CBNU at TREC 2019 Precision Medicine Track

Seung-Hyeon Jo, Kyung-Soon Lee

Division of Computer Science and Engineering, RCAIT

Jeonbuk National University, Republic of Korea

{jackaa, selfsolee}@jbnu.ac.kr

ABSTRACT

This paper describes the participation of the CBNU team at the TREC Precision Medicine Track 2019. We use the cancer-centered document clusters based on graph embedding. Documents are retrieved by re-ranking documents and pseudo-relevance feedback based on cancer-centered document clusters.

Keywords

Precision medicine, cancer-gene relation, cancer-centered document cluster, graph embedding

1. INTRODUCTION

In our participation to TREC 2019 Precision Medicine, we use the cancer-centered document clusters using clinical causal relationships and graph embedding for clinical document retrieval. In TREC 2019 PM, a biomedical document about patient cases typically describes a challenging medical case such as a patient's disease (type of cancer), the relevant genetic variants (which genes), basic demographic information (age, sex). Diseases can be detected using cancer-gene relation to a clinical query which is given a patient's disease and genes. Cancer-centered document clusters are constructed based on clinical causal relationships [1, 2, 3] and graph embedding.

2. SUBMITTED RUNS

In order to construct cancer-gene relation, we used 3,153 cancer gene's information and 181 cancers in Wikipedia. The cancer-gene relation has been represented as follows:

- **genes using a "genetic" field:** cancer genes are extracted in only 'genetic' field.
- **genes using all fields:** cancer genes are extracted in abstracts and 'genetic' field.

Clinical causal relationships were constructed using Unified Medical Language System (UMLS) and Wikipedia articles and cancer-gene relation has been constructed using TREC 2018 PM[1], GENT[4], GSEA[5], and ICD-10[6] information.

In order to create initial document clusters, four types of clinical causal relationships are used: disease-symptom, disease-test, disease-treatment and cancer-gene relationships. The retrieved documents can contain at least one of causal relationships.

For the other documents which are not belonging to the initial clusters, graph embedding method is applied for classification [7, 8]. For learning, the documents in an initial cluster are used as positive examples and other documents are used for negative examples. These documents are pseudo-relevant and pseudo-non relevant.

The detected diseases for a query are used to select particular document clusters and the clusters are used for pseudo-relevance feedback and re-ranking. Combining the initial retrieval results for an original query and the weights from the selected disease document clusters is applied.

$$Score_i(Q', D) = \lambda \cdot Score(Q, D) + (1-\lambda)\frac{1}{|C|}\sum_{i=1}^{|C|}\{Score_i(Q_{D-S}, C_i) + Score_i(Q_{D-T}, C_i) + Score_i(Q_{D-X}, C_i) + Score_i(Q_{C-G}, C_i))\} \tag{1}$$

where Q is an original query and |C| represents the number of document clusters. $Q_{D\text{-}S}$ represents Disease-Symptom relationships, $Q_{D\text{-}T}$ represents Disease-Test relationships, $Q_{D\text{-}X}$ represents Disease-Treatment relationships, and $Q_{C\text{-}G}$ represents cancer-gene relation. $Score_i(Q, D)$ is the initial document result. $Score_i(Q, D)$ is the initial document result. $Score(Q_{D\text{-}S}, C_i)$, $Score_i(Q_{D\text{-}T}, C_i)$, $Score_i(Q_{D\text{-}X}, C_i)$ and $Score(Q_{C\text{-}G}, C_i)$ represent the retrieval result for a Disease-Symptom relationships, Disease-Test relationships, Disease-Treatment relationships and Cancer-Gene relation of cancer *i*, respectively.

3. EXPERIMENTS

3.1 Run Description

Our experimental methods are described as follows:

- cbnuSA1: re-ranking documents for Scientific Abstracts (using BERT graph embedding)
- cbnuSA2: pseudo-relevance feedback for Scientific Abstracts (using BERT graph embedding)
- cbnuSA3: re-ranking documents for Scientific Abstracts (using med2vec graph embedding)
- cbnuSA4: pseudo-relevance feedback for Scientific Abstracts (using med2vec graph embedding)
- cbnuCT1: re-ranking documents for Scientific Abstracts (using BERT graph embedding)
- cbnuCT2: pseudo-relevance feedback for Scientific Abstracts (using BERT graph embedding)
- cbnuCT3: re-ranking documents for Scientific Abstracts (using med2vec graph embedding)
- cbnuCT4: pseudo-relevance feedback for Scientific Abstracts (using med2vec graph embedding)

3.2 Experimental Results

The experimental results for Scientific Abstracts are shown in Table 2.

RunID	infNDCG	P@10	R-Prec
cbnuSA1	0.4052	0.4575	0.2527
cbnuSA2	0.4097	0.4625	0.2483
cbnuSA3	0.3954	0.4275	0.2318
cbnuSA4	0.3960	0.4325	0.2317
Median	*0.4559*	*0.5450*	*0.2806*

Table 2. Experimental results for Scientific Abstracts

The experimental results for Clinical Trials are shown in Table 3.

RunID	infNDCG	P@10	R-Prec
cbnuSA1	0.5529	0.4842	0.4083
cbnuSA2	**0.5568**	**0.4921**	**0.4121**
cbnuSA3	0.5132	0.4289	0.3461
cbnuSA4	0.5107	0.4237	0.3461
Median	*0.5137*	*0.4658*	*0.3477*

Table 3. Experimental results for Clinical Trials

ACKNOWLEDGMENT

This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education(NRF-2017R1D1A1B03036275).

REFERENCES

[1] S. H. Jo, and K. S. Lee, “CBNU at TREC 2017 Clinical Decision Support Track”, In Proceedings of the 26th Text Retrieval Conference, 2017.

[2] S. H. Jo, and K. S. Lee, “CBNU at TREC 2016 Clinical Decision Support Track”, In Proceedings of the 25th Text Retrieval Conference, 2016.

[3] S. H. Jo, J. W. Seol and K. S. Lee, “CBNU at TREC 2015 Clinical Decision Support Track”, In Proceedings of the 24th Text Retrieval Conference, 2015.

[4] http://medical-genome.kribb.re.kr/GENT/

[5] http://software.broadinstitute.org/gsea/index.jsp

[6] https://icd.who.int/browse10/2016/en

[7] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”, In Proceedings of the NAACL-HLT 2019, 2019.

[8] E. Choi, M. T. Bahadori, E. Searles, C. Coffey, M. Thompson, J. J. Tejedor-Sojo, and J. Sun, “Multi-layer Representation Learning for Medical Concepts”, In Proceedings of the 22nd Knowledge Discovery and Data mining Conference, 2016