AccSent: Accurate Semantic Evaluation of Sentence Embeddings

Anonymous ACL submission

Abstract

Sentence embeddings, which encode arbitrary sentences as fixed-length numeric vectors, have shown promising results across a diverse range of semantic tasks. While prior work has demonstrated their effectiveness at capturing basic semantics, we present a new semantic evaluation set called *AccSent* for a more in-depth analysis of how accurately such embeddings reflect sentence semantics. We show that current embedding models are generally able to capture the broad semantic meaning of sentences, but that they are heavily affected by surface-level biases (such as lexical choices and sentence structures) instead of capturing the accurate semantic meaning of sentences. On our *AccSent* test set, sentence embedding models merely obtain an accuracy of 26.2% at evaluating the semantic similarity of sentences. We release our data and code on GitHub.

1 Introduction

Sentence embeddings (Conneau et al., 2017; Reimers and Gurevych, 2019; Li et al., 2020; Gao et al., 2021), which encode an arbitrary sentence as a fixed-length vector, have gained widespread use and achieved state-of-the-art results for NLP tasks such as measuring semantic textual similarity (STS; Reimers and Gurevych, 2019; Wang et al., 2022), finding matches in information retrieval (Kamalloo et al., 2023), evaluating translation quality (Rei et al., 2022), and supporting retrieval augmentated generation from large language models (Lewis et al., 2020). The promising results of sentence embeddings in downstream tasks prove that sentence embeddings are able to capture important aspects of sentence semantics, but it remains non-trivial to directly assess how accurately sentence embeddings capture detailed sentence semantics.

There are a number of methods designed to probe semantics contained in sentence embeddings. Adi et al. (2016) introduced a binary classification task that involves predicting whether a word w is contained in a sentence s based on its sentence embedding. Similarly, Conneau et al. (2018) proposed a classification task that requires predicting which one of a set of 1,000 words is included in a given sentence embedding. Others developed ways to investigate sentence embeddings by regenerating the sentence from left to right based on the sentence embedding (Bowman et al., 2015; Kerscher and Eger, 2020). Zhu et al. (2018) proposed contrastive similarity evaluations to assess to what extent sentence embeddings reflect negation.

$\mathbf{X}$	The **wildfire** in Japan last week killed 20 people.
$\mathbf{X}'$	20 people were killed by the **wildfire** last week in Japan.
$\mathbf{Y}$	The **tsunami** in Japan last week killed 20 people.

Table 1: The semantic similarity $\mathrm{sim}(\mathbf{X}, \mathbf{X}')$ between $\mathbf{X}$ and paraphrase $\mathbf{X}'$ should be higher than $\mathrm{sim}(\mathbf{X}, \mathbf{Y})$ and $\mathrm{sim}(\mathbf{X}', \mathbf{Y})$, since $\mathbf{Y}$ refers to a different event.

In this paper, we present a new semantic evaluation benchmark called *AccSent* and show that, even though sentence embedding models are generally able to capture the broad semantic meaning of sentences, they are heavily affected by surface-level biases (such as lexical choices and sentence structures) instead of capturing the accurate semantic meaning of sentences. *AccSent* includes sentences with near-identical semantics (such as $\mathbf{X}$, $\mathbf{X}'$ in Table 1) and also sentences that are semantically different (such as $\mathbf{X}$ vs. $\mathbf{Y}$ and $\mathbf{X}'$ vs. $\mathbf{Y}$). We used GPT4 to paraphrase sentences from an English NEWS corpus to obtain the near-equivalent sentences and created the semantically distinct sentences by manually changing an important word of the original NEWS sentences. We then evaluated the accuracy of state-of-the-art sentence embedding models at evaluating semantic similarity of different sentence pairs contained in *AccSent*. We find that sentence embedding models have a tendency to favor superficial biases as opposed to genuine semantics. For instance, $\mathbf{X}$, $\mathbf{Y}$ in Table 1 are semantically different but at the surface-level

keyword→replacement	keyword→replacement
cooperation→competition	celebration→contribution
nuclear→renewable	daily→monthly
Trump→Biden	lung→heart
vaccine→virus	south→north
wildfire→tsunami	knowledge→education
scientist→artist	integrity→beauty
Republicans→Democrats	convicted→accused
Jewish→Muslim	restrictions→suggestions
electronic→solar	terrorism→egotism
financial→cultural	unemployment→bankruptcy

Table 2: Keywords and their replacement words used to create our dataset.

differ in only one word and hence bear a significant degree of resemblance at a superficial level. We show that sentence embedding models are heavily biased by surface-level features and can only obtain an accuracy of 26.2% for evaluating the semantic similarity of sentences in *AccSent*. We release our dataset for future research.

