity benchmark datasets for each language. Since the GLUE (Wang et al., 2019) 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, including a benchmark dataset focusing on semantic similarity. Semantic similarity benchmarks for each language are created in two ways: by automatic translation or re-constructed by each language expert following the original method. The former is likely to fundamentally face the same problems as those in English benchmarks, though including the issue of translation quality. Regarding the latter, some innovations sometimes be seen from the original constructing method, such as the addition of more detailed instructions on label definition when annotating the similarity by non-expert, in the Korean GLUE (KLUE) (Park et al., 2021). There is momentum for the creation of language-specific GLUEs, but it is necessary to make such considerations for an appropriate benchmarks before following the original method when creating datasets on one's own.

A.2 Statistics of datasets and subsets in the experiments

Statistics of entire datasets. Table 3 shows statistics on the number of sentence pairs (dataset size), the number of words and sentence length for STS, MTM and PR. 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 tasks.

Statistics of subsets. The created subset statistics of sentence length distribution are shown in Table 4. The values in Table 4 are the means of s1 (or hyp, query) and s2 (or ref, passage) taken over the whole subset. As shown in this, it can be confirmed that the MTM shorter subsets such (0, 40), (5, 45) as is the nearly same distribution with respect to STS one. Statistics of the subset of sentence length, vocabulary coverage, and the granularity of similarity are shown in Table 4, 5, and 6, respectively. The values in Table 4 are the means of s1 (or hyp, query) and s2 (or ref, passage) taken over the whole subset.

A.3 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` module (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 7 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 belongs to "corpus-freq $-$ wordfreq > 0.001" for STS, MTM, and PR, respectively, were 43, 18, and 26 words (if excluding stopwords and punctuation, 28, 3, and 6 words, respectively). The examples of higher frequent words in each dataset are shown in Table 8. As shown in this, some domain-specific words (STS: image caption, MTM: news, PR: Question Answering) are particularly frequent in each corpus. STS seems to be biased toward certain words (e.g., relatively abstract nouns such as man and dog, colors, present progressive forms). The results show that STS has a high occurrence rate of relatively "easy" vocabulary, especially in

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

	STS (s1, s2)	MTM (hyp, ref)	PR (query, passage)
#sentence pairs	8,628	3,793	6,668,967
#sentences ({s, s'})	17,256	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	6.1764±2.6423
#words / s'	**11.437±6.099**	23.467±11.138	65.640±26.692

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

	MTM		PR	
	size	avg. len	size	avg. 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. len" means the average of sentence length for each subset.

the image captioning domain, which makes the lexical difficulty of the entire corpus easier than the application tasks.

Gap of proper noun in word representation distribution In actual semantic similarity prediction models, words are embed into a multi-dimensional space and treated as a soft distributed representation. In the soft representation, whether STS vocabulary deviates from the vocabulary of the application tasks? We confirm whether the model that treat soft vocabulary representations still results in a bias in the lexical distribution.

We visualize 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. 9 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). The non-overlapping clusters were found to be mainly proper nouns such as Columbus (in detail, see Appendix). 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.

A.4 Word distribution analysis

Fig. 10 shows enlarged views of the areas surrounded with red lines in the visualization of word distributions (Fig. 9). These areas mostly includes several proper nouns such as Columbus, Carolina, and Robin in all the datasets.

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 data for architectural pre-training, noting that NLI with the lowest lexical overlap between the two sentences was the most effective in pre-training. 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.

Length analysis. Fig. 11 shows a histogram of sentence length distribution including NLI. In general, NLI datasets have a relatively short sentence length distribution, similar to that of STS. Although MNLI contains relatively longer sentences than SNLI, when compared to the distributions of MT Metrics and Passage Retrieval, it can be read that there are still fewer examples of longer sentences than in the other application datasets.

Vocab analysis. In following, we check the vocabulary distribution on the NLI datasets.

