Sentence-Space Metrics (SSM) for the Evaluation of Sentence Comprehension

Jieyu Lin, Honghua Chen, Nai Ding*
Key Laboratory for Biomedical Engineering of Ministry of Education,
College of Biomedical Engineering and Instrument Sciences, Zhejiang University, Hangzhou, China
{jieyu_lin,honghuachen,ding_nai}@zju.edu.cn

Abstract

It is a fundamental challenge to evaluate whether a model can truly capture the meaning of sentences. Evaluation of whether a model well captures the meaning of individual words, however, can be effectively achieved by analyzing whether the model encodes words in a vector space where semantically similar words form clusters. Inspired by this approach, we propose the Sentence-Space Metrics (SSM) to evaluate model interpretation of sentences, and the sentence space is constructed based on the pairwise entailment relationships between all sentence pairs within a sentence pool. We use three metrics to evaluate a sentence space, i.e., (1) sparsity, (2) clustering of related sentences, and (3) similarity with the sentence space measured from humans. The SSM is applied to evaluate 20 models, including ChatGPT, 18 BERT-family models fine-tuned for Natural Language Inference (NLI) task, as well as SimCSE, a sentence representation model. The SSM reveals dramatic differences among models: Although all models achieve high accuracy on standard NLI datasets such as MNLI, none of them mirrors the human behavior under the SSM. These results demonstrate that, compared with traditional accuracy measures, the SSM considers pairwise relationships between hundreds of sentences and therefore provide a more fine-grained evaluation of model interpretation of sentences.

1 Introduction

How to represent the meaning of a word or a sentence is a classic question in philosophy, linguistics, and psychology (Rumelhart, 1986; Lund and Kevin, 1997; Martin, 2007; Mikolov et al., 2013b; Mikolov et al., 2013a). Converging evidence from Natural Language Processing (NLP) studies and human neuroimaging studies suggests that the meaning of a word can be well described by its relationship with other words. For example, word embedding algorithms project words into a semantic space in which words with more similar meanings are more closely located (Başkaya et al., 2013; Baroni et al., 2014; Th et al., 2015; Wang et al., 2019). Similarly, human neuroimaging studies have shown that semantic space for words can be constructed based on human rating of the similarities between words and predicts how similarly the brain responds to different words (Wang et al., 2018; Wang et al., 2020). In human language, words are only the basic elements to express meaning, while the unique expressive power relies on the construction of sentences - We use sentences to represent compositional semantic structures, including events and propositions. The complexity and abundancy in sentence meaning renders the evaluation of sentence meaning much more challenging than that of word meaning (Muennighoff et al., 2023).

Here, inspired by the analysis of word semantic space, we evaluate how well an NLP model captures sentence meaning by constructing a sentence space. Each sentence is a point in the vectorial space, and the distance between two sentences represents how strongly the two sentences are related. Just like the relationship between two words can be measured based on different features, e.g., animacy, abstractness, and part of speech information, the relationship between two sentences can be measured in many different ways. Here, we use the graded entailment score to quantify the relationship between two

* Corresponding author: Nai Ding

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 1334-1350, Taiyuan, China, July 25 - 28, 2024.
Volume 1: Main Conference Papers
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

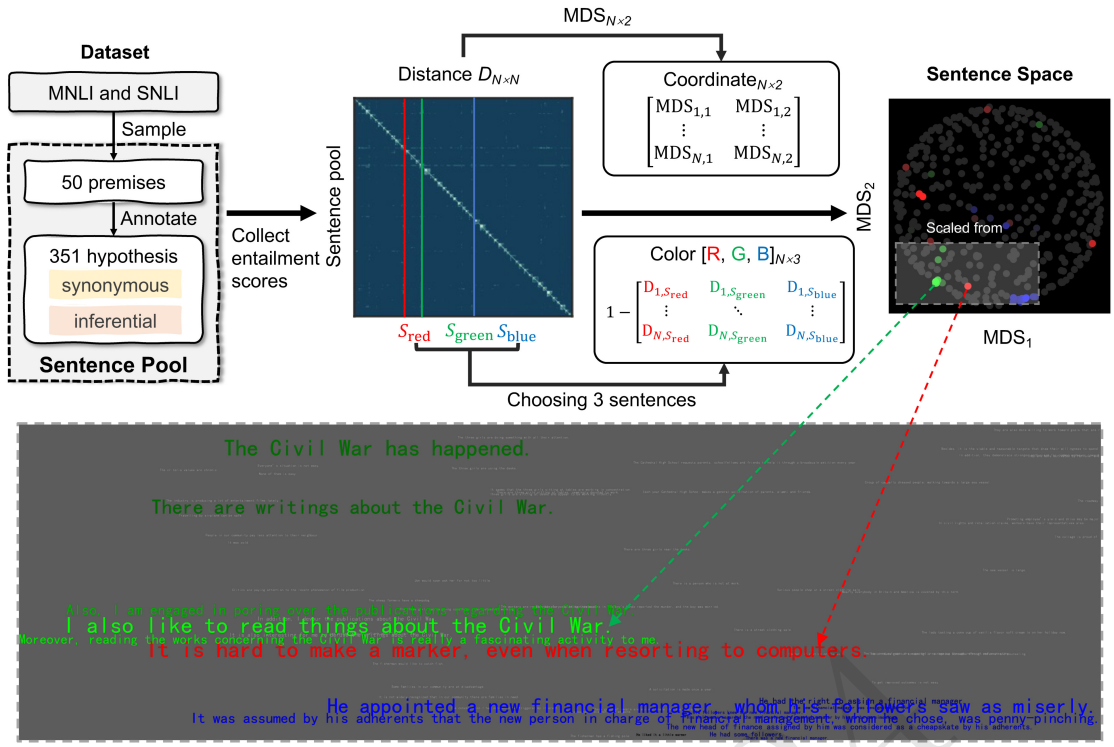

Figure 1: Construction of a semantic space for sentences. We construct a sentence pool, collect pairwise entailment scores to build a distance matrix, and use multidimensional scaling (MDS) to visualize the space. Each point in the sentence space in the upper right corner represents a sentence, and its color, i.e., RGB channels, illustrates how the sentence relates to 3 basis sentences that are randomly selected from the sentence pool. A portion of the sentence space is zoomed in in the lower panel, in which the font size is proportional to the shortest distance to a basis sentence – Larger fonts for sentences that are close to a basis sentence. For this illustration, the 3 basis sentences are: (R) Making a marker is a complicated task, even with modern computer assistance; (G) I also like to read things about the Civil War; (B) He appointed a new financial manager, whom his followers saw as miserly.

sentences (Jiang and de Marneffe, 2022). In other words, if a sentence could entail another sentence, the two sentences should have a shorter distance in the sentence space. We consider the semantic entailment score instead of other semantic similarity measures (Li et al., 2006; Mihalcea et al., 2006; Agirre et al., 2012) since semantic entailment is relatively more clearly defined while semantic similarity can be judged based on different sets of features (Deshpande et al., 2023). All the SSM, however, can be easily applied to sentence spaces constructed based on other measures.

