Evaluation of Large Language Model Sentence Embeddings with SentEval

Amritanj Ayush, Evan Ozaroff, Parisa Torshizi

Abstract

The generation of sentence embeddings is important for numerous NLP tasks including semantic textual similarity, semantic search, and paraphrase mining, and is reliant on the model's ability to extract the unique semantic information of a sentence. Many large pre-trained language models (PLMs) can be used to generate sentence embeddings from text, either as a direct output of the model or extracted from an intermediate hidden state. Naturally, the quality of these sentence embeddings can vary significantly.

The quality of sentence embeddings can be determined by extrinsic evaluation, which considers the effectiveness of the embeddings on downstream tasks, or intrinsic evaluation, which utilizes pre-determined ground truth labels. Our research compares the performance of numerous large language models such as BERT and GPT-2 on intrinsic probing tasks implemented by the SentEval library. Our analysis found that the BERT family of models outperformed other PLMs such as MPNet and GPT-2 on the SentEval intrinsic probing of sentence embeddings, with the base BERT model performing best.

1 Introduction

Natural language processing has made significant progress over the years, with Pre-trained Language Models (PLMs) being at the forefront of this advancement. These models have revolutionized NLP by enabling transfer learning, where models can be pre-trained on large amounts of text data and then fine-tuned for specific downstream tasks.

One important application of PLMs is the generation of sentence embeddings, which are numerical representations of sentences that can capture their semantic meaning. By generating high-quality sentence embeddings, these models can understand the semantic nuances of language, allowing for better analysis. However, the quality of these embeddings can vary significantly across different models.

High-quality sentence embeddings are essential as they can directly impact the performance of downstream models that rely on these representations. Better embeddings lead to improved text understanding, enabling models to more accurately perform tasks like text classification, and question-answering. Additionally, high-quality embeddings allow models to generalize better across different domains and adapt to new tasks more effectively.

We evaluated and compared the performance of seven large language models, namely BERT, MPNet, RoBERTa, ALBERT, DistilBERT, BigBird-RoBERTa, and GPT-2. Our goal was to identify which models produce high-quality sentence embeddings and to understand the factors that contribute to their performance. To achieve this, we used SentEval, a benchmarking tool that offers several probing tasks designed to evaluate the quality of embeddings. The main assumption behind probing tasks is that if a linear model is able to predict linguistic information, then the corresponding embeddings could encode this information as well [1].

2 Background/Related Work

In prior research, Conneau Alexis et al. [2] used the SentEval probing tasks to study embeddings generated by three encoders trained in numerous different ways and evaluate the resultant properties of these embeddings. The study was primarily concerned with two neural network encoders, BiLSTM-last/max and Gated ConvNet.

Our research is concerned with applying the same SentEval probing tasks to a slate of contemporary large language models, expanding upon their analysis.

3 Data

The datasets used in this project are taken from the SentEval library. SentEval, developed by Facebook AI Research, provides a set of probing tasks, along with their datasets to evaluate sentence embeddings. Probing tasks attempt to evaluate how much information can be extracted from given embeddings by fitting the predictor models to the embeddings and evaluating the accuracy of these predictors on labeled data.

The datasets for each probing task contain around 100k training samples, 10k validation samples and 10k test samples. Each row of the data contains three columns, the train/validation/test partition, the ground truth label, and the sentence. Due to computational constraints, we were forced to work with a subset of about 1/10th of the data for each task. A brief description of these probing tasks is given below:

Sentence Length:
This task is a classification task, and the labels in its data are bins referring to length intervals.

Word Content:
This is also a classification task, and the curated labels are target words picked among the most common words in the vocabulary.

Tree Depth:
Tree depth is another classification task in which labels are the minimum depth of the sentence's syntactic tree.

Bigram Shift:
This task classifies sentences by whether two succeeding tokens are being inverted or not.

