CS410 - Text Information Systems
Satyo Iswara (iswara2)

Technology Review - State of the Art of Medical Text Analysis

Introduction

In the past couple decades, digitalization of health data has been steadily increasing. In 2020 alone, it is estimated that globally 2,314 exabytes of health care data is generated [1]. This data is generated from multiple sources such as EHR(electronic health records), physician/ nurse notes, clinical research, radiology and ECG(Electrocardiogram) report just to name a few. With such a wealth of data additional knowledge can be gained from cross referencing data, however with such a extensive data, gathering association can be a challenge on its own. So with this challenges in mind, text analysis technique has emerge to help health care professional to improve patients outcome. In this paper, I want to look into text analysis techniques used for medical text including named entity recognition (NER), relationship identification and text summarization. In this paper, I also want to explore the commercial used of medical text data.

Analysis Methods

The first methods that I want to explorer is named entity recognition (NER). NER is the technique to identify a substring as a part of predefined category. Example of NER would be identification of the substring 'multiple sclerosis' and putting it as a disease category. Zweigenbaum et al. [2] indicated that Conditional Random Fields (CRFs) has historically been used to identify entity, however performance of NER is more heavily dependent on the text features. Jinhyuk Lee et al. [3] presented BioBERT (Bidirectional Encoder Representations from Transformers for Biomedical Text Mining) as a NER tool pre-trained on large scale biomedical corpora. BioBERT trained using domain specific data sets PubMed abstracts (PubMed) and PubMed Central full-text articles (PMC). For tokenization BioBERT uses WordPiece tokenization, which breaks down term into frequent sub term. This mitigates issue on out-of -vocabulary issue.

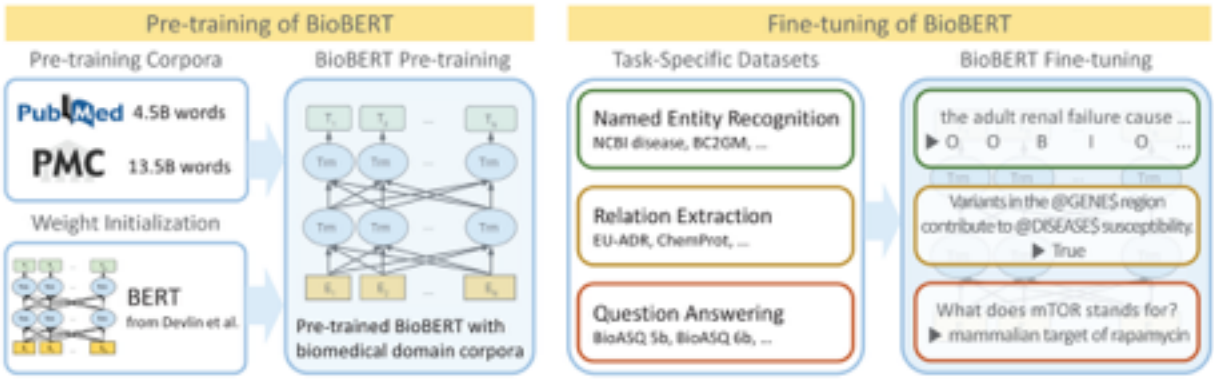

Figure 1. BioBERT overview

After NER, the second important method for medical text analysis is to identify the relationship between entities. Identifying relationship can give healthcare professional additional insight and knowledge from source text. The simplest approach is to see co-occurance of entities, however this is imperfect as co-occurance statistic can have a high recall, but have poor precision. So, pattern-based approach is more typically used to identify relationship between entities. Bunescu et al. [4] used weight identification for patterns, term and part of speech to derive relationship of entities from individual sentences. The confidence on the document is computed as maximum confidence over all sentences in the document. Another approach is to do a full parsing to generate a complete syntactic structure and derive relationship based on grammatical relationship (eg. subject to object relationship). This technique however is bounded by the size of corpus to be analyzed.

The third important analysis method is to summarize medical text. This is a important task as it not only help medical professional to quickly identified related information based on a given query. We should make a distinction how summarization generated from text analysis is differ from generic summary as generic summary does not make any assumption of the intended use of the documents, while summary generated by text analysis system takes into account users information needs. Muhammad Afzal et al.[5] provided a framework termed Biomed-Summarizer. Biomed-Summarizer use several steps to create summary. First, it utilize deep neural network binary classifier to filter scientific studies. Second, Biomed-Summarizer use a bidirectional long-short term memory recurrent neural network as classifier, this generates PICO(Patient/Problem, Intervention, Comparison, and Outcome) text sequences. Then in the third step Biomed-Summarizer computes similarity between user query and PICO test sequences using Jaccard similarity. And the last step is to generate representative summary from high scoring PICO text sequence.