2 *AccSent*: Accurate Sentence-Level Semantic Evaluation

Principle. Our new semantic evaluation set *AccSent* is designed to test how accurately sentence embeddings capture semantics as opposed to superficial similarities. For a sentence $\mathbf{X}$, *AccSent* provides two sentences $\mathbf{X}'$ and $\mathbf{Y}$ for semantic similarity evaluation. $\mathbf{X}$ and $\mathbf{X}'$ share a common meaning, but differ in lexical choice and sentence structure. In contrast, $\mathbf{Y}$ is obtained by changing an informative word in $\mathbf{X}$ and thus $\mathbf{Y}$ and $\mathbf{X}$ differ in just a single word, but have notably different semantics. We use sentence embedding models to predict the semantic similarity $\mathrm{sim}(\mathbf{X}, \mathbf{X}')$ between $\mathbf{X}$, $\mathbf{X}'$ and assess whether it is predicted to be greater than $\mathrm{sim}(\mathbf{X}, \mathbf{Y})$ and $\mathrm{sim}(\mathbf{X}', \mathbf{Y})$.

Data. To create our dataset, we need pairs of sentences with near-identical meaning. While a number of paraphrase resources exist (Dolan and Brockett, 2005; Freitag et al., 2020), upon closer inspection, many do not fulfill this criterion sufficiently well, as we find that the paraphrase sentences provided in these datasets frequently have notable semantic differences. Therefore, we used GPT4 to create paraphrase sentences for our dataset. We first collected sentences from an English NEWS corpus *News crawl 2020*[1] and then used GPT4 to paraphrase each collected sentence. When we collected sentences from the NEWS corpus, we used keyword-based collection. For each keyword shown in Table 2, we collected 50 short sentences (20–30 words) and 50 long sentences (50–60 words). We used 20 keywords in total, and therefore collected 1,000 short sentences and 1,000 long sentences in total. To create a paraphrase sentence for each collected sentence, we used GPT4 with the following prompt "Create a sentence with the exactly same meaning with this sentence: ...". We used the original sentence as $\mathbf{X}$ and the paraphrase sentence created by GPT4 as $\mathbf{X}'$. To obtain $\mathbf{Y}$, we changed the keyword contained in the original sentence into the replacement word as shown in Table 2, so that the meaning of the sentence diverges from that of the original.

[1] https://data.statmt.org/news-crawl/en/news.2020.en.shuffled.deduped.gz

Original
We will continue to condemn his horrific conduct and provide our full cooperation to law enforcement as it works to ensure that justice is served.
Paraphrase
We remain committed to denouncing his appalling behavior and will offer our complete support to the authorities in their efforts to see that justice is duly carried out.
Original
During a secret overnight flight to visit U.S. troops in Iraq on Christmas night in 2018, Trump sought input from Bolton and others on Air Force One about dumping Vice President Mike Pence from the 2020 ticket in favor of Nikki Haley, who had just stepped down as U.S. ambassador to the United Nations.
Paraphrase
While flying covertly to Iraq to meet with American soldiers on the night of Christmas in 2018, Trump consulted with Bolton and additional advisors aboard Air Force One regarding the possibility of replacing Vice President Mike Pence with Nikki Haley, the recently resigned U.S. ambassador to the United Nations, for the upcoming 2020 election campaign.

Table 3: Two paraphrase examples created by GPT4.

	short	long
cooperation	0	0
nuclear	0	5
Trump	1	1
vaccine	2	0
wildfire	0	3
average	0.6	1.8

Table 4: Human evaluation results for our *AccSent* dataset: numbers of invalid examples for each keyword.