Top Constituent:
This task identifies the existence of the constituent patterns in each sentence. These constituent patterns are the top most common patterns in the whole corpus.

Past Present:
Past present classifies sentences based on the tense of their main verb.

Subject Number:
This task labels sentences based on their subject being either singular or plural.

Object Number:
This task labels sentences based on their subject being either singular or plural.

Odd Man Out:
This binary task, labels sentences with the labels of being either original or changed, meaning a token has changed in the sentence compared to the source corpus.

Coordination Inversion:
In this task, if the two clausal conjoints in a sentence have been inverted, the sentence is labeled as inverted. Otherwise it is labeled as original.

4 Methods

For our study, we compare the quality of sentence embeddings produced by the following pre-trained large language models - family of BERT transformers, GPT-2, and MPNet.

BERT, developed by Google AI Language Team, is a transformer-based model pre-trained on large corpora such as BooksCorpus and English Wikipedia, using a masked language modeling and next sentence prediction task. BERT has approximately 110M parameters and uses WordPiece tokenization, which breaks down words into subwords and can handle out-of-vocabulary words. Many organizations and individuals have developed models which are built on the architecture and mechanism of BERT, forming the family of BERT transformers. A few such models that we have selected for our study are explained below.

RoBERTa (Robustly Optimized BERT approach), developed by Facebook AI Research, is another transformer-based model that improves upon BERT by using a larger pre-training corpus and modifying the pre-training process. It has about 125M parameters and uses dynamic masking as a pre-training task. The next-sentence prediction task is removed from its pre-training tasks. RoBERTa is trained on 160 GB of text, which includes the same corpora as BERT and additional web texts.

ALBERT (A Lite BERT), developed by Google Research, is a compact version of BERT that uses parameter sharing and factorization techniques to reduce the number of parameters. ALBERT also introduces a self-supervised loss that encourages cross-layer parameter sharing. The smallest version of ALBERT has 4 times fewer parameters than the base BERT model.

DistilBERT (Distilled BERT), developed by Hugging Face, is a smaller and faster version of BERT that is distilled from the base BERT model using a knowledge distillation technique. DistilBERT has approximately 40% fewer parameters than the base BERT model.

BigBird-RoBERTa, developed by Google Research, is a sparse-attention based transformer which extends Transformer based models, such as BERT to much longer sequences. Moreover, BigBird comes along with a theoretical understanding of the capabilities of a complete transformer that the sparse model can handle. It is a pretrained model on English language using a masked language modeling (MLM) objective and has approximately 86M parameters. BigBird-RoBERTa is pre-trained on a large corpus of text, similar to RoBERTa, but with additional web texts and other domain-specific texts.

MPNet, developed by Microsoft Research, is a transformer-based model with approximately 84 million parameters, pre-trained on a large corpus with a multitask learning objective that includes various language understanding tasks. Unlike BERT, MPNet uses byte pair encoding (BPE) tokenization, which allows for the representation of rare words and out-of-vocabulary tokens. Both models have similar architectures based on a transformer encoder with multiple layers, but MPNet's training process includes additional language understanding tasks, which may contribute to its improved performance in certain natural language processing tasks.

GPT or Generative Pre-trained Transformer is a neural language model that was first introduced by OpenAI in 2019. It consists of two training steps. In the first step, the model leverages the abundance of unlabeled corpora by training on this data and then initializing the model’s parameters. In the second step, the GPT model which was previously pre-trained is trained on labeled task-specific data. This approach addresses the challenge of scarcity of labeled data and thus, trains a more robust model on them [3]. This learning scheme allows the model to learn the inner representations of the language and be able to further extract features for various downstream tasks. The model that was used in this project, is the smallest version of GPT-2, with 124M parameters. GPT-2 is a modification of the GPT model. In GPT-2 an additional normalization layer was added to the final self-attention block , the vocabulary was expanded, and the context size and batch size was also increased . GPT-2 was trained on WebText, a curated dataset of webpages of Reddit. In order to maintain an appropriate quality, only the web pages that had more than three karmas were picked [4]. In the preprocessing, the texts were tokenized using Byte Pair Encoding (BPE) having the vocabulary size of 50257 tokens. This model was made available to use through the Hugging Face library. In addition to the GPT-2 raw model, Hugging Face provides other fine-tuned versions of GPT-2, including DialoGPT, a fine-tuned model for multiturn human conversation tasks. One limitation of the GPT-2 is that it may be prone to bias as it was trained on unfiltered data. it is recommended that these models should be approached with caution in human-sensitive cases.

