CBNU at TREC 2018 Precision Medicine Track

Seung-Hyeon Jo, Won-Kyu Choi, Kyung-Soon Lee
Division of Computer Science and Engineering, CAIIT
Chonbuk National University, Republic of Korea
{jackaa, smtj119, selfsolee}@chonbuk.ac.kr

ABSTRACT

This paper describes the participation of the CBNU team at the TREC Precision Medicine Track 2018. We propose construction of cancer-centered document clusters using gene information. Documents are retrieved by re-ranking documents and pseudo-relevance feedback based on cancer-centered document clusters.

Keywords

Precision medicine, diseases and genes, clinical causal knowledge, cancer-centered document cluster, Wikipedia

1. INTRODUCTION

The 2018 track focus on an important use case in clinical decision support: providing useful precision medicine-related information to clinicians treating cancer patient.

In our participation to TREC 2018 PM, we propose construction of cancer-gene relation for clinical document retrieval. In TREC 2018 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), and other potential factors that may be relevant. 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 cancer-gene relation.

2. SUBMITTED RUNS

In order to extract clinical terms belong to causal relationships, we have defined four clinical categories for UMLS semantic types: Disease, Symptom, Test, and Treatment.

Wikipedia articles are extracted by using UMLS terms belong to the Disease and Symptom clinical category. The following seven fields of the "contents" part are used for extraction of medical information: "Signs and symptoms", "Diagnosis", "Characteristics", "Complications", "Screening", and "Treatment". The clinical terms are extracted from fields. When the Wikipedia page does not have such fields in the contents, the terms for Symptom, Test, and Treatment category are extracted from the abstract part.

Cancer-gene relation has been constructed using cancer gene lists (Atlas [4], CANgenes [5], CIS [6], human Lymphoma, Miscellaneous, Sanger [7], Vogelstein [8]) and Wikipedia [9] articles.

The form of cancer gene's information is shown below. This information lets you know the genetic code and the name of the gene.

Gene Symbol	geneID	prevSymbols	Synonyms	Name	Organism
NKX2-2	4821	NKX2B	NKX2.2	NK2 HOMEOBOX 2	Human
MEN1	4221			MENIN 1	Human
FYN	2534		SYN, SLK, MGC45350	FYN PROTO-ONCOGENE, SRC FAMINY TYROSINE KINASE	Human
…	…	…	…	…	…

Table 1. Form of cancer gene's information

In this paper, we used 2,027 cancer gene's information and 181 cancers in Wikipedia. The cancer-gene relation has been defined in two ways. The number of cancer-gene relation is 181.

- **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) [10] and Wikipedia articles. The clinical causal relationships were represented as follows:

- **DISEASE-SYMPTOM** relation: < disease$_i$: symptom$_{i1}$, symptom$_{i2}$ … >
- **DISEASE-TEST** relation: < disease$_j$: test$_{j1}$, test$_{j2}$ … >
- **DISEASE-TREATMENT** relation: < disease$_k$: treatment$_{k1}$, treatment$_{k2}$ … >
- **CANCER-GENE** relation: < disease$_l$: gene$_{l1}$, gene$_{l2}$ … >

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.

Genes and names of genes can be obtained through genetic information. And other cancer information can be obtained through cancer-gene relation. In this paper, using the cancer-gene relation and cancer gene’s information, expansion terms are selected in a query. When using relation with all fields of Wikipedia, genes in “genetics” field are weighted 1, other genes are weighted 0.5.