Human Evaluation. With the keyword replacement method, we can make sure that $\mathbf{X}$ and $\mathbf{Y}$ are semantically different. However, $\mathbf{X}'$ is generated by GPT4. Although we find that GPT4 is good at paraphrasing as shown in Table 3, $\mathbf{X}'$ may not be a perfect paraphrase of $\mathbf{X}$ in some cases. To make sure that our *AccSent* dataset can be used as a valid semantic evaluation set, we performed a

human evaluation for 25% of our dataset, including both short and long sentences of 5 keywords as shown in Table 4. We asked 10 annotators to check whether $\mathbf{X}$ and $\mathbf{X}'$ are indeed semantically closer than $\mathbf{X}$ and $\mathbf{Y}$ for each sentence triplet. Each annotator evaluated sentences of one assigned keyword. Short and long sentences of the same keyword were assigned to two different annotators separately. Table 4 shows that there are roughly 0.6 out of 50 invalid short sentence examples for each keyword and there are roughly 1.8 out of 50 invalid long sentence examples for each keyword, which indicate that our dataset has a high quality and can generally be used as a valid test set for sentence-level semantic evaluation.

3 Experiments

3.1 Experimental Setup

We tested whether the following state-of-the-art sentence embedding models can capture the accurate meaning of the sentence using our *AccSent* evaluation set.

1. **SBERT:** From SBERT (Reimers and Gurevych, 2019), we consider the powerful *paraphrase-multilingual-mpnet-base-v2*[2] model to create embeddings with cosine similarity as the metric.
2. **LaBSE[3]:** Computes semantic similarity as the dot product of two embeddings and is mainly used for bitext mining (Feng et al., 2022).
3. **SimCSE:** From SimCSE[4], we use the *princeton-nlp/sup-simcse-roberta-large* model and compute the cosine similarity of two sentence embeddings as the similarity metric.
4. **COMET:**[5] Computes the similarity of two sentences using a regression model based on sentence embeddings and is mainly used for translation quality evaluation (Rei et al., 2022). We use the *Unbabel/wmt20-comet-qe-da* model[6].
5. **OpenAI:** OpenAI *text-embedding-ada-002* embeddings are provided by OpenAI[7] for diverse use cases. Cosine similarity is used to compute semantic similarity.

We tested these sentence embedding models for two semantic evaluation tasks using our dataset, Task 1: $\mathrm{sim}(\mathbf{X}, \mathbf{X}') > \mathrm{sim}(\mathbf{X}, \mathbf{Y})$ and Task 2: $\mathrm{sim}(\mathbf{X}, \mathbf{X}') > \mathrm{sim}(\mathbf{X}', \mathbf{Y})$.

[2] https://huggingface.co/sentence-transformers/paraphrase-multilingual-mpnet-base-v2

[3] https://huggingface.co/sentence-transformers/LaBSE

[4] https://github.com/princeton-nlp/SimCSE

[5] https://unbabel.github.io/COMET/

[6] Newer COMET models are not publicly available.

[7] https://platform.openai.com/docs/guides/embeddings

	$\mathbf{XX}' > \mathbf{XY}$			$\mathbf{XX}' > \mathbf{X}'\mathbf{Y}$		
	T_a	T_s	T_l	T_a	T_s	T_l
SBERT	0.202	0.293	0.111	0.920	0.951	0.890
LaBSE	0.034	0.058	0.011	0.938	0.957	0.919
SimCSE	0.223	**0.354**	0.092	0.919	0.965	0.873
COMET	**0.262**	0.302	**0.223**	0.832	0.846	0.818
OpenAI	0.065	0.093	0.038	**0.953**	**0.976**	**0.931**

Table 5: Accuracies of sentence embedding models for semantic evaluation using our *AccSent* test set. T_a/T_s/T_l denotes the accuracy for all/short/long sentences in the dataset.