To explain in more detail, in the first step of the training, GPT uses a standard language modeling objective to maximize the likelihood of predicting the next word given the previous k words. With k being the size of a context window as an input and its following word being the output. A multi-layer Transformer decoder, a variant of the transformer, was added to the language model. This applies a multi-headed self-attention over the input tokens. In the second step of the training, the supervised fine-tuning, the new inputs are given to the model. And then the outputs of the last transformer block’s activation are fed into a linear layer to predict the final outputs of the model. The objective of the second layer is similar to the first layer, achieved by using gradient descent.

GPT-3 is another variation of the GPT model, introduced in 2020 by OpenAI. This new model is an autoregressive language model with 175 billion parameters, 10x more than previously proposed language models. Unlike GPT-2, GPT-3 doesn’t need fine-tuning or gradients, but it is trained with the few-shot learning. And interestingly, this few-shot approach proved to perform strongly on many tasks, including question-answering, and specifically tasks that require on-the-fly reasoning [5]. The GPT-3 model is reachable via the OpenAI REST API, which has limitations in terms of the number of user requests per minute. Although we were able to retrieve sentence embeddings for single texts from this model, we were not able to apply it on the probing tasks due to reaching the rate limit.

These models were selected due to their popularity and superior performance on various NLP benchmarks. We intend to evaluate the ability of these models to extract linguistic information from textual inputs by running them on SentEval's probing tasks.

5 Experiments

We compare the quality of sentence embeddings produced by the models described using the SentEval library. The SentEval predictors fitted for each task were selected to be logistic regression classifiers that take as an input the sentence embeddings and predict the class labels. Training was performed for 4 epochs and validation was performed using 10 folds.

5.1 Results

Prior to evaluating our models performance on the SentEval probing tasks, we perform a qualitative analysis of our sentence embeddings using the SentEval datasets. We investigate the models' ability to encode linguistic information in sentence embeddings by observing the t-SNE projections of two models, BERT and MPNet.

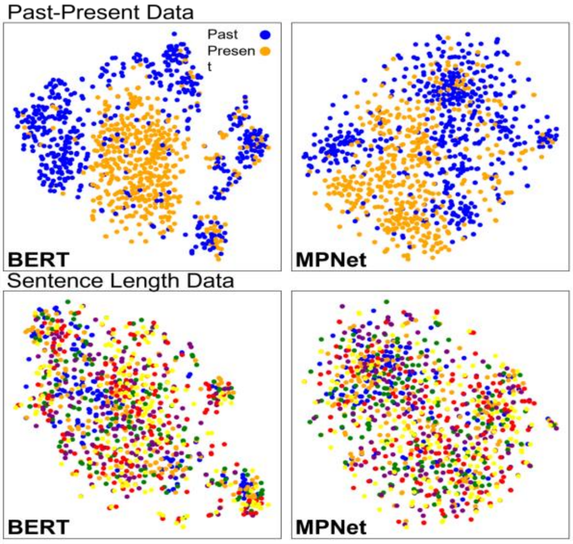


Figure 1: Comparison of t-SNE projections of BERT and MPNet embeddings on SentEval data

Figure 1 displays the comparison of t-SNE projections for BERT and MPNet on two SentEval datasets, Sentence Length and Tense. Already we can make inferences on these models' varying ability to capture linguistic information in their embeddings. Both models seem to cluster past and present classes well in the Tense projection space, however, separation appears cleaner in the projection of the BERT embeddings. This may suggest that BERT sentence embeddings more effectively capture information regarding tense when compared to MPNet embeddings.