Semantic entailment, also referred to as the Natural Language Inference (NLI) task, is widely used to evaluate the sentence comprehension ability of models (Bowman et al., 2015; Williams et al., 2018). Previous studies, however, mostly focus on the entailment relationship between pairs of sentences that are constructed either manually (Bowman et al., 2015; Williams et al., 2018; Nie et al., 2020) or based on templates (McCoy et al., 2019; Luo et al., 2022). For example, large datasets such as MNLI (Bowman et al., 2015) and SNLI (Williams et al., 2018) have been constructed by collecting premises from corpora, and asking annotators to construct hypotheses that have the required relationship with the premise (i.e., entailment, contradiction, or neutral). Transformers-based language models have achieved high accuracy on mainstream NLI datasets (Devlin et al., 2019; Liu et al., 2019; He et al., 2021), but their performance is sensitive to data corruption, distribution shift, and data manipulation (Poliak et al., 2018; Sanchez et al., 2018; Wallace et al., 2019a; Jin et al., 2020). Consequently, it is of significant interest to design more comprehensive methods to evaluate model performance on the NLI task (Naik et al., 2018; McCoy et al., 2019; Bartolo et al., 2020; Kiela et al., 2021; Liu et al., 2023). The SSM proposed here differs from

Premise	Hypothesis
Making a marker is a complicated task, even with modern computer assistance.	Producing a marker is complex, even with the help of computers. (0.93, 0.83)
Making a marker is a complicated task, even with modern computer assistance.	Producing a marker is even trickier without computers. (0.80, 0.39)
A fisherman with his friend is setting up his pole.	A fisherman and his friend are preparing his fishing rod. (0.7, 0.76)
A fisherman with his friend is setting up his pole.	The fisherman would like to catch fish. (0.76, 0.33)

Table 1: Examples of human-written hypotheses based on 2 premises. The numbers in parentheses represent the human entailment scores of the sentence pairs as well as the reversed pairs.

previous methods to evaluate sentence interpretation by jointly considering the relationships between all possible pairs of sentences within a sentence pool. Full paring of sentences naturally introduces sentence pairs that have diverse relationships, some of which can violate the heuristic cues in current large NLI datasets.

The procedure to construct a sentence space is illustrated in Figure 1. We constructed a sentence pool in which sentences fall into 50 semantically related clusters (6-11 sentences per cluster), calculated the distance between these sentences, i.e., the graded entailment score, and analyzed the space defined by the distance matrix. An example sentence space is visualized in Figure 1. We applied the method to test 20 language models, as well as human participants for comparison. For a model that well captures sentence meaning, we expect the following three properties. First, short distance should be sparse since most sentence pairs are unrelated. Second, semantically related sentences should form clusters. Third, the model sentence space should align with the human sentence space. Therefore, we use (1) sparsity, (2) clustering, and (3) similarity with human to evaluate a sentence space.

2 Method

2.1 Dataset

We construct a sentence pool, S, and perform the SSM based on S. The sentence pool S differs from standard NLI datasets in three ways. First, since the SSM aims to examine the clustering of semantically related sentences in a sentence space, sentences in S should form clusters, i.e., groups of sentences that are strongly related. Second, since sentence entailment is a directional measure, we distinguish *synonymous* sentences, i.e., two sentences can entail each other, and *inferential* sentences, i.e., one sentence can entail the other but not vice versa. Third, if there are N sentences in S, the SSM considers all possible pairs from S, i.e., N^2 pairs of sentences. Therefore, N must be relatively small to make the annotation of the entailment relationships between N^2 pairs feasible.

We initialize S by randomly selecting 80 premises from the dev split of SNLI and MNLI under the constraints that the sentence contains 5-20 words. All premises are proofread to ensure that they are grammatical, meaningful, and semantically irrelevant with each other. We then expand S with hand-written hypotheses. The 80 premises are divided into ten nonoverlapping sets, each containing 8 premises. Each set is given to an annotator and ten annotators are recruited in total. For each premise, the annotator has to write 2-5 *synonymous* sentences and 3-5 *inferential* sentences. See Table 1 for example. Since it may be challenging to write the hypotheses for some premises, the annotator is only required to finish the task for 5 out of the 8 premises. All the annotators are proficient in English and have passed CET-6, a standardized English test in China.

Altogether, 351 sentences are constructed after we filter out duplicated sentences and sentences that contain less than 3 content words. These 351 sentences, together with the 50 premises (39 from MNLI and 11 from SNLI), constitute the sentence pool S. Finally, S contains 401 sentences that form 50 semantically related clusters around the 50 initial premises (6-11 sentences per cluster). The 401 sentences

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 1334-1350, Taiyuan, China, July 25 - 28, 2024.
Volume 1: Main Conference Papers

lead to 401×401 sentence pairs, and each pair is denoted as (s_i, s_j) for $s_i, s_j \in S$, where s_i is the premise and s_j is the hypothesis. The sentence pairs fall into 5 categories: The first two categories are *inferential* pairs and *synonymous* pairs, which include the initial premise and the corresponding inferential and synonymous hypotheses written by human annotators. Another two categories are *inferential (reversed)* pairs and *synonymous (reversed)* pairs, which switch the premise and hypothesis in *inferential* and *synonymous* pairs. Sentence pairs not falling into the four categories are referred to as *irrelevant* pairs. The entailment relationship between a sentence pair is characterized by a graded measure between 0 and 1, i.e., $R(i, j)$ (Chen et al., 2020b). If s_i can entail s_j with high probability, $R(i, j)$ is near 1. In contrast, if the two sentences are unrelated or contradict each other, $R(i, j)$ is near 0. The pairwise entailment scores $R(i, j)$ form a matrix $R \in \mathbb{R}^{N \times N}$, where N is the sentence number in S, i.e., 401.

2.2 Human Entailment Score

We collect human entailment scores, i.e., $R(i, j)$ for human. Since it is time-consuming to annotate the entailment relationships between all possible 401×401 sentence pairs, we only annotate a subset of sentence pairs. This subset includes all *Inferential*, *Synonymous*, *Inferential (reversed)* and *Synonymous (reversed)* pairs (702 pairs in total), as well as *Irrelevant* pairs that are judged to have an entailment score higher than 0.5 by at least 9 of the 18 BERT-family models. In total, 3,855 pairs of sentences are selected. The instruction for annotators is "For each pair of sentences, can the second sentence be inferred by the first sentence?". They choose from four options: (A) cannot be inferred, (B) can be inferred with low probability, (C) can be inferred with high probability, or (D) can be inferred. For the four options, $R(i, j)$ is scored as 0, 0.3, 0.7, and 1, respectively. Each sentence pair is scored by 8-11 annotators. Annotations by annotators with entailment accuracy below 0.6 are removed. Finally, $R(i, j)$ for human is averaged across annotators, where outliers beyond 3 standard deviation from the mean are removed. For unlabeled data, we assign $R(i, j)$ as 1 for identical sentence pairs (since a sentence always entails itself), and as 0 for the remaining *Irrelevant* pairs.

