ReligioBERT - MSiA 414 Text Analytics Final Project

Nathan Franklin
Northwestern University
NathanFranklin2019@u.northwestern.edu
4 December 19

Abstract

"Life is suffering" - The First Noble Truth of Buddhism. This project is an attempt to use machine learning and natural language processing to deal with this inconvenient truth by continuing the pre-training step of a BERT language embedding model on a religious text corpus, resulting in the ReligioBERT model. After this pre-training step, I created a REST api deployment of both the base-BERT model and the ReligioBERT model which takes user input and responds with a passage from a religious text, in this case The Confucian Analects. I then compared various performance measures of the baseBERT model and the ReligioBERT model.

1 Introduction

As unenlightened humans, we are faced with a barrage of environmental stimuli, our brain's interpretations of those stimuli, challenges in life, important dilemmas, and many more things which buffet us in our daily lives. Over the millennia, human civilization has created religion and philosophy to try to best navigate through the precarious circumstances we find ourselves in. Religious and philosophical texts are the storage and delivery vehicles of humanity's most well known answers and strategies, and thus having an easy way to find a relevant passage from one of these texts could come in handy for humans in their day to day lives. This project attempts to use techniques in machine learning and natural language processing to develop a model which can make it easier for someone to find the right topic or passage for their situation. The related work section details previous relevant work done in the natural language processing field of machine learning, while the method section details how I went about training and deploying my model for the aforementioned goal.

2 Related Work

This project's main technical approach is focused on extracting vectorized representations of words and/or sentences. With this in mind, I did some research into different ways to accomplish this task.

One of the most popular approaches to this task in the NLP community was using either a pre-trained word2vec model or a custom trained word2vec model, as [1] Mikolov et al. (2013) explains. The approach taken by these researchers was to build a neural network, which after training, would be able to produce unique vectors for any word used in its training vocabulary. These embeddings are word-by-word, and cannot be extended directly to sentence embeddings, one drawback of this method. The embeddings generated

also are not context-dependent, meaning the word "cell" (or any other homonym groups) will get only one vector representation, where optimally we could get different vector representations for its different uses in everyday language (context-dependent). Word2vec took a creative approach to generate these embeddings, using 2 different model architecture and vector extraction strategies. The CBOW model took as input words which appeared near a masked word, and tried to predict this masked word. The skip-gram model took as input one word, and tried to predict the context words around it.

An improvement in some ways over word2vec came from [2] Peters et al. (2018) with the introduction of ELMo. As previously noted, one of the weaknesses of word2vec was the context-independent nature of the embeddings. ELMo uses a bi-directional LSTM structure in order to be able to take into account the context of a word being used in a sentence. Similar to how a deep learning network uses earlier layers to recognize edges and simpler structures, and later layers to recognize more complex structures, the ELMo architecture can be thought of as using its earlier layers to get the context of a word in a sentence, and then the later layers to get the specific definition for that word in a given context.

While ELMo was an improvement over word2vec, BERT, layed out in [3] Devlin et al. (2018), is the latest improvement to this strategy. The context-dependent embeddings are retained, while the complex, bi-directional-lstm architecture is eschewed in favor of a simpler Transformer approach using language attention. Because of these improvements, I chose BERT as my model to accomplish my goal.

3 Dataset

I used a corpus of 10 religious texts from world religions for pre-training of ReligioBERT. See Appendix [5]-[13] for a list of the texts. The texts were cleaned and stripped of extraneous information including stop words, punctuation, capitalization, metadata, and more before training of ReligioBERT. The details are specified in the associated code repository (link found at the end of paper).

4 Method

The end vision of the ReligioBERT model was to have a way for a user to be able to enter in some topic, problem, question, or situation in their lives, and to have the application provide suggestions for which passage from the available religious/philosophical literature would be most relevant to a user's input.

While baseBERT can provide us high dimensional vector representations of words and sentences which we can use for this task, the model was not trained on religious texts, and as such, it may not have good representations for much of the topics/syntax/vocabulary present in this corpus. Inspired by the idea of training a BERT architecture model on a specific knowledge domain corpus in order to achieve better understanding as displayed in [4] Beltagy et al (2019) with their deployment of SciBERT, I followed a similar approach. I continued the pre-training of a baseBERT model on my own religious text corpus in order to train and deploy ReligioBERT. I further pre-trained baseBERT for three epochs, which took about two hours on a single Nvidia GeForce RTX 2080 Ti GPU, to create ReligioBERT.