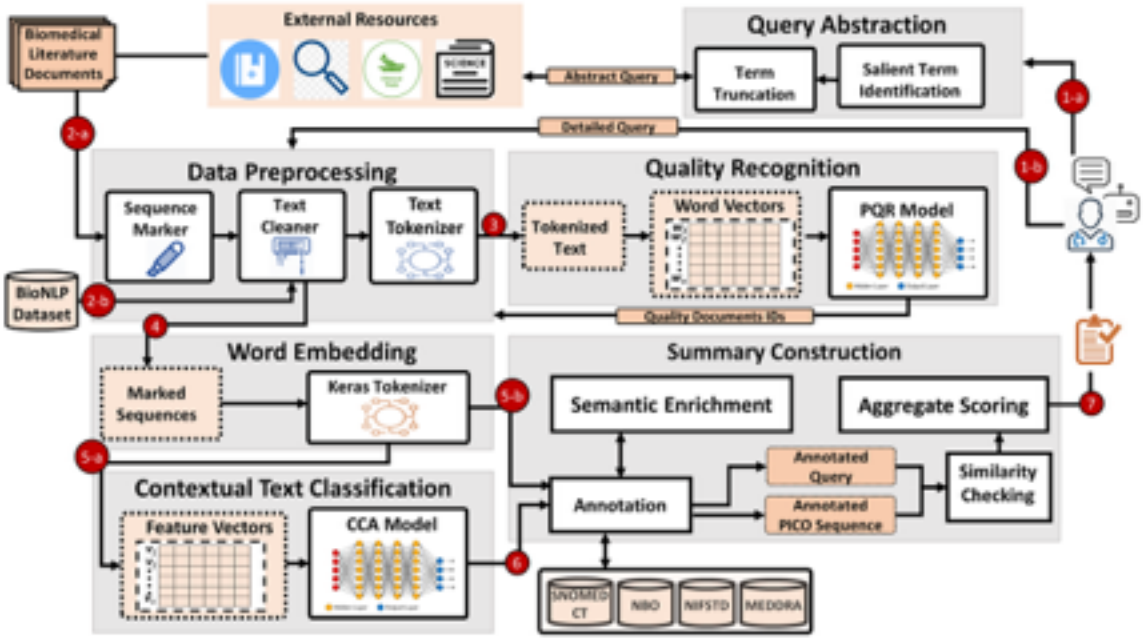

Figure 2. Biomed-Summarizer architecture

Conclusion

With the rise of digitalization of medical records, the needs of medical text analysis will continue to increase. In my research, there are several notable company that include medical text analysis as their product offering. Quertle is a biomedical search engine based of Henderson, Nevada. Quartle main product Qinsight provided medical professional biomedical literature search engine that also provide recommendation based on their previous research. Qinsight corpus contains more than 44 million documents, including biomedical literature, journal articles, grant proposals, clinical trials and patent applications. Another company that utilizing medical text analysis is MedBioInformatics. MedBioInformatics is a company that focus on genomics analysis based of Barcelona, Spain. Its main offering, Disgenet plus is based on large volume of genome sequence and large volume of publication on disease genomics. Disgenet plus provided medical professional and drug makers to easily identified related published work. Quartle and MedBioInformatics are just some of many companies that using medical text analysis. There is a lot more other challenges on medical text analysis that is not covered in this paper, but looking into all of this, I am excited for the development medical text analysis in the future.

References

[1] Manfye Goh, "Medical Text Analytic Techniques And Its Applications", towards data science, November 14, 2020

[2] Pierre Zweigenbaum, Dina Demner-Fushman, Hong Yu and Kevin B. Cohen, "Frontiers of biomedical text mining: current progress", August 15, 2007

[3] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim,Sunkyu Kim, Chan Ho So and Jaewoo Kang, "BioBERT: a pre-trained biomedical language representation model for biomedical text mining", September 5, 2019

[4] Bunescu R, Mooney R, Ramani A, et al. "Integrating co-occurrence statistics with information extraction for robust retrieval of protein interactions", MEDLINE, 2006

[5] Muhammad Afzal, Fakhare Alam, Khalid Mahmood Malik, Ghaus M Malik. "Clinical Context–Aware Biomedical Text Summarization Using Deep Neural Network: Model Development and Validation", Journal of Medical Internet Research, October 23, 2020