CBNU at TREC 2017 Precision Medicine Track

Seung-Hyeon Jo, Kyung-Soon Lee

Division of Computer Science and Engineering, CAIIT
Chonbuk National University, Republic of Korea
{jackaa, selfsolee}@chonbuk.ac.kr

ABSTRACT

This paper describes the participation of the CBNU team at the TREC Precision Medicine Track 2017. We have constructed cancer-centered document clusters using cancer-gene relation and clinical causal information. A query has been expanded with disease terms and pseudo-relevance feedback is applied for cancer disease document clusters.

Keywords

Precision medicine, cancer-gene relationship, clinical causal knowledge, cancer document cluster

1. INTRODUCTION

The TREC Precision Medicine (PM) Track 2017 focuses on the case of providing clinical decision support to cancer patients with genetic variations. In our participation to TREC PM Track 2017, we propose construction of cancer disease clusters based on cancer-gene relation for re-ranking and pseudo-relevance feedback. A clinical document for 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 for a clinical query which is given a patient's disease and genes. Cancer-centered document clusters are constructed based on cancer-gene relation and clinical causal relationships [1, 2].

2. Construction of Clusters Based on Cancer-Gene and Clinical Causal Relationship

Cancer-gene relation is extracted from the cancer gene lists (Atlas [3], CANgenes [4], CIS [5], human Lymphoma, Miscellaneous, Sanger [6], Vogelstein [7]) and Wikipedia [8] articles. The examples of cancer gene's information are shown in Table 1. This information provides the genetic code and the name of the gene. The total number of cancer gene's information we used is 2027.

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. Examples of cancer gene's information

For our experiments, the cancer-gene relation is extracted from the Genetic field and other fields in Wikipedia. The number of cancer-gene relation is 181.

- **genes from Genetic field:** cancer genes are extracted from only the 'Genetic' field.
- **genes from all fields:** cancer genes are extracted from abstracts and 'Genetic' field.

The cancer-gene relation extracted is represented as follows:

- **Cancer-Gene** relation: < $disease_j$: $gene_{j1}$, $gene_{j2}$ … >

Clinical causal relationships were constructed in our previous experiments [1, 2] using Unified Medical Language System (UMLS) [9] and Wikipedia articles. The clinical causal relationships were represented as follows:

- **SYMPTOM-DISEASE** relation: < $symptom_i$: $disease_{i1}$, $disease_{i2}$ ... >

- **DISEASE-SYMPTOM** relation: < $disease_j$: $symptom_{j1}$, $symptom_{j2}$ … >
- **TEST-DISEASE** relation: < $test_k$: $disease_{k1}$, $disease_{k2}$ ... >
- **DISEASE-TEST** relation: < $disease_l$: $test_{l1}$, $test_{l2}$ … >
- **TREATMENT-DISEASE** relation: < $treatment_m$: $disease_{m1}$, $disease_{m2}$ ... >
- **DISEASE-TREATMENT** relation: < $disease_n$: $treatment_{n1}$, $treatment_{n2}$ … >

In order to create initial disease clusters, three types of clinical causal relationships are used: disease-symptom, disease-test, and disease-treatment relationships. The retrieved documents contain all the causal relationship terms.

- **Disease-Symptom** relationships: <$disease_x$: $symptom_{x1}$, $symptom_{x2}$, … >
- **Disease-Test** relationships: <$disease_x$: $test_{x1}$, $test_{x2}$, ... >
- **Disease-Treatment** relationships: <$disease_x$: $treatment_{x1}$, $treatment_{x2}$, ... >

Figure 1 shows the clinical causal relationships with cancer-gene relation. These causal relations are used for cancer-centered document clustering and expanding disease terms.