After training, now in possession of a baseBERT model as well as a ReligioBERT model, I created 2 versions of the Confucian

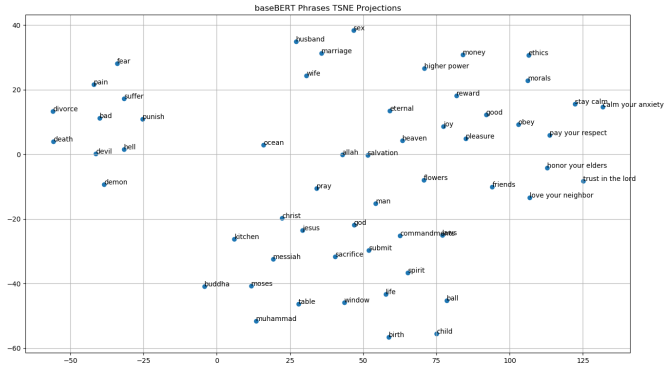

Figure 1: TSNE Projection from baseBERT Embeddings.

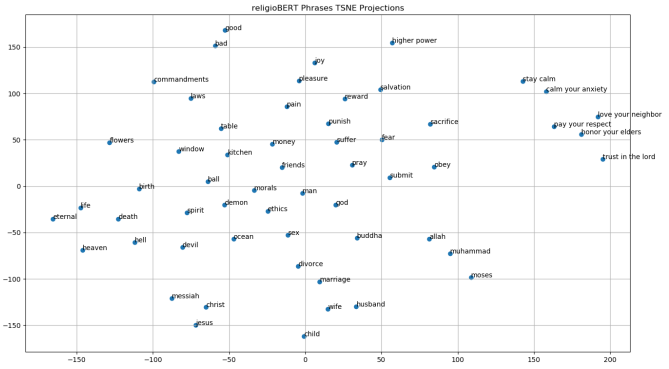

Figure 2: TSNE Projection from ReligioBERT Embeddings.

Analects text, one with an associated baseBERT embedding and one with an associated ReligioBERT embedding for every sentence. My intuition was that these BERT embeddings can capture meaning/syntax/topics and more from a sentence, and thus I could approximate how ”applicable” a sentence from a religious text was to user input by comparing the cosine distance between the 768-dimensional (base/Religio)BERT embedding of the user input with the 768-dimensional (base/Religio)BERT embedding of the sentences in the Confucian Analects document. The sentence from Analects which had the smallest cosine distance was taken to be the most ”relevant” sentence, and I had my model return both 2 sentences before and after this most relevant sentence in order to give a passage back. Ideally this input would be directly relevant to a user and thus will be worth reading for that person.

I performed experiments on my baseBERT and ReligioBERT models where I created embeddings for a group of phrases, and used TSNE projections to map them on to 2D space, where I could analyze how well the

models were grouping words. I also performed a comparison of the passages given back by the model for various user input examples for the two models. The results are detailed below.

5 Results

My experiments compared the performance of the uncased base BERT pre-trained model (12-layer, 768-hidden, 12-heads, 110M parameters) with my custom ReligioBERT model. Below is a comparison of these models fared in two tasks: TSNE Projections (Figures 1 and 2) and Religious Text Passage Recommendations (Figures 3, 4, 5, and 6).

5.1 TSNE Projections

The first task involved analyzing TSNE projections for the two models. BaseBERT TSNE projections (Figure 1) actually separated things pretty intuitively, even with these specialized vocabulary words. For instance, negative connotation things like divorce/death/bad/suffer/pain/demon are grouped together, separate from others. Then more spiritual things were grouped together decently, with messiah/moses/jesus/pray/salvation and more all being somewhat close to each other. Interestingly the longer phrases were all grouped together.

ReligioBERT TSNE projections (Figure 2) were fairly similar in its projection, and maybe even having a bit less intuitive projection groupings. Of course it's important to remember TSNE is a stochastic process and so there is some element of luck/randomness in creating and interpreting these. In fact the negative things, like punish/suffer/fear/death/divorce were much more spread out in the ReligioBERT projection. One hypothesis for this is that in normal text, these bad things may be less frequent, and more clustered together, but in religious texts, because these themes are more common, it may be that they are less strongly associated with each other.

5.2 Religious Passage Recommendations

Figure 3 shows the recommended passage from baseBERT based on the user-input: "my parent recently died and I am very sad and lonely". BaseBERT replied with a passage about "superiors" and "departure", but it wasn't particularly relevant. ReligioBERT (Figure 4) responded with a passage about "master" but it also was not particularly helpful.

The second user-input ("treat other people how you wish to be treated, with respect") gave more interesting results. BaseBERT (Figure 5) sent back a passage about a superior man being respectful to all others. ReligioBERT (Figure 6) seemed to give a relevant passage, talking about advice when meeting new people, to always behave as if you are receiving many guests. It also mentions to "not to do to others as you would not..." which is directly relevant to the topic the user input.