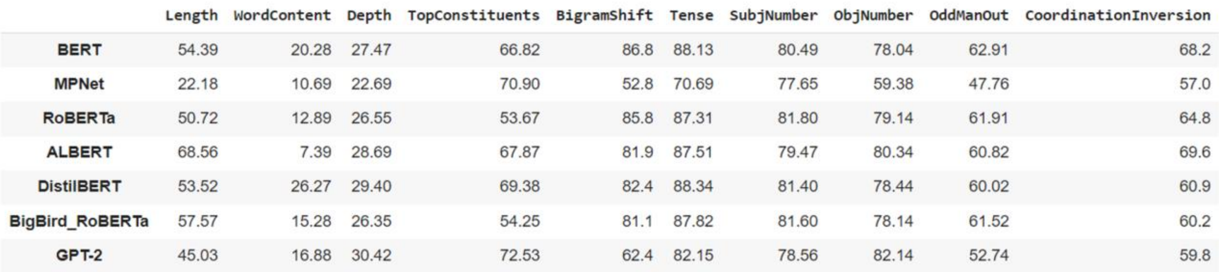
Projections of the embeddings for Sentence Length do not exhibit the same clustering behavior. This may indicate that neither BERT nor MPNet sentence embeddings capture sentence length effectively.

Indeed, after generating the results for our models on the SentEval tasks shown in Figure 2, we can see that both of these inferences were correct. BERT outperforms MPNet on the Tense task (88.1% to 70.7% accuracy) and all models perform more poorly on the Sentence Length task overall (50.3%, Figure 4).

We observe that the BERT family of models all have comparable average performance, an accuracy around 60% on average for the probing tasks, shown in Figure 3. These models share similar architecture, training corpus', and pre-training tasks, specifically masked language modeling. These similarities in training may lead to the similarity in embedding performance that we observe here. The original BERT model performs the best of all models.

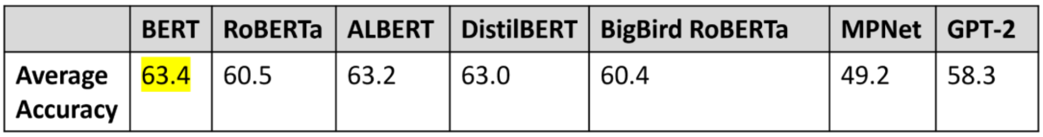
MPNet performed worst on the probing tasks overall, with an average accuracy of 49%. This finding was surprising, as prior work on evaluation of sentence embeddings notes MPNet as one of the best language models available [6]. There are two possible explanations for the discrepancy we observe between our findings and previous work. The referenced work compares models on a set of downstream tasks, which is a fundamentally different approach to embeddings evaluation than the probing tasks we employ (extrinsic vs. intrinsic). Additionally, our results may differ from previous findings due to our use of a smaller training set.

GPT-2 was initially constructed as a language model to predict the next word. And it has proven to perform better in language modeling than the corresponding downstream tasks. Especially, if one does not account for the manifold in the GPT-2 embeddings, the embeddings will not perform well on downstream tasks.



