Exploring Frame Semantics in Vector Space Models

Group 13

Abstract

Language models that generate vector spaces for word embeddings show exciting promise for semantic tasks, such as predicting analogies. However, less work has been done on how these models might capture other semantic information that is useful for many natural language applications, such as information extraction and machine translation. In this paper, we explore the use of vector space models to capture semantic frames–linguistic semantics related to encyclopedic knowledge–in sentences. We use the state of the art Sentence BERT model to generate sentence embeddings for sentences in the FrameNet dataset that contains over 200,000 hand-labeled sentences with semantic frames, and perform cluster analysis on the vector space to determine how well clustered different semantic frames are. We find that, overall, clusters are not well formed in these vector spaces. We discuss the reasons for these results, and present future work for continuing this study.

1 Introduction

Frame semantics theory enables linguists to relate linguistic semantics with encyclopedic knowledge such that similar situations have the same semantic frames. For example, the sentences "John sold a car to Mary" and "Mary bought a car from John" essentially represent the same situation – a transaction between two people occurring – from different perspectives. These semantic frames are useful for many applications including information extraction and machine translation. As such, initiatives such as FrameNet (ICSI UC Berkeley, 2020) have attempted to build a large semantic frame data set that researchers can use as part of their research, but generating the data set has required manual annotation of over 200,000 sentences.

In this paper, we explore the ability of vector space models of semantics to capture semantic frames of sentences. Vector space models have previously shown to effectively learn the semantic relationship between single words (Mikolov et al., 2013) and between documents of arbitrary length (Le and Mikolov, 2014). We hypothesize that the vector space generated using such models may cluster sentences of the same frame together, such as the examples with John and Mary earlier. If this were accurate, it would allow initiatives such as FrameNet to greatly expand the size of their corpus with potentially related sentences since sentence embeddings models work on unlabeled text data, potentially improving the effectiveness of applications that use semantic frames down the line.

To conduct our study, we use pre-trained, state of the art Sentence Bert models to generate word embeddings for each sentence in the FrameNet corpus (Reimers and Gurevych, 2019, 2020). The FrameNet dataset has labels for words that are lexical units or those that have some semantic frame that other words or concepts relate to. These frames span many categories, including places, actions (such as manufacturing), emotion, and many more. We use these frames to conduct a cluster analysis of the embedding spaces to identify if individual semantic frames are well clustered from one another. We also examine how individual frames different in their clustering between different Sentence BERT models.

Our results show that overall clustering performance is not very good. We hypothesize that this could be due to the high-dimensionality of the sentence BERT models, or because our models were not fine-tuned on the FrameNet dataset we used. We discuss how future work continuing this exploration may want to address the limitations of our work.

Corpus	Document	Sentence	Semantic frames	Words corresponding to the semantic frames	Phrase types corresponding to the words
ANC	110CYL067	I wanted to be there ... I had my second chance to change my life	'Desiring', 'Locative_relation', 'Possession', 'Ordinal_numbers', 'Cause_change', 'Opportunity'	'want', 'there', 'have', 'second', 'change', 'chance'	'verb', 'adverb', 'verb', 'adjective', 'verb', 'noun'

Figure 1: Example sentence from FrameNet annotated with semantic frames.

2 Model

In our analysis, we study how well vector spaces generated by neural language models cluster together sentences containing the same semantic frames. While we initially wanted to use a pretrained doc2vec model for our analysis, we decided against it since doc2vec models can be notoriously poor performing on corpus that they are not trained on. Instead, we use Sentence-Bert (SBERT), which is a version of BERT fine-tuned with a siamese network structure (Reimers and Gurevych, 2019, 2020). We make this choice because (1) SBERT offers major computational advantages over BERT, reducing the sentence similarity task from 65 hours in BERT to 5 seconds with SBERT; (2) it recently, in 2019, achieved state-of-the-art results on several sentence-embedding tasks; and (3) it has several reliable, state-of-the-art pretrained models available, released by the authors themselves.