6 Discussion

It's unclear if the ReligioBERT provided any substantial improvement over the baseBERT model on these tasks. More numerous and rigorous tests would need to be undertaken in order to draw a strong conclusion. The specific way the BERT embeddings were used also may not be ideal for my task, as just because analects has a sentence with a close BERT embedding to a user input, it may not necessarily be helpful or relevant to the user. A more complete understanding of BERT embeddings and the distance between them could help with this task. The size of the BERT model is also a drawback, as it is relatively time consuming to perform the training, and even in some cases

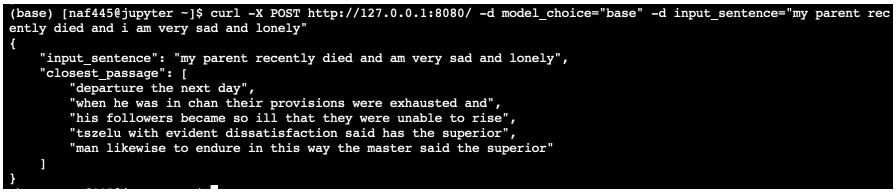

Figure 3: BaseBERT Passage Recommendation

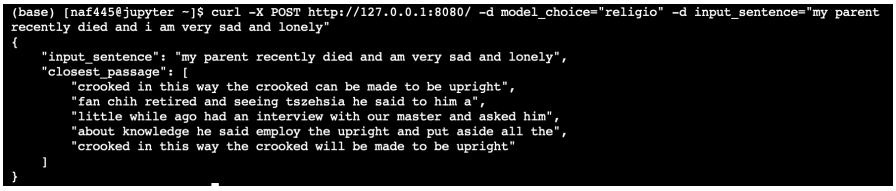

Figure 4: ReligioBERT Passage Recommendation

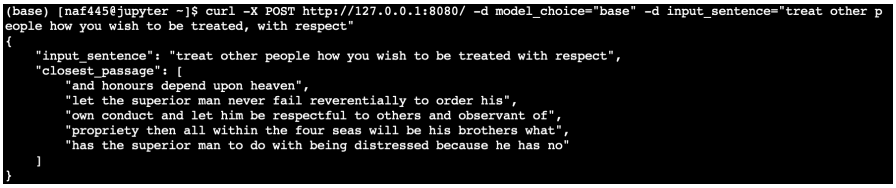

Figure 5: BaseBERT Passage Recommendation

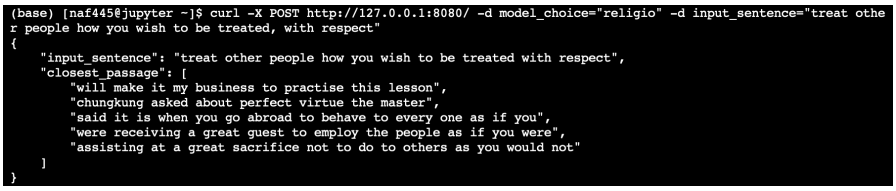

Figure 6: ReligioBERT Passage Recommendation

to get an embedding on a new user input. The size also makes it difficult to host on the available cloud based deployment services. In the future, I think trying longer training and more rigorous evaluation methods would help to see whether this approach is worthwhile. I guess for now, you will still have to go talk to your local religious scholar...

Acknowledgments

Special thanks to Emilia Apostolova and Timo Wang for direction and help!

Appendices

6.1 References

References

[1] Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013b). Distributed representations of words and phrases and their compositionality. In Advances in Neural Information Processing Systems, pages 3111–3119.

[2] Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018b. Deep

Contextualized Word Representations. arXiv preprint arXiv:1802.05365.

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of Deep Bidirec- tional Transformers for Language Understanding. arXiv preprint arXiv:1810.04805, 2018.

[4] Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. Scibert: Pretrained Language Model for Scientific Text. In EMNLP.

[5] Christian/Jewish: King James Bible

`https://www.kaggle.com/tentotheminus9/religious-and-philosophical-texts`

[6] Hindu: Bhagavad Gita

`https://www.gutenberg.org/files/2388/2388-h/2388-h.htm`

[7] Islamic: The Quran

`https://www.kaggle.com/tentotheminus9/religious-and-philosophical-texts`

[8] Buddhist: The Gospel of Buddha

`https://www.kaggle.com/tentotheminus9/religious-and-philosophical-texts`

[9] Confucianism: Analects

`http://www.gutenberg.org/cache/epub/3330/pg3330.txt`

[10] Egyptian: Egyptian Book of the Dead

`https://www.gutenberg.org/files/28282/28282-8.txt`

[11] Mormon: Book of Mormon

`https://www.kaggle.com/tentotheminus9/religious-and-philosophical-texts`

[12] Tibetan: The Tibetan Book of the Dead

`https://archive.org/stream/TheTibetanBookOfTheDead/The-Tibetan-Book-of-the-Dead_djvu.txt`

[13] Stoic: Meditations

`https://www.kaggle.com/tentotheminus9/religious-and-philosophical-texts`

6.2 Code Repository

`https://github.com/naf445/naf445_msia414_project`