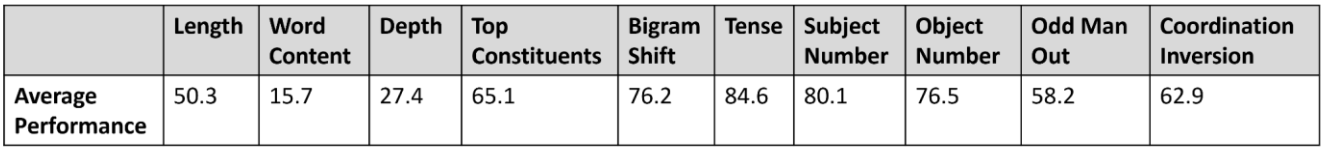
	Length	WordContent	Depth	TopConstituents	BigramShift	Tense	SubjNumber	ObjNumber	OddManOut	CoordinationInversion
BERT	54.39	20.28	27.47	66.82	86.8	88.13	80.49	78.04	62.91	68.2
MPNet	22.18	10.69	22.69	70.90	52.8	70.69	77.65	59.38	47.76	57.0
RoBERTa	50.72	12.89	26.55	53.67	85.8	87.31	81.80	79.14	61.91	64.8
ALBERT	68.56	7.39	28.69	67.87	81.9	87.51	79.47	80.34	60.82	69.6
DistilBERT	53.52	26.27	29.40	69.38	82.4	88.34	81.40	78.44	60.02	60.9
BigBird_RoBERTa	57.57	15.28	26.35	54.25	81.1	87.82	81.60	78.14	61.52	60.2
GPT-2	45.03	16.88	30.42	72.53	62.4	82.15	78.56	82.14	52.74	59.8

Figure 2: Accuracy results for all models and all tasks



	BERT	RoBERTa	ALBERT	DistilBERT	BigBird RoBERTa	MPNet	GPT-2
Average Accuracy	63.4	60.5	63.2	63.0	60.4	49.2	58.3

Figure 3: Average accuracy on all tasks for each model



	Length	Word Content	Depth	Top Constituents	Bigram Shift	Tense	Subject Number	Object Number	Odd Man Out	Coordination Inversion
Average Performance	50.3	15.7	27.4	65.1	76.2	84.6	80.1	76.5	58.2	62.9

Figure 4: Average accuracy of all models for each task

Performance of all models on tasks such as Tense and Subject Number was impressive, above 80% accuracy on average. These specific tasks evaluate the embeddings ability to capture grammatical information from the input sentences. This may suggest that these types of large language models more readily encode this type of grammatical information when compared to other types of information, like word content or sentence length.

Performance on the Word Content task was particularly low, at around 15% accuracy on average. This however, was to be expected: due to computational constraints we were forced to make use of only a subset of the labeled data for each probing task, selecting around 10,000 training examples. The Word Content probing task is structured as a classification task over a vocabulary of 1,000 words, and thus would require a larger training set in order to produce accurate predictors using the embeddings from each model.

6 Conclusions

Our analysis found that the BERT family of models outperformed other PLMs such as MPNet and GPT-2 on the SentEval intrinsic probing of sentence embeddings. The base BERT model performed the best overall. We found that all models more readily encode grammatical information such as tense and subject number in their embedding spaces, than non-grammatical information such as sentence length and vocabulary.

Given additional computational resources and time, a more rigorous testing of the model's embeddings could have been performed. In the future we would be interested in training and testing our models using the full SentEval datasets available and fitting more complex predictors and expressive to our embeddings: neural networks as opposed to the logistic regression used.

References

[1] Ethayarajh, Kawin. "How contextual are contextualized word representations? comparing the geometry of BERT, ELMo, and GPT-2 embeddings." arXiv preprint arXiv:1909.00512(2019).

[2] Conneau, Alexis, and Douwe Kiela. "Senteval: An evaluation toolkit for universal sentence representations." arXiv preprint arXiv:1803.05449 (2018).

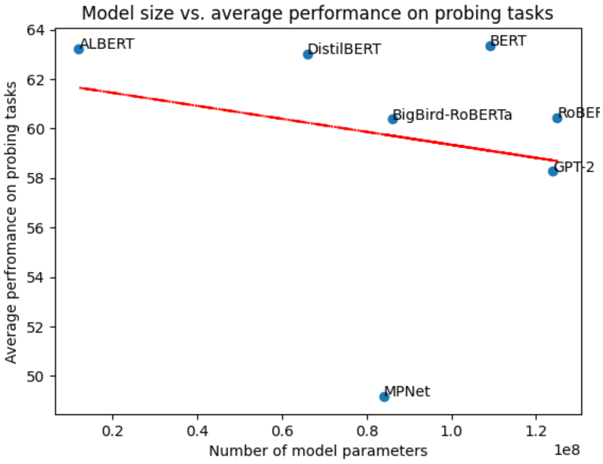
[3] Radford, Alec, et al. "Improving language understanding by generative pre-training." (2018).