The pre-trained models released were all trained on the NLI (Natural Language Inference) dataset, which contains 570k human-generated English sentence pairs, manually labeled with one of three categories: entailment, contradiction and neutral. They choose the NLI dataset because previous work has shown that it can be highly effective in generating universal sentence embeddings (Conneau et al., 2018). The specific models they release are described in Table 3, trained on both the base (dim = 768) and large versions (dim = 1024) of BERT that both use mean pooling. We also analyze models that were fine-tuned on the STS Benchmark dataset, which makes them better suited for semantic textual similarity and–we hypothesize–better for clustering semantic frames (Cer et al., 2017).

3 Dataset

The FrameNet dataset is a large-scale effort of manually annotating sentences with semantic frames that represent common semantic situations found in text, such as having possession of an object. We use Release 1.7 of the FrameNet data, maintained by the International Computer Science Institute (ICSI) in affiliation with UC Berkeley (ICSI UC Berkeley, 2020). This dataset contains 7 corpuses and 102 documents, with 4938 total sentences and 797 unique frames.

Figure 1 contains an example row of the parsed dataset. In this example, we see that the sentence is part of the corpus ANC. Each corpus has multiple documents, and this sentence is part of the document 110CYL067. The annotators labeled this sentence with 6 frames, each corresponding to a different word in the sentence.

4 Evaluation

The goal of our evaluation is to see if:

1. vector semantic spaces form good clusters for different semantic frames
2. any patterns exist in which frames cluster together more or less tightly than others
3. there are differences in clusters depending upon which model we use

4.1 Setup

To cluster data, we first generate sentence embeddings for all sentences in our dataset using the SBERT models described earlier. Then, we select data based on different clusters we want to examine (e.g., semantic frame), and label sentences with those facets (e.g., sentences from Desiring frame; sentences from Possession frame; etc.). Finally, we compute our evaluation metrics on each of these groups, which are described in the following section.

4.2 Measures

To evaluate how well corpuses, documents, and semantic frames are clustered together in the vector semantic space, we borrow measures from cluster analysis how well clustered sentences are in the space. We compute a silhouette score for each cluster labeled by a semantic frame (Rousseeuw, 1987). The silhouette score takes into account both how cohesive individual clusters are ($a(i)$) and how separated clusters are from one another ($b(i)$). The measure for a single score is computed as:

$$s(i) = \frac{b(i) - a(i)}{max\{a(i), b(i)\}}, \text{ if } |C_i| > 1$$

$$a(i) = \frac{1}{|C_i| - 1} \sum_{j \in C_i, i \neq j} dist(i, j)$$

$$b(i) = \min_{k \neq i} \frac{1}{C_k} \sum_{j \in C_k,} dist(i,j)$$

Silhouette scores range from -1 to 1, where -1 indicates little separation between clusters and high dissimilarity between points in a cluster, and 1 indicates good separation between cluster and low dissimilarity within the cluster. To compute a score for an entire cluster, the average of scores within that cluster are taken. To compute a score for the entire dataset (i.e., performance across clusters), the max of average scores is taken.

As the silhouette coefficient is a composite measure of clustering performance, we also separately compute additional intra-cluster measures for how tightly sentences within a cluster are together using. First, we compute the average diameter distance (ADD)[1], or the average distance between all objects belonging to the same cluster which is defined as:

$$ADD(C) = \frac{1}{|C| * (|C| - 1)} \sum_{i,j \in C, i \neq j} dist(i,j)$$

Second, we compute the complete diameter distance (CDD), or the maximum distance between two points in a cluster, which is defined as:

$$CDD(C) = \max_{i,j \in C} dist(i,j)$$

These measures give us additional details into how cohesive clusters are for different semantic frames.

For all distance calculations, we use Manhattan distance instead of Euclidean distance due to the dimensionality of the sentence embedding vectors (Aggarwal et al., 2001).

5 Results

Below we present our findings on semantic frame clustering. We present overall clustering performance for each model, and analyze the performance on individual clusters.