For the other documents which are not belong to the initial clusters, CNN(Convolutional Neural Networks) method is applied for classification [11, 12]. 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.

$$Score_i(Q', D) = Score_i(Q_{C-G}, C_i))\} \quad (1)$$

where Q is an original query and |C| represents the number of document clusters. Q_{D-S} represents Disease-Symptom relationships, Q_{D-T} represents Disease-Test relationships, Q_{D-X} represents Disease-Treatment relationships, and Q_{C-G} represents cancer-gene relation. $Score_i(Q, D)$ is the initial document result. $Score_i(Q, D)$ is the initial document result. $Score(Q_{D-S}, C_i)$, $Score_i(Q_{D-T}, C_i)$, $Score_i(Q_{D-X}, C_i)$ and $Score(Q_{C-G}, C_i)$ 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: query expansion + re-ranking documents for Scientific Abstracts
- cbnuSA2: query expansion + pseudo-relevance feedback for Scientific Abstracts
- cbnuSA3: query expansion (age and treatment terms priority) + re-ranking documents for Scientific Abstracts
- cbnuSA1: query expansion + re-ranking documents for Clinical Trials
- cbnuSA2: query expansion + pseudo-relevance feedback for Clinical Trials
- cbnuSA3: query expansion (age and treatment terms priority) + re-ranking documents for Clinical Trials

3.2 Experimental Results

The experimental results for Scientific Abstracts are shown in Table 2. The cbnuSA1 and cbnuSA2 shows significant improvement over the median.

RunID	infNDCG	P@10	R-Prec
cbnuSA1	0.4523	0.5520	0.2992
cbnuSA2	0.4347	0.5440	0.2765
cbnuSA3	0.1880	0.2300	0.0985
Median	0.4291	0.5460	0.2672

Table 2. Experimental results for Scientific Abstracts

The experimental results for Clinical Trials are shown in Table 3. The cbnuCT1 and cbnuCT2 shows improvement over the median.

RunID	infNDCG	P@10	R-prec
cbnuCT1	0.4572	0.4680	0.3382
cbnuCT2	0.4597	0.4620	0.3493
cbnuCT3	0.2722	0.2340	0.1887
Median	0.4297	0.4680	0.3268

Table 3. Experimental results for Clinical Trials

4. CONCLUSIONS

In this paper, we propose construction of cancer-gene relation for clinical document retrieval and construction of cancer-centered clusters using clinical causal relationships and convolution neural networks.

5. ACKNOWLEDGEMENTS

This research was supported by the MSIT(Ministry of Science and ICT), Korea, under the ITRC(Information Technology Research Center) support program (IITP-2017-2015-0-00378) supervised by the IITP(Institute for Information & communications Technology Promotion), and 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] J. L. Huret, S. L. Minor, F. Dorkeld, P. Dessen, and A. Bernheim. “Atlas of genetics and cytogenetics in oncology and haematology, an interactive database”, Nucleic Acids Research, 2000.

[5] K. Akagi, T. Suzuki, R. M. Stephens, N. A. Jenkins, and N. G. Copeland. “RTCGD: retroviral tagged cancer gene database”, Nucleic Acids Research, 2004.

[6] J. M. Coffin, S. H. Hughes, and H. E. Varmus, “Retroviruses”, Cold Spring Harbor Press, Cold Spring Harbor, 1997.

[7] M. E. Higgins, M. Claremont, J. E. Major, C. Sander, and A. E. Lash. “CancerGenes: a gene selection resource for cancer genome projects”, Nucleic Acids Research, 2007.

[8] B. Vogelstein, N. Papadopoulos, V. E. Velculescu, S. Zhou, L. A. Diaz Jr, and K. W. Kinzler. “Cancer genome landscapes”, Science. 2013.

[9] http://en.wikipedia.org

[10] O. Bodenreider. “The Unified Medical Language System(UMLS): intergrating biomedical terminology”. Nucleic Acids Research, vol. 32, pp. D267–D270, 2004.

[11] J. Xu, P. Wang, G. Tian, B. Xu, J. Zhao, F. Wang, H. Hao. “Short Text Clustering via Convolutional Neural Networks”, In Proceedings of NAACL-HLT 2015.

[12] H. He, K. Gimpel, and J. Lin. “Multiperspective sentence similarity modeling with convolutional neural networks”, In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, 2015.