2.3 Model

2.3.1 BERT-family NLI models

We use BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), and DeBERTa-v3 (He et al., 2021) as pre-trained models, and consider both the base version and large version. The pre-trained models are provided by Huggingface (Wolf et al., 2019) and fine-tuned based on an NLI dataset, which could be SNLI, MNLI, or a combination of SNLI, MNLI, and ANLI (Nie et al., 2020). For concision, we use ANLI to refer the combination of SNLI, MNLI, and ANLI in subsequent sections. During fine-tuning, each model takes in a pair of sentences (s_i, s_j), formats it as $[CLS, s_i, SEP, s_j, SEP]$, and encodes them together. The final embedding of CLS token is run through a linear layer to obtain three probability scores $(P^e_{i,j}, P^c_{i,j}, P^n_{i,j})$, each denoting the probability of the relationship between the sentence pair being entailment, contradiction, or neutral (Devlin et al., 2019; Falke et al., 2019). The training details are shown in Appendix A. Here, we only use $P^e_{i,j}$ as the entailment score, i.e., $R(i, j)$.

2.3.2 SimCSE

SimCSE provides a universal vectorial sentence representation (Gao et al., 2021). The SimCSE model used in this work is based on RoBERTa-large (Liu et al., 2019), and is fine-tuned on SNLI and MNLI with a contrastive loss (Chen et al., 2020a). The model encodes each sentence as a vector, and the entailment score, i.e., $R(i, j)$, is defined as the cosine similarity between a pair of sentences. SimCSE is a model to evaluate semantic textual similarity, instead of an NLI model. Therefore, the model cannot possibly discriminate the directionality of the entailment relationship. We consider this model since it can possibly describe the synonymity of sentences.

2.3.3 ChatGPT

ChatGPT is a large language model developed by OpenAI upon the InstructGPT (Ouyang et al., 2022), and we use the GPT-3.5-turbo API. Here, we directly ask ChatGPT to judge the entailment relationship between a sentence pair under a zero-shot setting, using the multiple-choice questions as same as in the

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 1334-1350, Taiyuan, China, July 25 - 28, 2024.
Volume 1: Main Conference Papers

human experiment. Each sentence pair is evaluated through an independent query to ChatGPT, to avoid the influence of previous samples. The prompts are shown in Appendix B.

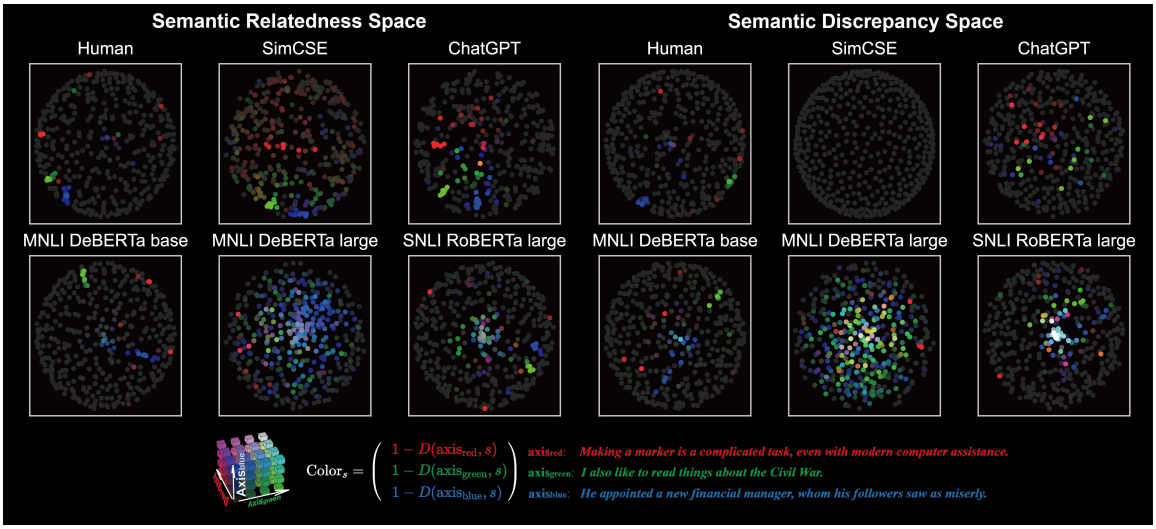

Figure 2: Visualization of the semantic relatedness space and semantic discrepancy space.

2.4 Entailment Accuracy

We turn the graded entailment score into a binary value to calculate the entailment accuracy. The BERT-family models are fine-tuned on the NLI task, and could predict a label, i.e., entailment or non-entailment (pooling the original 'contradiction' and 'neutral' labels). For human annotators, SimCSE, and ChatGPT, the relationship between a sentence pair is judged as entailment if and only if the entailment score is higher than 0.5. The mean human entailment score is taken as the ground truth.

2.5 Distance in Sentence Space

The entailment relationship between two sentences is directional – For a pair of sentences, one sentence may entail the other but not vice versa. In other words, there are two entailment scores $R(i,j)$ and $R(j,i)$ for each pair of sentences. We converted the two entailment scores into two unidirectional distances to construct two separate sentence spaces:

The semantic relatedness distance is defined as the average of the two entailment scores, i.e., $D_{rel}(i,j) = 1 - (R(i,j) + R(j,i))/2$, which describes general semantic similarity between two sentences. For example, the sentences on each row of Table 1 have relatively high entailment scores.

The semantic discrepancy distance is defined as the absolute difference of the two entailment scores, i.e., $D_{dis}(i,j) = 1 - |R(i,j) - R(j,i)|$, which describes how well the directional entailment relationship is captured. For example, on the even rows of Table 1, the premise entails the inferential sentence but not vice versa, resulting in a relatively large difference entailment score; while on the odd rows of Table 1, the sentences can entail each other and lead to a relatively low difference entailment score.

In the following, the calculations are applied to both D_{rel} and D_{dis}, and the two distance matrices are also both referred to as D.

2.6 2D Visualization of Sentence Space

We visualize the sentence space by showing each sentence as a colored dot in a 2-D space. The coordinate of a sentence is determined by the 2-D multidimensional scaling (MDS) (Mead, 1992) based on D. The RGB color indicates its distance to each of 3 basis sentences that are randomly chosen from S. For example, if s_{green} is selected as the basis for the green color, the value for the G channel is $1 - D(s_{\text{green}}, s_i)$ for $s_i \in S$. Consequently, sentences that have shorter distance to s_{green} and much longer distance to s_{red} and s_{blue} will have greenish color on the graph. If a sentence is close to none of the three basis sentences, its color is black. In contrast, if the sentence is close to all three basis sentences, its color is white.

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 1334-1350, Taiyuan, China, July 25 - 28, 2024.
Volume 1: Main Conference Papers

Model	Dataset	Synonymous	Synonymous (reversed)	Inferential	Inferential (reversed)	Irrelevant
BERT base	ANLI	0.748	0.771	0.650	0.791	0.460
	MNLI	0.733	0.756	0.645	0.791	0.720
	SNLI	0.664	0.763	0.686	0.786	0.382
BERT large	ANLI	0.840	0.847	0.714	0.809	0.417
	MNLI	0.794	0.832	0.664	0.795	0.691
	SNLI	0.802	0.832	0.714	0.768	0.388
DeBERTa base	ANLI	0.847	0.832	0.727	0.809	0.405
	MNLI	0.840	0.840	0.723	0.809	0.716
	SNLI	0.847	0.855	0.705	0.800	0.398
DeBERTa large	ANLI	0.840	0.840	0.723	0.809	0.436
	MNLI	0.809	0.855	0.686	0.800	0.692
	SNLI	0.855	0.847	0.718	0.805	0.419
RoBERTa base	ANLI	0.817	0.855	0.673	0.809	0.457
	MNLI	0.817	0.840	0.691	0.805	0.645
	SNLI	0.794	0.748	0.705	0.773	0.443
RoBERTa large	ANLI	0.802	0.824	0.718	0.809	0.530
	MNLI	0.824	0.817	0.691	0.805	0.495
	SNLI	0.878	0.832	0.700	0.827	0.414
SimCSE		0.893	0.885	0.686	0.468	0.806
ChatGPT		0.916	0.885	0.759	0.282	0.819
Human		0.805	0.809	0.780	0.792	0.841

Table 2: Accuracy for the entailment judgement.

The three basis sentences are only used for visualization and the quantitative analyses are based on all sentences as detailed in the following.

2.7 Sentence-Space Metrics (SSM)

2.7.1 Sparsity

In S, most sentence pairs are unrelated. Therefore, we expect a well-constructed sentence space to be sparse in the sense that most sentence pairs are unrelated and the distance between them should be near its maximal value, i.e., 1. We use the averaged l^1 norm (Hurley and Rickard, 2009) to quantify the sparsity of distance matrices $D \in \mathbb{R}^{N\times N}$. Formally, we calculate $\frac{\|1-D\|_1}{N^2}$. Smaller value indicates higher sparsity – The distance between most sentences is near 1.

2.7.2 Clustering

Clustering of semantically similar words is often used as a method to evaluate word vector space (Baroni et al., 2014). Here, since the hypotheses are written based on 50 premises, they should fall into 50 clusters, with all the hypotheses written for the same premise and premise itself belonging to the same cluster. To capture how well a model clusters semantically related sentences, we calculate the Calinski-Harabasz Index (CH) (Caliński and JA, 1974). Specifically, for each sentence s_i in S, its vector representation $d_i = [d_{i,1}, d_{i,2}, \cdots, d_{i,N}]$ is the corresponding row in the distance matrix, and it is classified into one of the 50 predefined clusters. The CH Index measures the ratio of the between-cluster isolation, the sum of the squared distance between the center of each cluster and the center of S, to the within-cluster coherence, the sum of the squared distance between each sentence and the center of its cluster. A larger value of CH indicates better cluster performance.

2.7.3 Similarity with Human

We compare the similarity of sentence space between the model and human using the representational similarity analysis (RSA) score (Kriegeskorte et al., 2006; Kriegeskorte et al., 2008). The RSA concerns whether the distance matrix D rated by human and models are consistent or not. Specifically, the RSA method flattens the upper diagonal part of D into a vector and calculate the Pearson correlation coefficient (Pearson, 1896; Huitson et al., 1976; Rodgers et al., 1988) between the vectors.

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 1334-1350, Taiyuan, China, July 25 - 28, 2024.
Volume 1: Main Conference Papers

Model	Dataset	Sparsity (L1)		Clustering (CH)		Similarity (RSA)	
		rel	dis	rel	dis	rel	dis
BERT base	ANLI	0.04	0.06	6.09	3.15	0.63	0.55
	MNLI	0.02	0.02	7.69	4.10	0.76	0.67
	SNLI	0.10	0.13	5.87	2.99	0.44	0.40
BERT large	ANLI	0.06	0.09	5.66	2.95	0.61	0.54
	MNLI	0.04	0.06	6.12	3.94	0.52	0.39
	SNLI	0.09	0.13	5.60	2.93	0.46	0.39
DeBERTa base	ANLI	0.05	0.07	5.79	2.85	0.63	0.56
	MNLI	0.03	0.04	8.48	4.08	0.79	0.71
	SNLI	0.08	0.10	4.95	2.68	0.51	0.45
DeBERTa large	ANLI	0.06	0.09	5.47	2.92	0.57	0.51
	MNLI	0.13	0.15	6.34	3.92	0.26	0.29
	SNLI	0.06	0.08	6.29	2.85	0.59	0.49
RoBERTa base	ANLI	0.04	0.07	6.50	3.07	0.67	0.58
	MNLI	0.04	0.05	7.60	3.78	0.73	0.63
	SNLI	0.09	0.14	5.56	3.41	0.41	0.32
RoBERTa large	ANLI	0.04	0.06	7.74	3.31	0.73	0.64
	MNLI	0.06	0.09	5.40	2.87	0.58	0.50
	SNLI	0.08	0.13	4.48	2.48	0.46	0.37
SimCSE		0.07	/	28.04	/	0.56	/
ChatGPT		0.07	0.07	12.50	3.25	0.50	0.12
Human		0.01	0.01	10.73	4.85	1.00	1.00

Table 3: Model performance under the SSM. Rel and dis represent the semantic relatedness space and the semantic discrepancy space, respectively.

3 Results

3.1 Entailment Accuracy

We first evaluate how accurately models could judge the entailment relationship between sentence pairs in our dataset (Table 2). Most models reach relatively high accuracy on human-written sentence pairs as well as the reversed pairs (*synonymous*, *synonymous (reversed)*, *inferential*, and *inferential (reversed)* pairs). For *synonymous* and *synonymous (reversed)* pairs, ChatGPT and SimCSE outperform BERT-family models but the difference is less than 15%. For *inferential (reversed)* pairs, however, ChatGPT and SimCSE perform much worse than BERT-family models (>30% difference). For SimCSE, $R(i,j) = R(j,i)$ and therefore the model cannot correctly judge the entailment relationship in *inferential (reversed)* pairs. ChatGPT, however, perform even worse than SimCSE in *inferential (reversed)* pairs, indicating a failure to detect bidirectional entailment relationship between sentences (see, e.g., the even rows of Table 1).

For *irrelevant* sentence pairs, the performance of BERT-family models is generally low, i.e., below 50% for more than half of the models, and their performance is not clearly higher for larger models (i.e., large vs. base model) and larger fine-tuning datasets (i.e., ANLI vs. MNLI/SNLI). The performance of SimCSE and ChatGPT is much higher than BERT-family models and is comparable to human performance.

3.2 Sentence-Space Metrics (SSM)

Next, we evaluate the models based on the SSM. We construct and visualize the sentence space following the procedure in Figure 1. Figure 2 separately visualize the semantic relatedness space and semantic discrepancy space for humans and a few representative models. The sentence space for other models is shown in Appendix C.

Since the three basis sentences are independently chosen and semantically irrelevant, in general, a sentence can possibly relate to at most one of them. In other words, we expect the color of most sentences to be black, and a few sentences could be red, green, or blue. For some BERT-family models, however, the sentence spaces tend to have numerous colored dots (Figure 2). More importantly, some dots are

nearly white, indicating the implausible condition that these sentences have short distances to all three basis sentences. The visualization in Figure 2 is colored based on three randomly chosen basis sentences. To more comprehensively evaluate the sentence space properties, we evaluate the entire sentence space under the SSM.