The statistics on NLI's vocabulary distribution are shown in Table 9. The Herdan's C of NLI is

	MTM				PR			
	size	avg. Recall(s, s')	avg. zipffreq	avg. word len	size	avg. Recall(s, s')	avg. zipffreq	avg. word len
all	3,793	0.882±0.084	3.680±1.544	7.471±2.790	6,614	0.835±0.079	3.220±1.404	7.334±3.165
High	100	1.000±0.000	5.110±1.004	5.706±2.422	100	0.988±0.011	4.858±0.986	5.955±2.471
Low	100	0.631±0.060	3.655±1.879	6.598±2.914	100	0.572±0.051	3.445±1.734	6.659±3.280

Table 5: stats. of vocabulary subsets for MTM and PR.

STS			MTM		
	size	avg. similarity		size	avg. similarity
[0, 1]	1182	0.655±0.280	Sim-Low: (-2, -0.47]	950	-0.820±0.266
(1, 2]	1348	1.631±0.285	Sim-MidLow: (-0.47, -0.03]	948	-0.240±0.126
(2, 3]	1672	2.653±0.291	Sim-MidHigh: (-0.03, 0.42]	943	0.193±0.127
(3, 4]	2317	3.614±0.287	Sim-High: (0.42, 1.5]	952	0.683±0.183
(4, 5]	1491	4.619±0.304	-	-	-

Table 6: Dataset size (#sentence pairs) and average & standard derivation of gold-standard similarity scores on STS and MTM subsets.

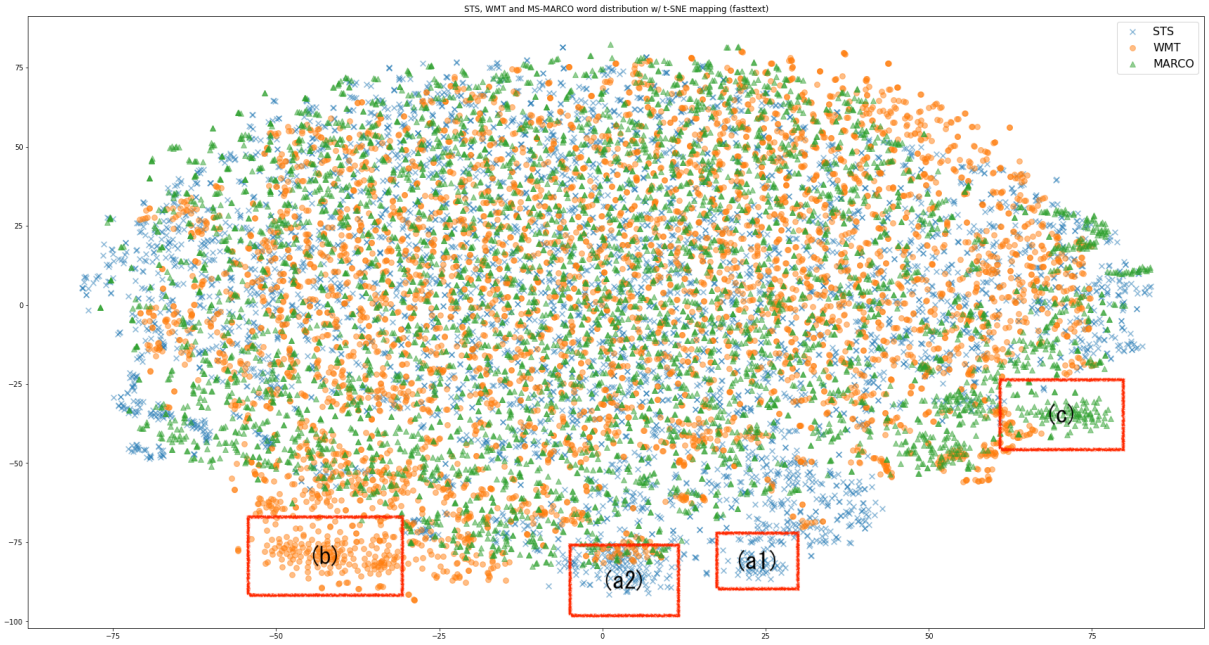


Figure 9: Word distribution of fasttext model in three datasets, STS (blue), MT Metrics (orange) and Passage Retrieval (green).

