SemSe - A Multilingual Semantic Search Application

Akshay Shinde, Aratrika Basu, Nitish Surana, Omkar Sabnis, Sai Srihitha Reddy Thummala
University of Southern California
[akshinde, aratrika, nitishsa, osabnis, saisrihi]@usc.edu

1 Motivation

Semantic search describes a search engine's attempt to generate the most accurate results possible by understanding the query based on the searcher's intent, query context and relationship between words. Most search engines and techniques emphasize keyword matching and optimizing exact pattern searching. At times, keyword search does not help the end-user as the document may not contain the exact search phrase but may contain its semantic equivalent.

Our project focuses on enabling multilingual semantic search for documents and websites where normal browser search features fail to identify results using traditional keyword matching.

1.1 Novel Contributions

We focus on the multi-lingual aspect of the document as part of our novel contributions. A transliterated search query would enable the user to quickly get search results without worrying about the language barrier or having knowledge of the vocabulary used.

We particularly target popular languages that the students at USC are familiar with like English, Hindi and Spanish. We also want to make an API that will not require any work from the user, thereby being almost unsupervised in nature.

2 Design

2.1 Materials

Our API will be fine-tuned on data that is provided by the website. We will use a standard pre-trained model such as BERT or mBert (Xu et.al, 2021) which will provide us with a wide representation of the languages chosen. Our API will have multiple endpoints for parsing, preprocessing and training making the process unsupervised - not requiring any user intervention (Wang et.al, 2021).

2.2 Methods

The semantic search API will have two phases - a training phase and a querying phase. In the training phase, our API will parse the website, extracting the content from the different pages of the website. The content will be preprocessed and converted into the correct format as required by our model. We then fine-tune the model on the extracted data which will make the general model more effective. Using the fine-tuned model, a semantic index, a unique representation of the website content is created and stored.

In the querying phase, the API will accept multilingual user queries. The query will run through the model encoder - converting the query into the index representation. The query embedding will be compared with the semantic index to find the content that is most similar to the query. We return the top 5 most similar embeddings. The decoded content is then returned to the website, which the website will display as hyperlinks for the users to click on.

We plan on creating a small custom corpus of multilingual queries that the user can ask the semantic search system. We will be gathering multilingual search queries through surveys. We will then annotate these search queries and feed them to our model for finetuning and testing purposes. We will start by fine-tuning our model in English, followed by Hindi and Spanish in the end.

2.3 Baselines and Evaluation Protocols

We test the model's performance on standard datasets like SentEval (Conneau and Kiela, 2018), Microsoft Research Paraphrase Corpus (MRPC) (Dolan and Brockett, 2005), Quora Question Pairs dataset (Sharma et.al, 2019), etc. that are widely used to compare semantic textual similarity. We will compare the model performance on these datasets with BERT's performance. This will be our baseline for evaluation.

Evaluation Protocols: We will be using an 80:10:10 split for our training, validation & test data. We will be using accuracy and F1 score as the metric for evaluation to match the standard metrics used on these datasets. We will also evaluate the top-5 and top-1 result performance of our model on the custom corpus and compare it to the performance of BERT.

3 Timeframe and Division of Work

We plan on providing a fine-tuned semantic search model by the end of the project. The potential difficulties that can arise while working on these projects include making the corpus diverse, annotation procedure, and the size of the corpus. Our alternative plan would be to create a bilingual model (considering only 2 languages) in case of time restrictions.

All team members will work in every phase of the project. Our work is divided into three phases.

1. Akshay Shinde will lead the first phase which involves data gathering & corpus creation. It will also include scraping website content, surveying for the test corpus, data preprocessing and corpus creation. (Till April 1)
2. Aratrika Basu and Omkar Sabnis will lead the second phase which consists of model building, fine-tuning and API creation. (Till April 10)
3. Nitish Surana and Sai Srihitha Reddy Thummala will lead the third phase which involves evaluating & comparing the model's performance with the baseline and visualizations of the results. (Till April 20)

References

Haoran Xu, Benjamin Van Durme and Kenton Murray, 2021. BERT, mBERT or Bi-BERT? A study on contextualized embeddings for Neural Machine Translation. In *The 2021 Conference on Emperical Methods in Natural Language Processing.*

Kexin Wang, Nils Reimers and Iryna Gurevych, 2021. TSDAE: Using Transformer-based Sequential Denoising Auto-Encoder for Unsupervised Sentence Embedding Learning. In *Findings of the 2021 Conference on Emperical Methods in Natural Language Processing.*

Alexis Conneau and Douwe Kiela. 2018. SentEval: An evaluation toolkit for universal sentence representations. arXiv preprint arXiv:1803.05449.

William B. Dolan and Chris Brockett. 2005. Automatically Constructing a Corpus of Sentential Paraphrases. I*n Proceedings of the Third International Workshop on Paraphrasing (IWP2005)*.

Lakshay Sharma, Laura Graesser, Nikita Nangia, Utku Evci. 2019. Natural language understanding with the quora question pairs dataset. arXiv preprint arXiv:1907.01041.

Pratik Bhavsar, 2021. Better Semantic Search with Unsupervised Training of Sentence Encoder. https://pakodas.substack.com/p/unsupervised-training-of-sentence

4 Appendix

The following is a block diagram of our approach: