**Finding Topical Similarity in Responsa Using Transformers**

*Jacob Khalili and Yehuda Snow*

*Advised under Prof. Jonathan Schler*

One of the cornerstones of modern halachic literature is responsa. Careful learning, analysis and understandings of these scholarly letters inform not only the recipient on how they are to act, but establish important halachic precedence and principles used to decide complex questions for future generations.

Imagine a world, where when you are learning helpful suggestions of sources to look at can help aid in your understanding; a world where it's simple and easy to find the opinions of other great poseqim (adjudicators of Jewish law) on similar issues and other cases where similar concepts are at play.

Using modern natural language processing (NLP) techniques, building such a tool on a large scale has only recently become possible. In 2018, Google Research published a state of the art pre-trained network for NLP, called Bidirectional Encoder Representations from Transformers, or BERT. BERT is an unsupervised machine learning model, which is used to create an encoding-vector which represents the texts. BERT’s bidirectionality is one of the key reasons for its success. When representing a word, BERT not only looks at the word itself, but the context of the word, in both directions (previous and future words).

In April of 2021, researchers at Bar Ilan University published AlephBERT, a model based on Google’s BERT, but trained on modern Hebrew literature. The model was trained on Twitter, Wikipedia, and OSCAR (a large dataset of multilingual data crawled from the internet.) The model was trained using Masked-Language Modeling (MLM), where every sentence fed to the model contained a blank word, which the model tried to predict; the weights of the model were optimized to predict the missing words correctly. This gave the model a sense of linguistic patterns of the language it was trained on.

Whether doing training or inference, the model requires a tokenizer to convert the words into a vector representation of the text. This vector representation is then fed into BERT to generate an encoding. When training, that encoding is then fed to a second half of the network, which predicts the missing word, followed by a loss calculation, and the updating of weights. When doing inference, the distance between the current paragraph’s vector representations, and all other vector representations is calculated, and the closest paragraphs are represented to the user. After being trained in this fashion, the middle layer’s output is used as a vector encoding of the input text.

In April of 2021, the Bar Ilan NLP Lab released AlephBERT, a BERT model designed for the Hebrew Language. It was trained on OSCAR’s Hebrew section (a collection of articles in Hebrew from the internet), Hebrew Wikipedia, and Hebrew twitter. Because of the training data provided, the model learned very modern colloquial hebrew. However, for use in this task AlephBert would need to learn a more formal, and much older hebrew used in rabbinic responsa.

This was done by fine-tuned AlephBERT via MLM on a subset of the Bar Ilan Responsa Project’s collection of rabanic responsa. Through this process, AlephBERT became accustomed to the dialect of Hebrew used specifically in responsa.

The model was then able to generate embedding in vector space that accurately represented the topics of each document. Calculating the magnitude of the distance between two vectors to represent how topically different they are. For computational efficiency, this process was implemented using matrix operations, which were performed on a GPU. Now an efficient method for comparing a paragraph to all the other paragraphs in courpas of responsa and finding the paragraphs that is the most topically similar to that paragraph was possible.

The results received were reasonable with this method, though the measurements are not objective. For any given paragraph about 2 of the top 3 suggestions were deemed relevant.

Generally in rabbinic responsa, the main points in the document are expressed in the beginning when the question is asked and at the end in the conclusion. So, in order to improve the search results document embeddings were also generated. This was done by using the first 250 tokens and the last 250 tokens and in the document. This size was chosen because the maximum pre-trained AlephBERT model uses is 512 tokens.

This method on its own generated reasonable results which were similar in quality to the previous model. After summing both scores for any given paragraph (the sum of the paragraph distance and the document distance), when then sorted by overall distance, the results were promising, though similar. There is much difficulty in assessing model improvement unless there are very large changes, noticeable in reviewing a few dozen documents. In the future, once users are available, user feedback can be a critical tool in assessing how relevant a particular article is.

Overall, while the results for topical similarity were promising, there is still more work to be done. Sometimes results contained no paragraphs that were similar or just one or two similar results, even though there were actually many paragraphs that were similar. This could potentially be improved by increasing the number of training epochs, using an ensemble of models, and/or implementing a discriminator network.