3.2 Evaluation Results

Table 5 reports the accuracy of sentence embedding models in recognizing $\mathbf{X}$, $\mathbf{X}'$ as semantically more similar than $\mathbf{X}$, $\mathbf{Y}$ (or $\mathbf{X}'$, $\mathbf{Y}$). We observe that all tested models achieve a reasonably high accuracy (over 80%) for recognizing that $\mathbf{X}$, $\mathbf{X}'$ bear a greater semantic similarity than $\mathbf{X}'$, $\mathbf{Y}$, even for long sentences, which indicates that sentence embedding models can capture the general semantic meaning of sentences. However, Table 5 shows that sentence embedding models struggle to recognize that $\mathbf{X}$, $\mathbf{X}'$ are semantically closer than $\mathbf{X}$, $\mathbf{Y}$ (even for short sentences), most likely because, although $\mathbf{X}$, $\mathbf{Y}$ are semantically different, $\mathbf{X}$, $\mathbf{Y}$ are highly similar at the surface level (differing in just a single word). The fact that sentence embedding models generally predict $\mathbf{X}$, $\mathbf{Y}$ to be more similar than $\mathbf{X}$, $\mathbf{X}'$ shows that sentence embedding models exhibit a clear bias towards surface-level similarity (e.g., lexical choices and sentence structures) rather than genuine semantic similarity.

It is worth noting that the best-performing model COMET for the $\mathbf{XX}' > \mathbf{XY}$ task is the worst-performing model for the $\mathbf{XX}' > \mathbf{X}'\mathbf{Y}$ task. Also, the two best-performing models LaBSE and OpenAI for the $\mathbf{XX}' > \mathbf{X}'\mathbf{Y}$ task are the two worst-performing models for the $\mathbf{XX}' > \mathbf{XY}$ task, which suggests that it may be very difficult for a single sentence embedding model to capture the complete and accurate meaning of sentences and therefore different sentence embedding models tend to capture different sentence-level semantic features.

To obtain a more in-depth analysis of how sentence embedding models performed on our *AccSent* dataset, we show the results for subsets of the data with different keywords separately in Figure 1. Interestingly, we find that the keyword choice is a cru-

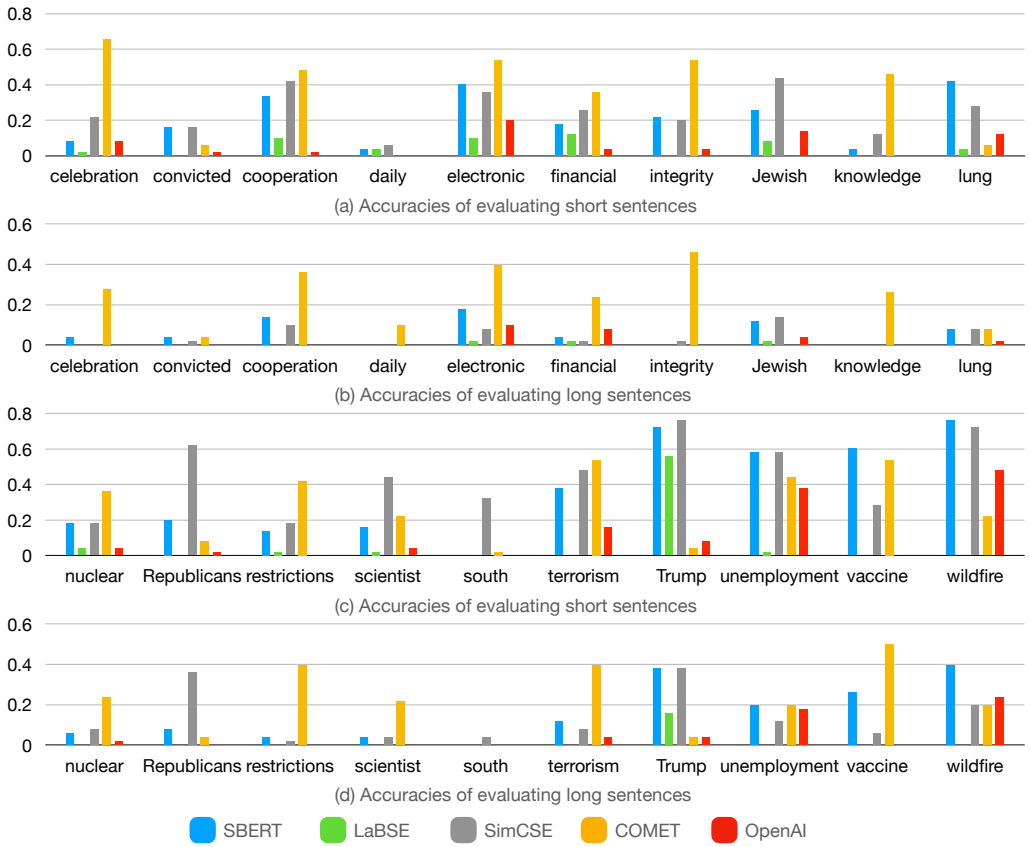

Figure 1: Accuracies of different sentence embedding models for evaluating short and long sentences with different keywords.