5.1 Overall Clustering Performance

Overall silhouette scores on semantic frame clusters show that SBERT models do not cluster semantic frames well; see Figure 2. Both the base and large model reported negative silhouette scores, which indicate that the clusters are likely not very separated. The models fine-tuned on the STS benchmark reported slightly better silhouette scores, but were still negative overall, indicating poor separation ($b(i)$) and cohesion ($a(i)$). We believe this may be occurring due to the high-dimensionality of the SBERT models making the vector space very large. In the next section, we focus our individual cluster analysis on the fine-tuned models since they had slightly better performance.

[1]This measure is equivalent to $a(i)$ in the silhouette coefficient when averaged over all points in a cluster.

Model	Overall Silhouette Score
Base NLI	-0.2594
Large NLI	-0.2534
Base NLI + STS	-0.2131
Large NLI +STS	-0.2044

Figure 2: Average Silhouette scores across all sentences in each model.

Model	Semantic frame	Average diameter distance (ADD)	Complete diameter distance (CDD)	Number of sentences
Base NLI + STS	Measure_volume	100.833171	151.249756	3
	Simultaneity	114.949028	114.949028	2
	Stinginess	222.565430	278.914703	3
	Come_down_with	227.048874	227.048874	2
	Fastener	270.318176	270.318176	2
Large NLI + STS	Measure_volume	148.002533	222.003799	3
	Simultaneity	176.880142	176.880142	2
	Stinginess	275.551554	336.213959	3
	Come_down_with	369.407043	369.407043	2
	Bail_decision	411.489685	411.489685	4

5 smallest semantic frame clusters by average diameter distance

Model	Semantic frame	Average diameter distance (ADD)	Complete diameter distance (CDD)	Number of sentences
Base NLI + STS	Manner	494.585968	494.585968	2
	Expected_location_of_person	477.026794	477.026794	2
	Trying_out	461.139282	461.139282	2
	Timespan	459.536499	486.449677	4
	Medium	458.254272	458.254272	2
Large NLI + STS	Guilt_or_innocence	754.963013	754.963013	2
	Carry_goods	729.914001	729.914001	2
	Spatial_co-location	706.898112	783.406982	3
	Board_vehicle	706.556519	706.556519	2
	Timespan	703.577067	770.423401	4

5 largest semantic frame clusters by average diameter distance

Figure 3: Largest and smallest semantic frame clusters by average diameter distance

5.2 Analyzing Individual Clusters

Beyond overall clustering effectiveness, we evaluate what semantic frames are better or worse clustered by their intra-cluster measures. Figure 3 shows the 5 smallest clusters and 5 largest clusters for the base and large SBERT models trained on the NLI dataset and fine-tuned with the STS benchmark.

Interestingly, both models share 4 out of 5 of the smallest frames, but only 1 of the largest frames (Timespan). However, it is important to also note

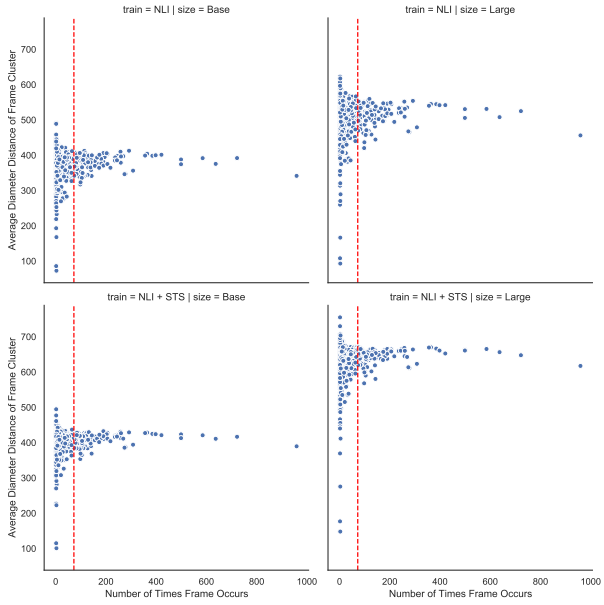

Figure 4: Impact of occurrence count of frames on intra-cluster distance. Red line indicates the mean number of sentences frames occurred in.

the number of sentences that contain those semantic frames. On average, semantic frames were found in 36 sentences, which is much higher than how often any of the frames above were found in sentences. This tells us that the sentences that contain these frames are either (1) very similar sentences with similar embeddings and thus are clustered closely together; or (2) very different sentences with rarely used semantic frames and thus clustered far from one another.