3.2.1 Sparsity

In the sentence pool, most sentence pairs are unrelated, and the distance between them is supposed to near its maximal value, i.e., 1. To characterize whether the distance is truly around 1 for most sentence pairs, we calculate the averaged l^1 norm of 1 minus the distance matrix (Table 3). As expected, the sentence space for humans has an averaged l^1 norm near 0, suggesting a highly sparse space. For models such as the MNLI BERT base, and MNLI DeBERTa base, the sentence space also has an averaged l^1 norm near 0, lower than the l^1 norm of SimCSE and ChatGPT. However, for some other models, e.g., MNLI DeBERTa large, the sentence space has a larger averaged l^1 norm, indicating a bias to judge unrelated sentences to have an entailment relationship.

3.2.2 Clustering

In our sentence pool, 5-10 hypotheses are composed for each of the 50 premises. Hypotheses associated with the same premise are semantically related and should cluster in the sentence space. We quantify whether these sentences truly cluster using the CH Index. Specifically, we define 50 categories based on the 50 premises and separately calculate the CH Index in the semantic relatedness and discrepancy spaces (Table 3). A CH Index near 0 indicates that the 50 clusters are not separated, suggesting semantically related sentences are not more close to each other in the sentence space compared with irrelevant sentences. In contrast, a higher CH Index indicates better separability of the 50 clusters. In the semantic relatedness space, the CH Index is the highest for SimCSE, ChatGPT, and humans. In the semantic discrepancy space, the CH Index is the highest for humans and a couple of BERT-family models.

3.2.3 Similarity with Human

Finally, we calculate the similarity of sentence space between the model and human by calculating the Pearson correlation coefficient between two distance matrices (Table 3). The results indicate that MNLI DeBERTa base is most similar to humans, followed by MNLI BERT base. The correlation between MNLI DeBERTa base and human reaches about 0.7 for both the semantic similarity and semantic discrepancy spaces. In semantic relatedness space, MNLI DeBERTa large perform the worst, suggesting that it could not well capture the degree of entailment. In semantic discrepancy space, ChatGPT perform the worst, suggesting that it cannot well capture the directional entailment relationship.

4 Related Work

4.1 Natural Language Inference

We consider a graded instead of categorical label, since Pavlick and Kwiatkowski (2019) and Zhang et al. (2021) found inherent disagreements in human textual inferences, and graded entailment score can capture more fine-grained relationship between sentences. Current NLI models are susceptible to distribution shifts and data manipulation (Hendrycks and Gimpel, 2017; Gururangan et al., 2018; Ebrahimi et al., 2018; Wallace et al., 2019a; Miller et al., 2020). To evaluate the models, previous studies have constructed challenging sets through, e.g., altering sentence structure (e.g., by changing the voice, clause structure, and negation etc.), replacing words with semantically similar or dissimilar words, altering sentence length by appending a tautology sentence (Naik et al., 2018; Glockner et al., 2018; McCoy et al., 2019; Luo et al., 2022; Liu et al., 2023). Previous studies have also used adversarial attack to test the robustness of models (Jia and Liang, 2017; Wallace et al., 2019a; Behjati et al., 2019; Jin et al., 2020). Consistent with our result, previous studies have also found that models tend to judge unrelated sentences to be entailment (Belinkov et al., 2019; Song et al., 2020; Lin et al., 2021; Luo et al., 2022).

To avoid predefined templates or fixed sentence patterns, other studies suggest to involve humans to construct challenging examples (Wallace et al., 2019b; Nie et al., 2020; Bartolo et al., 2020; Kiela et al., 2021). Our method only asks the annotators to generate a few hypotheses for each premise, not requiring

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 1334-1350, Taiyuan, China, July 25 - 28, 2024.
Volume 1: Main Conference Papers

the hypotheses to be challenging and therefore reducing the demand of human labor. The resulting sentence pairs exhibit natural variations in terms of syntactic, lexical, and semantic relationship. Since the SSM considers the relationship between all sentence pairs from a sentence pool, it involves more pairs of comparison. The large number of comparisons can reveal more detailed model performance, but also imposes a high labor cost for the labeling of ground truth. The cost of providing labels, however, is much lower than the cost to construct challenging samples.

Our results suggest that ChatGPT exhibits high accuracy on human-written pairs, but shows limited performance under the SSM. Qin et al. (2023) show that ChatGPT performs well but still underperforms fine-tuned models in two NLI datasets (i.e., RTE and CommitmentBank). Zhong et al. (2023) evaluate ChatGPT on GLUE benchmark, and find that ChatGPT have comparable performance with BERT-family models on NLI task. Some studies also suggest that ChatGPT has blind spots with certain types of samples (Basmov et al., 2023).

4.2 Sentence Vector Space

There are many ways to construct a sentence embedding. For example, some studies build upon the distributional hypothesis and learn a sentence embedding by predicting surrounding sentences (Kiros et al., 2015; Hill et al., 2016; Logeswaran and Lee, 2018). Other studies focus on deriving a sentence embedding by combining word embeddings (Mitchell and Lapata, 2008; Kalchbrenner et al., 2014; Wieting et al., 2016). Recently, contrastive training has been introduced to construct sentence embeddings (Gao et al., 2021; Ni et al., 2022). Previous studies evaluated the sentence embedding by a downstream task, e.g., semantic textual similarity. Here, however, we aim to utilize the sentence space to evaluate a downstream model.

5 Summary

We construct a sentence space to evaluate the sentence interpretation ability of models. The underlying philosophy of the method is that the meaning of a sentence is well described by its relationship with other sentences. Here, we use the graded entailment relationship between sentences as a distance measure to construct the sentence space, but the method can be readily used for other sentence similarity or relatedness measures. The evaluation results show that current models, including ChatGPT, still do not resemble the human ability to make language-based inference, even when the testing datasets do not involve challenging world knowledge or multi-media information. In sum, by constructing a sentence space and calculating metrics based on the space, we propose a new method that can more comprehensively evaluate sentence processing abilities of language models.

6 Limitations

Sentence spaces are only used for model evaluation, and future work will be directed toward explaining why model perform suboptimally in the sentence space and designing methods to improve model performance in the sentence space.

Acknowledgements

The study is supported by the National Key Research and Development Program of China (No. 2021ZD0204105).

References

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.

Marco Baroni, Georgiana Dinu, and Germán Kruszewski. 2014. Don't count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of the 52nd Annual Meeting of*

the Association for Computational Linguistics (Volume 1: Long Papers), pages 238–247, Baltimore, Maryland, June. Association for Computational Linguistics.

Max Bartolo, Alastair Roberts, Johannes Welbl, Sebastian Riedel, and Pontus Stenetorp. 2020. Beat the AI: investigating adversarial human annotation for reading comprehension. *Trans. Assoc. Comput. Linguistics*, 8:662–678.

Osman Başkaya, Enis Sert, Volkan Cirik, and Deniz Yuret. 2013. AI-KU: Using substitute vectors and co-occurrence modeling for word sense induction and disambiguation. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 300–306, Atlanta, Georgia, USA, June. Association for Computational Linguistics.