cial factor for the sentence embedding model performance. As shown in Figure 1, if a sentence embedding model was reasonably good at evaluating short sentences for a given keyword, then it most likely also got a relatively high accuracy for evaluating long sentences for the same keyword. The keyword choice affected the sentence embedding accuracies more significantly than the influence of sentence length for sentence embedding accuracies. Figure 1 also shows that different sentence embedding models performed very differently for different keywords (e.g., COMET performed reasonably well across most of the considered keywords, but obtained quite low accuracies for “Trump", while SBERT and SimCSE obtained relatively high accuracies for “Trump"), which again could suggest that different sentence embedding models tend to learn different types of semantic features of sentences. This could be a result of the capacity limitation of sentence embedding models, which can be improved by scaling up the model size along with larger and more diverse training data.

Comparison with Previous Work. Compared to previous work that probed general semantics contained in sentence embeddings, including the popular STS tasks, our *AccSent* evaluation set studies in particular how surface-level biases affect the effectiveness of sentence embedding models for capturing real semantics. Our experiments clearly show that current sentence embedding models are prone to neglect semantic changes in the sentences when surface-level biases exist, for both short and long sentences, which reveals a substantial shortcoming of current sentence embedding models.

4 Conclusion

This paper presents *AccSent*, a new dataset for assessing how accurately sentence embeddings are able to capture detailed sentence semantics. Our evaluation encompasses several state-of-the-art sentence embedding models, revealing that the effectiveness of sentence embedding models is heavily affected by surface-level biases and the types of semantic differences between sentences. Being affected by surface-level biases, the best-performing model COMET for our dataset only obtained an accuracy of 26.2% at making correct sentence-level semantic similarity predictions.

Limitations

Currently, *AccSent* only contains English sentences and we plan to extend *AccSent* to other languages so that *AccSent* can be used to evaluate sentence embedding models of different languages.

References

Yossi Adi, Einat Kermany, Yonatan Belinkov, Ofer Lavi, and Yoav Goldberg. 2016. Fine-grained analysis of sentence embeddings using auxiliary prediction tasks. *arXiv preprint arXiv:1608.04207*.

Samuel R Bowman, Luke Vilnis, Oriol Vinyals, Andrew M Dai, Rafal Jozefowicz, and Samy Bengio. 2015. Generating sentences from a continuous space. *arXiv preprint arXiv:1511.06349*.

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. 2017. Supervised learning of universal sentence representations from natural language inference data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 670–680, Copenhagen, Denmark. Association for Computational Linguistics.

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

William B. Dolan and Chris Brockett. 2005. Automatically constructing a corpus of sentential paraphrases. In *Proceedings of the Third International Workshop on Paraphrasing (IWP2005)*.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891, Dublin, Ireland. Association for Computational Linguistics.

Markus Freitag, David Grangier, and Isaac Caswell. 2020. BLEU might be guilty but references are not innocent. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 61–71, Online. Association for Computational Linguistics.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. Simcse: Simple contrastive learning of sentence embeddings. *arXiv preprint arXiv:2104.08821*.

Ehsan Kamalloo, Xinyu Zhang, Odunayo Ogundepo, Nandan Thakur, David Alfonso-hermelo, Mehdi Rezagholizadeh, and Jimmy Lin. 2023. Evaluating embedding APIs for information retrieval. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 5: Industry Track)*, pages 518–526, Toronto, Canada. Association for Computational Linguistics.

Martin Kerscher and Steffen Eger. 2020. Vec2Sent: Probing sentence embeddings with natural language generation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1729–1736, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474. Curran Associates, Inc.

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

Ricardo Rei, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. 2022. COMET-22: Unbabel-IST 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Bin Wang, C.-C. Jay Kuo, and Haizhou Li. 2022. Just rank: Rethinking evaluation with word and sentence similarities. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6060–6077, Dublin, Ireland. Association for Computational Linguistics.

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