Following our realization that the number of sentences in a semantic frame impacted the clustering results, we explored the extent of this impact. Figure 4, shows the distribution of how intra cluster distance (measured by average diameter distance of frame cluster) varies based on the occurrence count of the frame cluster (number of times frame occurs). We see that, for frames close to the mean occurrence count (the red dotted line), there is very little spread, approximately just 100 units of distance, which suggests to us that all frames might be almost equally well-clustered. However, we do not know how much better a cluster 100 units more tightly clustered is than another, and need better baselines to draw more meaningful conclusions.

6 Limitations and Future Work

A key limitation of our current work is that the SBERT model used was not fine-tuned on the FrameNet dataset, which previous work has shown can yield a large performance benefit if done (Devlin et al., 2019). In future work, we plan to set aside 80% of the data we have (approximately 4000 sentences and their semantic frames) to train, and evaluate on the remaining 20%; we will replicate the analysis above with the remaining 20% so that the results are comparable. We will also try to fine-tune and analyze the performance of an SBERT model designed for a multi-label classification task where we try to predict what semantic frames are present in a given sentence. Another approach to improving performance might be to use models with lower dimensionality to address any potential issues with the curse of dimensionality we might be experiencing here.

Lastly, this work can be extended by looking at co-occurrence and intersection of semantic frames in clusters. Since sentences can have multiple frames in them, it might be valuable to see if co-occurring frames are more closely clustered than individual frames. In future work, we also plan to perform a co-occurrence and intersection analysis to see how similarity between pairs of sentences changes as sentences have more common frames, and to see which combinations (co-occurrences) of shared frames most greatly impacts similarity, and whether any patterns exist.

7 Conclusion

In this paper, we attempt to explore how vector space models can cluster semantic frames together in the embedding space. We perform cluster analysis using FrameNet data with hand-labeled semantic frames and Sentence BERT models to generate sentence embeddings. Though we found that overall clustering performance was not good when using Sentence BERT models, we hope that this line of work can be further refined as detailed above. Intuitively, we still feel that semantic frames must be represented to some capacity in these vector space models. In exploring this further, we hope that mixed-initiative systems can be built to expand the FrameNet labeling initiative.

References

Charu C. Aggarwal, Alexander Hinneburg, and Daniel A. Keim. 2001. On the Surprising Behavior of Distance Metrics in High Dimensional Space. In Gerhard Goos, Juris Hartmanis, Jan van Leeuwen, Jan Van den Bussche, and Victor Vianu, editors, *Database Theory — ICDT 2001*, volume

1973, pages 420–434. Springer Berlin Heidelberg, Berlin, Heidelberg.

Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. 2017. SemEval-2017 Task 1: Semantic Textual Similarity Multilingual and Crosslingual Focused Evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 1–14, Vancouver, Canada. Association for Computational Linguistics.

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loic Barrault, and Antoine Bordes. 2018. Supervised Learning of Universal Sentence Representations from Natural Language Inference Data. *arXiv:1705.02364 [cs]*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

ICSI UC Berkeley. 2020. FrameNet. https://framenet.icsi.berkeley.edu/fndrupal/.

Quoc V. Le and Tomas Mikolov. 2014. Distributed Representations of Sentences and Documents. *arXiv:1405.4053 [cs]*.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. *arXiv:1301.3781 [cs]*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. *arXiv:1908.10084 [cs]*.

Nils Reimers and Iryna Gurevych. 2020. Making Monolingual Sentence Embeddings Multilingual using Knowledge Distillation. *arXiv:2004.09813 [cs]*.

Peter J. Rousseeuw. 1987. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65.