Victoria Basmov, Yoav Goldberg, and Reut Tsarfaty. 2023. ChatGPT and Simple Linguistic Inferences: Blind Spots and Blinds. *arXiv e-prints*, page arXiv:2305.14785, May.

Melika Behjati, SeyedMohsen MoosaviDezfooli, Mahdieh Soleymani Baghshah, and Pascal Frossard. 2019. Universal adversarial attacks on text classifiers. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2019, Brighton, United Kingdom, May 12-17, 2019*, pages 7345–7349. IEEE.

Yonatan Belinkov, Adam Poliak, Stuart M. Shieber, Benjamin Van Durme, and Alexander M. Rush. 2019. On adversarial removal of hypothesis-only bias in natural language inference. In Rada Mihalcea, Ekaterina Shutova, Lun-Wei Ku, Kilian Evang, and Soujanya Poria, editors, *Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics, *SEM@NAACL-HLT 2019, Minneapolis, MN, USA, June 6-7, 2019*, pages 256–262. Association for Computational Linguistics.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal, September. Association for Computational Linguistics.

Tadeusz Caliński and Harabasz JA. 1974. A dendrite method for cluster analysis. *Communications in Statistics - Theory and Methods*, 3:1–27, 01.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020a. A simple framework for contrastive learning of visual representations. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 1597–1607. PMLR, 13–18 Jul.

Tongfei Chen, Zhengping Jiang, Adam Poliak, Keisuke Sakaguchi, and Benjamin Van Durme. 2020b. Uncertain natural language inference. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8772–8779, Online, July. Association for Computational Linguistics.

Ameet Deshpande, Carlos E. Jimenez, Howard Chen, Vishvak Murahari, Victoria Graf, Tanmay Rajpurohit, Ashwin Kalyan, Danqi Chen, and Karthik Narasimhan. 2023. CSTS: conditional semantic textual similarity. *CoRR*, abs/2305.15093.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Javid Ebrahimi, Anyi Rao, Daniel Lowd, and Dejing Dou. 2018. Hotflip: White-box adversarial examples for text classification. In Iryna Gurevych and Yusuke Miyao, editors, *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 2: Short Papers*, pages 31–36. Association for Computational Linguistics.

Tobias Falke, Leonardo F. R. Ribeiro, Prasetya Ajie Utama, Ido Dagan, and Iryna Gurevych. 2019. Ranking generated summaries by correctness: An interesting but challenging application for natural language inference. In Anna Korhonen, David R. Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 2214–2220. Association for Computational Linguistics.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. Simcse: Simple contrastive learning of sentence embeddings. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 6894–6910. Association for Computational Linguistics.

Max Glockner, Vered Shwartz, and Yoav Goldberg. 2018. Breaking NLI systems with sentences that require simple lexical inferences. In Iryna Gurevych and Yusuke Miyao, editors, *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 2: Short Papers*, pages 650–655. Association for Computational Linguistics.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel R. Bowman, and Noah A. Smith. 2018. Annotation artifacts in natural language inference data. In Marilyn A. Walker, Heng Ji, and Amanda Stent, editors, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 2 (Short Papers)*, pages 107–112. Association for Computational Linguistics.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: decoding-enhanced bert with disentangled attention. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Dan Hendrycks and Kevin Gimpel. 2017. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

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, June. Association for Computational Linguistics.

Alan Huitson, Olive Jean Dunn, and Virginia A. Clark. 1976. Applied statistics: Analysis of variance and regression. *The Statistician*, 25:236–236.

Niall P. Hurley and Scott T. Rickard. 2009. Comparing measures of sparsity. *IEEE Trans. Inf. Theory*, 55(10):4723–4741.

Robin Jia and Percy Liang. 2017. Adversarial examples for evaluating reading comprehension systems. In Martha Palmer, Rebecca Hwa, and Sebastian Riedel, editors, *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*, pages 2021–2031. Association for Computational Linguistics.

Nanjiang Jiang and Marie-Catherine de Marneffe. 2022. Investigating reasons for disagreement in natural language inference. *Trans. Assoc. Comput. Linguistics*, 10:1357–1374.

Di Jin, Zhijing Jin, Joey Tianyi Zhou, and Peter Szolovits. 2020. Is BERT really robust? A strong baseline for natural language attack on text classification and entailment. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 8018–8025. AAAI Press.

Nal Kalchbrenner, Edward Grefenstette, and Phil Blunsom. 2014. A convolutional neural network for modelling sentences. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014, June 22-27, 2014, Baltimore, MD, USA, Volume 1: Long Papers*, pages 655–665. The Association for Computer Linguistics.

Douwe Kiela, Max Bartolo, Yixin Nie, Divyansh Kaushik, Atticus Geiger, Zhengxuan Wu, Bertie Vidgen, Grusha Prasad, Amanpreet Singh, Pratik Ringshia, Zhiyi Ma, Tristan Thrush, Sebastian Riedel, Zeerak Waseem, Pontus Stenetorp, Robin Jia, Mohit Bansal, Christopher Potts, and Adina Williams. 2021. Dynabench: Rethinking benchmarking in NLP. In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tür, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou, editors, *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 4110–4124. Association for Computational Linguistics.

Ryan Kiros, Yukun Zhu, Ruslan Salakhutdinov, Richard S. Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Skip-thought vectors. In Corinna Cortes, Neil D. Lawrence, Daniel D. Lee, Masashi Sugiyama, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 3294–3302.

Nikolaus Kriegeskorte, Rainer Goebel, and Peter Bandettini. 2006. Information-based functional brain mapping. *Proceedings of the National Academy of Sciences*, 103(10):3863–3868.

Nikolaus Kriegeskorte, Marieke Mur, and Peter A Bandettini. 2008. Representational similarity analysis-connecting the branches of systems neuroscience. *Frontiers in systems neuroscience*, page 4.

Yuhua Li, David McLean, Zuhair Bandar, James O'Shea, and Keeley A. Crockett. 2006. Sentence similarity based on semantic nets and corpus statistics. *IEEE Trans. Knowl. Data Eng.*, 18(8):1138–1150.

Jieyu Lin, Jiajie Zou, and Nai Ding. 2021. Using adversarial attacks to reveal the statistical bias in machine reading comprehension models. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 2: Short Papers), Virtual Event, August 1-6, 2021*, pages 333–342. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Wei Liu, Ming Xiang, and Nai Ding. 2023. Adjective scale probe: Can language models encode formal semantics information? *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(11):13282–13290, Jun.

Lajanugen Logeswaran and Honglak Lee. 2018. An efficient framework for learning sentence representations. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

Curt Burgess Lund and Kevin. 1997. Modelling parsing constraints with high-dimensional context space. *Language and cognitive processes*, 12(2-3):177–210.

Cheng Luo, Wei Liu, Jieyu Lin, Jiajie Zou, Ming Xiang, and Nai Ding. 2022. Simple but challenging: Natural language inference models fail on simple sentences. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Findings of the Association for Computational Linguistics: EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 3449–3462. Association for Computational Linguistics.

Alex Martin. 2007. The representation of object concepts in the brain. *Annu. Rev. Psychol.*, 58:25–45.