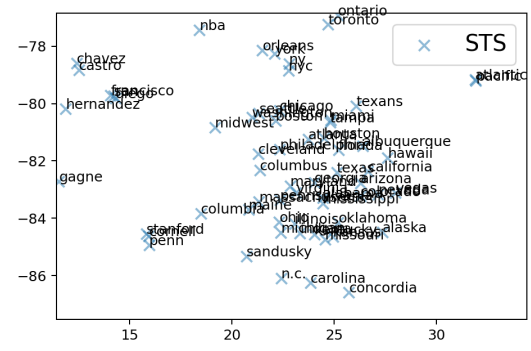
	STS	MTM	PR
zipffreq (↑)	**3.59±1.24**	3.45±1.54	1.29±1.74
length (↓)	**6.97±2.76**	7.34±2.83	10.1±4.83

Table 7: 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.

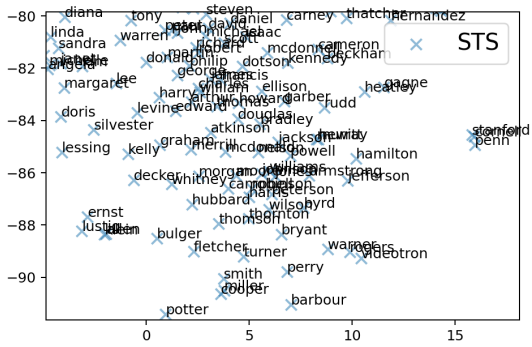
lower than that of STS and close to that of MT Metrics in TTR. As the word familiarity distribution of NLI, the average of zipffreq shows that more high-frequency words appear in both SNLI and MNLI than in STS. However, the average of word length of NLI is close to that of MT Metrics. These results indicate that although NLI has a fairly high frequency of occurrence, its word length distribution is on the longer side compared to STS. The visualization of the soft word distribution including NLI is shown in Fig. 12. As this figure shows, the actual distribution of the NLI vocabulary is such that it covers STS. This trend might contribute to the improvement of performances of NLI-supervised models such as SentenceBERT on

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, ⋯)

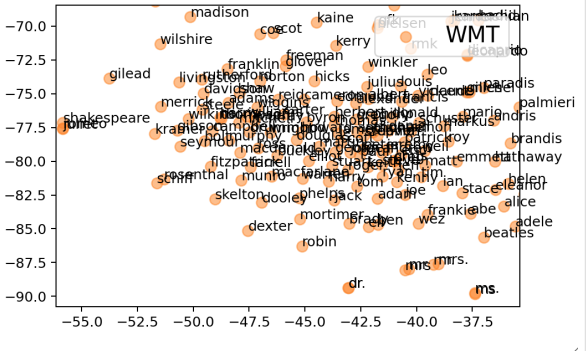
Table 8: Examples of higher frequency words for STS, MT Metrics: MTM, Passage Retrieval: PR (stopwords in parentheses).



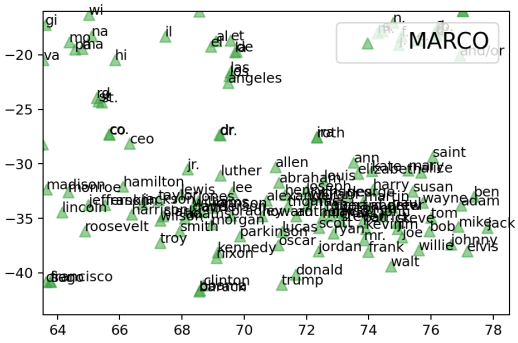
(a1) STS



(a2) STS



(b) Metrics



(c) Passage Retrieval

Figure 10: Expanded areas in the visualization of word distribution (Fig. 9).

