Information Retrieval and Extraction on COVID-19 Clinical Articles Using Graph Community Detection and Bio-BERT Embeddings

Debasmita Das, Yatin Katyal, Janu Verma Shashank Dubey, Aakash Deep Singh,
Kushagra Agarwal, Sourojit Bhaduri, Rajesh Kumar Ranjan
Mastercard AI Garage, Gurgaon, India
{firstname.secondname}@mastercard.com

Abstract

In this paper, we present an information retrieval system on a corpus of scientific articles related to COVID-19. We build a similarity network on the articles where similarity is determined via shared citations and biological domain-specific sentence embeddings. Ego-splitting community detection on the article network is employed to cluster the articles and then the queries are matched with the clusters. Extractive summarization using BERT and PageRank methods is used to provide responses to the query. We also provide a Question-Answer bot on a small set of intents to demonstrate the efficacy of our model for an information extraction module.

1 Methodology

We briefly describe our method here.

1. **Network of the articles:** We build a *citation graph* of the articles in the corpus where nodes corresponds to the papers and the edges are determined by the citations of the papers. If two papers have a common citation, then we add an edge between them. In addition, we include edges between papers if their semantic similarity is greater than a threshold. W use *BioSentVec* (Chen et al., 2019) which is trained on a corpus of about 30 million clinical and bio-medical research articles from the public databases - PubMed and MIMIC-III. *BioSentVec* provides *700-dimensional* sentence embeddings.

2. **Clustering of the articles:** We use **ego-splitting** (Epasto et al., 2017) based *community detection* algorithm to partition the articles in the network into clusters. The clusters are then studied qualitatively and we manually assign appropriate labels to the clusters.

3. **Mapping queries to the clusters:** A a given query, we find clusters that are closely related to the query. This mapping is facilitated by BioBERT embeddings (Lee et al., 2020) of the queries and the article titles. This helps in reducing the search space of the query to only within the most relevant clusters.

4. **Information Retrieval:** Focusing on the articles in the clusters relevant to the query, we use BioBERT embeddings of the whole articles (title, abstract, and body) to find the best matched articles within the clusters. Using a PageRank based procedure, we also extract best matching sentences to the query from within the most similar documents.

References

Qingyu Chen, Yifan Peng, and Zhiyong Lu. 2019. Biosentvec: creating sentence embeddings for biomedical texts. In *2019 IEEE International Conference on Healthcare Informatics (ICHI)*, pages 1–5. IEEE.

Alessandro Epasto, Silvio Lattanzi, and Renato Paes Leme. 2017. Ego-splitting framework: From non-overlapping to overlapping clusters. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 145–154.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.