Tom McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In Anna Korhonen, David R. Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 3428–3448. Association for Computational Linguistics.

Andrew W. Mead. 1992. Review of the development of multidimensional scaling methods. *The Statistician*, 41:27–39.

Rada Mihalcea, Courtney D. Corley, and Carlo Strapparava. 2006. Corpus-based and knowledge-based measures of text semantic similarity. In *Proceedings, The Twenty-First National Conference on Artificial Intelligence and the Eighteenth Innovative Applications of Artificial Intelligence Conference, July 16-20, 2006, Boston, Massachusetts, USA*, pages 775–780. AAAI Press.

Tomás Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. In Yoshua Bengio and Yann LeCun, editors, *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*.

Tomás Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013b. Distributed representations of words and phrases and their compositionality. In Christopher J. C. Burges, Léon Bottou, Zoubin Ghahramani, and Kilian Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States*, pages 3111–3119.

John Miller, Karl Krauth, Benjamin Recht, and Ludwig Schmidt. 2020. The effect of natural distribution shift on question answering models. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 6905–6916. PMLR.

Jeff Mitchell and Mirella Lapata. 2008. Vector-based models of semantic composition. In Kathleen R. McKeown, Johanna D. Moore, Simone Teufel, James Allan, and Sadaoki Furui, editors, *ACL 2008, Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics, June 15-20, 2008, Columbus, Ohio, USA*, pages 236–244. The Association for Computer Linguistics.

Niklas Muennighoff, Nouamane Tazi, Loïc Magne, and Nils Reimers. 2023. MTEB: massive text embedding benchmark. In Andreas Vlachos and Isabelle Augenstein, editors, *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2023, Dubrovnik, Croatia, May 2-6, 2023*, pages 2006–2029. Association for Computational Linguistics.

Aakanksha Naik, Abhilasha Ravichander, Norman Sadeh, Carolyn Rose, and Graham Neubig. 2018. Stress test evaluation for natural language inference. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2340–2353, Santa Fe, New Mexico, USA, August. Association for Computational Linguistics.

Jianmo Ni, Gustavo Hernández Ábrego, Noah Constant, Ji Ma, Keith B. Hall, Daniel Cer, and Yinfei Yang. 2022. Sentence-t5: Scalable sentence encoders from pre-trained text-to-text models. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Findings of the Association for Computational Linguistics: ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 1864–1874. Association for Computational Linguistics.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2020. Adversarial NLI: A new benchmark for natural language understanding. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel R. Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 4885–4901. Association for Computational Linguistics.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *NeurIPS*.

Ellie Pavlick and Tom Kwiatkowski. 2019. Inherent disagreements in human textual inferences. *Trans. Assoc. Comput. Linguistics*, 7:677–694.

Karl Pearson. 1896. Mathematical contributions to the theory of evolution. iii. regression, heredity, and panmixia. *Philosophical Transactions of the Royal Society A*, 187:253–318.

Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. 2018. Hypothesis only baselines in natural language inference. In Malvina Nissim, Jonathan Berant, and Alessandro Lenci, editors, *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics, *SEM@NAACL-HLT 2018, New Orleans, Louisiana, USA, June 5-6, 2018*, pages 180–191. Association for Computational Linguistics.

Chengwei Qin, Aston Zhang, Zhuosheng Zhang, Jiaao Chen, Michihiro Yasunaga, and Diyi Yang. 2023. Is chatgpt a general-purpose natural language processing task solver? *CoRR*, abs/2302.06476.

Joseph Lee Rodgers, W Alan Nicewander, and David C. Blouin. 1988. Thirteen ways to look at the correlation coefficient. *The American Statistician*, 42:59–66.

David E Rumelhart. 1986. Learning internal representations by back-propagating errors. *Parallel distributed processing: Explorations in the microstructure of cognition*, pages 318–362.

Ivan Sanchez, Jeff Mitchell, and Sebastian Riedel. 2018. Behavior analysis of NLI models: Uncovering the influence of three factors on robustness. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1975–1985, New Orleans, Louisiana, June. Association for Computational Linguistics.

Congzheng Song, Alexander M. Rush, and Vitaly Shmatikov. 2020. Adversarial semantic collisions. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 4198–4210. Association for Computational Linguistics.

Muneeb Th, Sunil Kumar Sahu, and Ashish Anand. 2015. Evaluating distributed word representations for capturing semantics of biomedical concepts. In Kevin Bretonnel Cohen, Dina Demner-Fushman, Sophia Ananiadou, and Jun'ichi Tsujii, editors, *Proceedings of the Workshop on Biomedical Natural Language Processing, BioNLP@IJCNLP 2015, Beijing, China, July 30, 2015*, pages 158–163. Association for Computational Linguistics.

Eric Wallace, Shi Feng, Nikhil Kandpal, Matt Gardner, and Sameer Singh. 2019a. Universal adversarial triggers for attacking and analyzing NLP. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 2153–2162. Association for Computational Linguistics.

Eric Wallace, Pedro Rodriguez, Shi Feng, Ikuya Yamada, and Jordan L. Boyd-Graber. 2019b. Trick me if you can: Human-in-the-loop generation of adversarial question answering examples. *Trans. Assoc. Comput. Linguistics*, 7:387–401.

Xiaosha Wang, Wei Wu, Zhenhua Ling, Yangwen Xu, Yuxing Fang, Xiaoying Wang, Jeffrey R Binder, Weiwei Men, Jia-Hong Gao, and Yanchao Bi. 2018. Organizational principles of abstract words in the human brain. *Cerebral Cortex*, 28(12):4305–4318.

Bin Wang, Angela Wang, Fenxiao Chen, Yuncheng Wang, and C.-C. Jay Kuo. 2019. Evaluating word embedding models: Methods and experimental results. *CoRR*, abs/1901.09785.

Xiaoying Wang, Weiwei Men, Jiahong Gao, Alfonso Caramazza, and Yanchao Bi. 2020. Two forms of knowledge representations in the human brain. *Neuron*, 107(2):383–393.

John Wieting, Mohit Bansal, Kevin Gimpel, and Karen Livescu. 2016. Towards universal paraphrastic sentence embeddings. In Yoshua Bengio and Yann LeCun, editors, *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*.

Adina Williams, Nikita Nangia, and Samuel R. Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In Marilyn A. Walker, Heng Ji, and Amanda Stent, editors, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, pages 1112–1122. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *CoRR*, abs/1910.03771.

Shujian Zhang, Chengyue Gong, and Eunsol Choi. 2021. Learning with different amounts of annotation: From zero to many labels. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7620–7632, Online and Punta Cana, Dominican Republic, nov. Association for Computational Linguistics.

Qihuang Zhong, Liang Ding, Juhua Liu, Bo Du, and Dacheng Tao. 2023. Can chatgpt understand too? A comparative study on chatgpt and fine-tuned BERT. *CoRR*, abs/2302.10198.

Appendix A Training Details

Parameter	Training Data	BERT		RoBERTa		DeBERTa	
		base	large	base	large	base	large
Learning rate	MNLI	2E-05	2E-05	2E-05	6E-06	2E-05	6E-06
	SNLI	3E-05	3E+05	2E-05	6E-06	2E-05	5E-06
	ANLI	3E-05	3E-05	2E-05	6E-06	2E-05	5E-06
Train epochs	MNLI	3	3	3	2	3	2
	SNLI	2	2	3	2	2	2
	ANLI	2	2	3	2	2	2
Batch size	MNLI	32	32	32	64	64	32
	SNLI	32	32	32	64	64	32
	ANLI	32	32	32	64	64	32
Weight decay	MNLI	0.01	0.01	0.10	0.00	0.00	0.00
	SNLI	0.10	0.10	0.01	0.00	0.00	0.00
	ANLI	0.10	0.10	0.01	0.00	0.00	0.00

Table 4: Hyperparameters for fine-tuning on SNLI, MNLI, and ANLI.

Model	Training Data	SNLI	MNLI	ANLI (R1)	ANLI (R2)	ANLI (R3)
BERT base	ANLI	0.915	0.84	0.556	0.457	0.43
	MNLI	0.792	0.84	0.244	0.28	0.312
	SNLI	0.899	0.723	0.263	0.312	0.314
BERT large	ANLI	0.924	0.871	0.611	0.45	0.45
	MNLI	0.837	0.862	0.294	0.274	0.313
	SNLI	0.919	0.776	0.29	0.31	0.327
DeBERTa base	ANLI	0.936	0.909	0.73	0.505	0.511
	MNLI	0.892	0.902	0.475	0.329	0.342
	SNLI	0.933	0.842	0.39	0.328	0.335
DeBERTa large	ANLI	0.94	0.918	0.819	0.647	0.615
	MNLI	0.908	0.912	0.575	0.407	0.408
	SNLI	0.939	0.878	0.541	0.427	0.413
RoBERTa base	ANLI	0.844	0.879	0.59	0.446	0.406
	MNLI	0.841	0.878	0.302	0.308	0.27
	SNLI	0.91	0.788	0.321	0.325	0.316
RoBERTa large	ANLI	0.884	0.908	0.686	0.375	0.386
	MNLI	0.89	0.903	0.471	0.254	0.269
	SNLI	0.927	0.842	0.383	0.305	0.295

Table 5: The accuracy evaluated on standard dataset.

NLI fine-tuned models were initialized using pre-trained models and further fine-tuned on NLI datasets, implemented on the transformer package provided by Huggingface (Wolf et al. 2019). We fine-tune with Adam optimizer using the hyperparameters listed in table 4, where the training objective is the cross-entropy loss between the labels and the predictions. After fine-tuning, the models achieve promising accuracy in validation set, as shown in Table 5.

Appendix B Details for Entailment Scores Collection

In this work, nearly 200 human annotators were involved in, resulting in 33975 pairwise entailment scores in total. For ChatGPT, we adopted the prompt with zero-shot setting as following:

Instruction

Please judge whether the first sentence can deduce the second sentence in the following sentence pairs, please choose from the following four options:
A. cannot be inferred
B. low probability to be inferred
C. high probability to be inferred
D. can be inferred
First sentence: *A number of kids gets instruction from band instructor.*
Second sentence: *There is more than one person on the street.*

Appendix C The Sentence Spaces of All Models

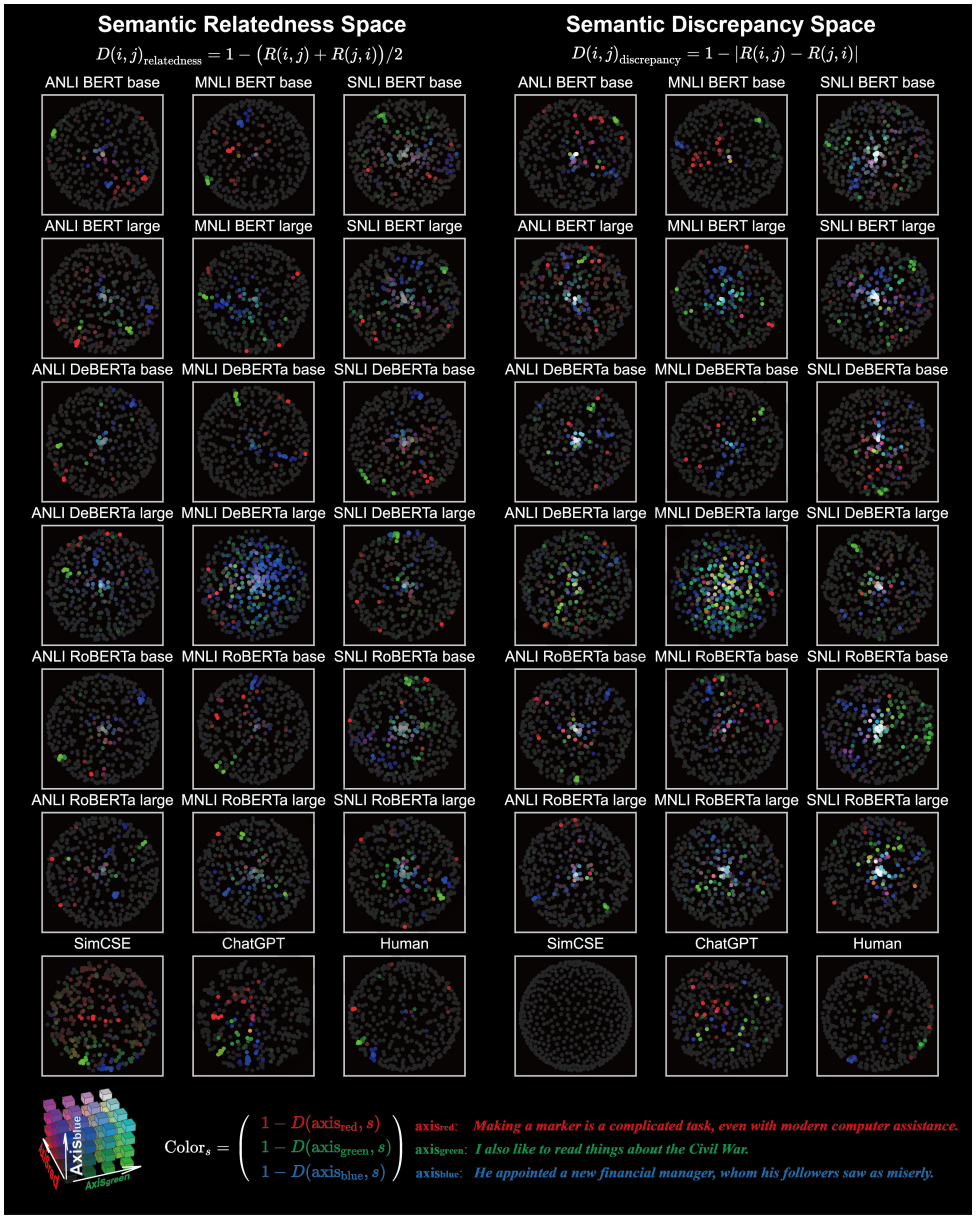

Figure 3: Visualization of the sentence space of semantic relatedness and semantic discrepancy respectively.