Through Time with BERT

Does pre-trained English BERT-base embed temporal information?

UNIVERSITÄT TRIER

Introduction

The **B**idirectional **E**ncoder **R**epresentations from **T**ransformers (BERT) is a widely used language model to generate contextual embeddings for NLP tasks.
Recent studies address the question what type of linguistic or structural information is represented in BERT embeddings.

In this work, we research the extent to which discrete linguistic tense information presents itself in raw contextual embeddings of the pre-trained BERT-base model. In particular, the following questions are examined:

- Do the embeddings of the different main tenses (Past, Present and Future) contain semantic tense information?
- Do the embeddings store further information about the aspects of each detailed tense, i.e., Simple, Perfect, Continuous and Perfect Continuous?
- How does the representation of the embeddings change when temporal adverbs are added?

Conclusion

Our experiment shows that BERT embeddings not only contain information about the main tenses, but also to some extent about the detailed tenses. We conclude that digits, adverbial structures, as well as syntactic similarity, length, and the amount of morphological information in a sentence all have an impact on BERT's sentence representations.

References

A comprehensive list of the references is available upon request.
Please scan the QR code for more information about BERT.

Discussion & Results

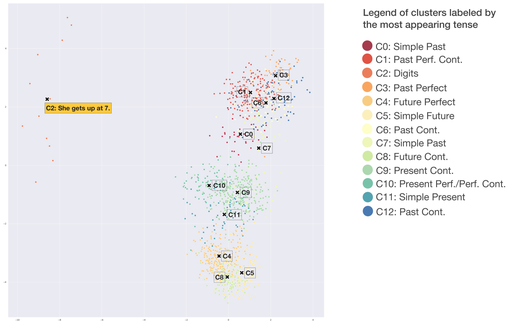
General observations

The clustering has a higher accuracy in distinguishing three clusters with and without adverbs (99.71%) than thirteen clusters without adverbs (86.9%).

Digits

Digits are noise in regards to our research, because they are mostly put into one or only very few different clusters and confuse the correct clustering by our defined tenses.

Figure 1: Principal Component Analysis with 13 clusters w/o adverbs



Morphology and Morpho-Syntax

The distances between the clusters in main as in detailed tenses often depend on shared morpho-syntactical forms.
More linguistic similarity results in less distance between the clusters of different tense forms. Therefore, the uniqueness and the number of morphological cues appears to be crucial.

Adverbs

Since some of the adverbs were used across different tenses, one can assume that these syntactic resemblances have an influence on the embeddings making them more similar and therefore reducing the distances between the referring clusters.

BERT Pre-Training

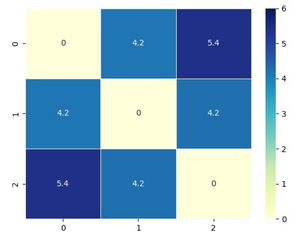
Although pre-training tasks do not explicitly consider tense, it seems that temporal knowledge is learned implicitly through both Next Sentence Prediction and Masked Language Modeling.

Table 1: Faiss k-Means clustering for Past, Present and Future sentences w/o adverbs using three distinct clusters

Faiss k-Means Clustering
3 Clusters without Adverbs

	C0	C1	C2	Total
Past				
Past Continuous Tense	0	0	114	114
Past Perfect Continuous Tense	0	0	114	114
Past Perfect Tense	0	0	114	114
Simple Past Tense	3	0	111	114
All	3	0	453	456
Present				
Present Continuous Tense	114	0	0	114
Present Perfect Continuous Tense	114	0	0	114
Present Perfect Tense	114	0	0	114
Simple Present Tense	113	1	0	114
All	455	1	0	456
Future				
Future Continuous Tense	0	114	0	114
Future Perfect Continuous Tense	0	114	0	114
Future Perfect Tense	0	114	0	114
Simple Future Tense	0	114	0	114
All	0	456	0	456
Total	458	457	453	1,368

Figure 2: Euclidean distances for 3 clusters with adverbs (C0: Future, C1: Present, C2: Past)



Methodology

I Dataset

114 simple English sentences were created. Then, all sentences were phrased in each of the 12 tenses. This dataset formed the basis for two corpora: one with and one without adverbs.

How

1. Corpus No. 1: GPT-3 via one-shot learning, manually filtered
2. Corpus No. 2: Added matching temporal adverbs to generated sentences using a self-written Python script

II Embeddings

A 768-dimensional vector representation of each sentence was created.

How

1. Generated via BERT-base uncased from bert-as-service
2. Extracted [CLS] token as sentence representation from second-to-last layer

III Clustering

Four clusterings were performed: 3 clusters with adverbs, 3 clusters w/o adverbs, 13 clusters with adverbs, 13 clusters w/o adverbs.

How

1. Applied k-Means from Facebook AI Similarity Search (FAISS)
2. Applied fuzzy clustering via Gaussian Mixture Model (GMM) for verification of clustering

IV Dimension Reduction

The 768-dimensional vectors were reduced to two dimensions for plotting with Matplotlib and Seaborn.

How

1. Used Principal Component Analysis (PCA) from FAISS
2. Used a scatter plot for visualizing the clustering

A project within the master program Computational Linguistics in the winter term 2021/22.
Authors: Laura Bernardy, Nora Dirlam, Maya Gálvez Wimmelmann, Johanna Garthe, Vivien Meyer, Ann-Jasmin Schneider, Viktoria Weber
Graphic: SeekPNG (2019): Bert – Bert and Ernie. Retrieved from SeekPNG.