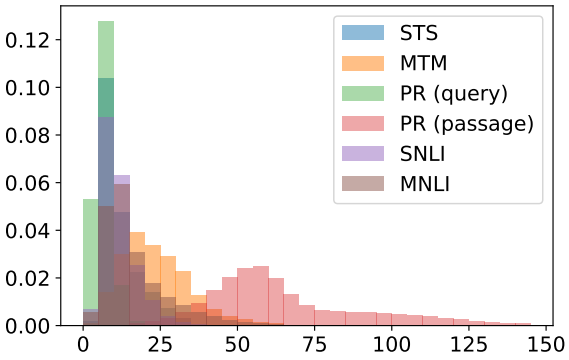


Figure 11: Histgram of sentence length in the datasets includes NLI.

	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 9: Statistics of vocabulary distribution on NLI datasets.

A.6 Model description

Table 10 shows the descriptions of the models used in this paper.

STS.

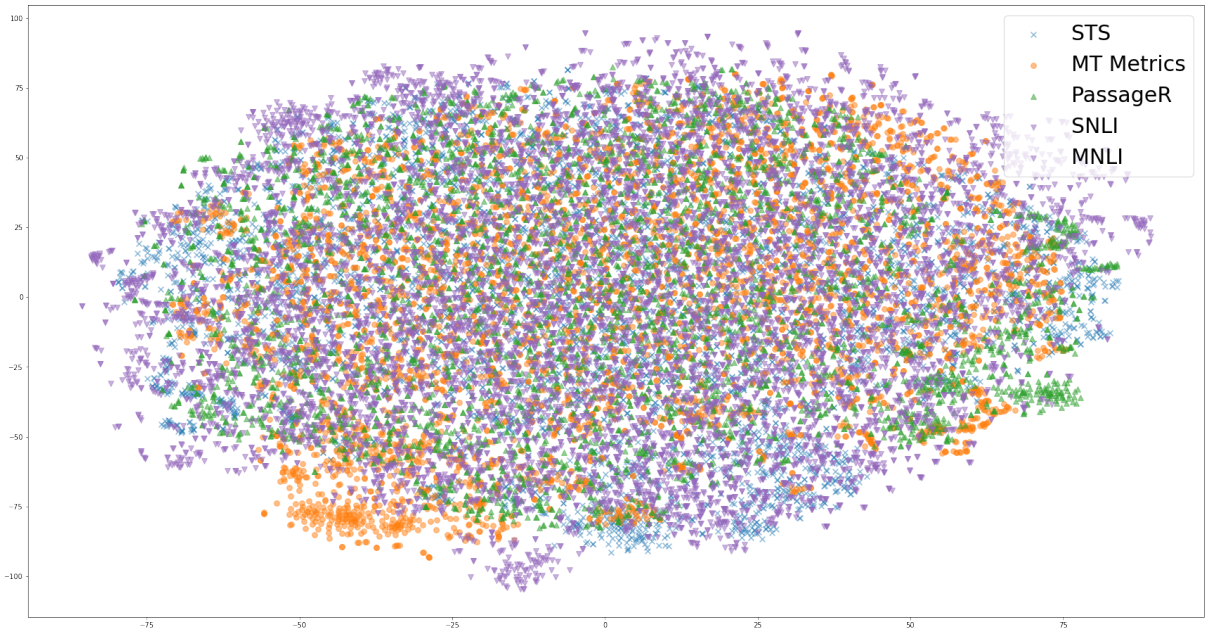


Figure 12: Word distribution of fasttext model in three datasets, STS (blue), MT Metrics (orange), Passage Retrieval (green) and NLI (purple).

	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 10: Text similarity model descriptions.