[4] Radford, Alec, et al. "Language models are unsupervised multitask learners." OpenAI blog 1.8 (2019): 9.

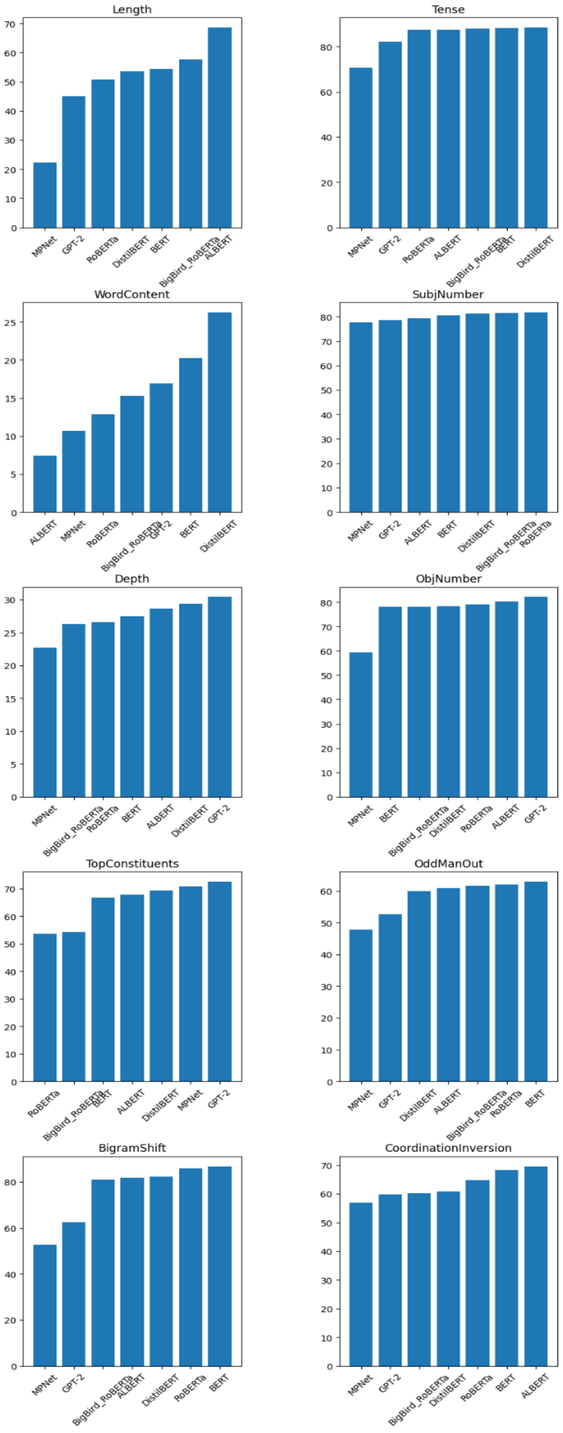
[5] Brown, Tom, et al. "Language models are few-shot learners." Advances in neural information processing systems 33 (2020): 1877-1901.

[6] SBERT.net. (n.d.). *Pretrained models¶*. Pretrained Models - Sentence-Transformers documentation. Retrieved April 16, 2023, from https://www.sbert.net/docs/pretrained_models.html

Appendices



Appendix 1: Model size in number of parameters vs. average performance on SentEval probing tasks



Appendix 2: Accuracy of all models on all probing tasks