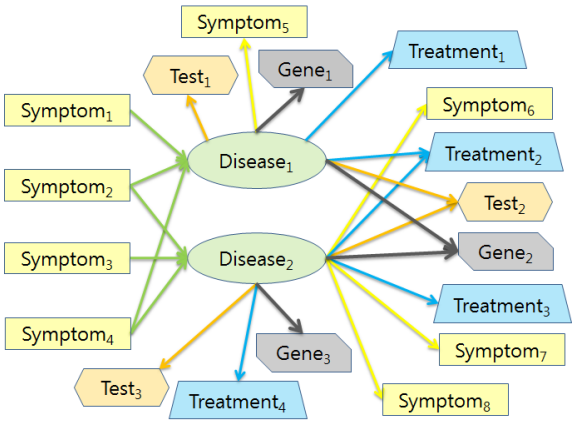

Figure 1. Clinical causal relationships including cancer-gene information

An initial cancer disease cluster contains documents which have all four types of relationships such as Disease-Symptom, Disease-Test, and Disease-Treatment, and Cancer-Gene relation terms. In Figure 2, the selected documents are used for constructing an initial cancer cluster. The number of initial and final cancer clusters is 181, which is the number of cancers extracted. The number of documents of initial cancer clusters is 5,283 for Scientific Abstracts collection and 2,376 for Clinical Trials collection.

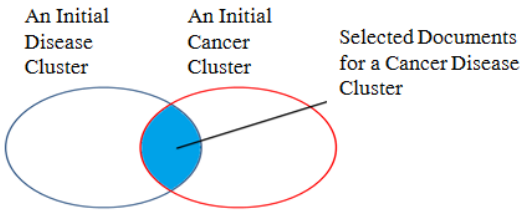

Figure 2. Creating a cancer disease cluster

For the other documents which are not belong to the initial clusters, CNN(Convolutional Neural Network) method is applied for classification[10]. 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. The selected clusters are used for pseudo-relevance feedback and re-ranking. For pseudo-relevant documents, the top 25 documents are selected based on Re-ranking scores. For re-ranking, the initial retrieval result for an original query and the scores from the selected clusters is combined as the following Equation.

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

where Q is an original query and |C| represents the number of document clusters. $Q_{D\text{-}S}$, $Q_{D\text{-}T}$, $Q_{D\text{-}X}$ *and* $Q_{C\text{-}G}$ represents a query using Disease-Symptom, Disease-Test, Disease-Treatment, and Cancer-Gene relationship terms, respectively. *Score(Q, D)* is the initial retrieval score for a query. $Score(Q_{D\text{-}S}, C_i)$, $Score(Q_{D\text{-}T}, C_i)$, $Score(Q_{D\text{-}X}, C_i)$ and $Score(Q_{C\text{-}G}, C_i)$ represents the retrieval result of a document D for the cluster C_i in a Disease-Symptom relationship, Disease-Test relationship, Disease-Treatment relationship and Cancer-Gene relation for cancer i, respectively. In our experiments, Indri search engine [11] is used for baseline retrieval.

3. SUBMITTED RUNS

3.1 Run Description

Our experimental methods are described as follows.

- **cbnuSA1**: Re-ranking based on the cancer-gene relation for Scientific Abstracts (genes from "genetic" field)
- **cbnuSA2**: Re-ranking based on the cancer-gene relation for Scientific Abstracts (genes from all fields)
- **cbnuSA3**: Pseudo-Relevance Feedback for cbnuSA2
- **cbnuSA4** (non-submitted): Pseudo-Relevance Feedback for cbnuSA1
- **cbnuCT1**: Re-ranking based on the cancer-gene relation for Clinical Trials (genes from "genetic" field)
- **cbnuCT2**: Re-ranking based on the cancer-gene relation for Clinical Trials (genes from all fields)
- **cbnuCT3**: Pseudo-Relevance Feedback for cbnuCT2
- **cbnuCT4** (non-submitted): Pseudo-Relevance Feedback for cbnuCT1

3.2 Experimental Results

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

RunID	infNDCG	P@10	R-Prec
cbnuSA1	**0.3139**	**0.4483**	**0.2135**
cbnuSA2	0.1143	0.1621	0.0928
cbnuSA3	0.1364	0.1897	0.1020
cbnuSA4	**0.3218**	**0.4614**	**0.2287**
Median	0.2685	0.3586	0.1739
Best	0.5782	0.8552	0.3928

Table 2. Experimental results for Scientific Abstracts

The experimental results for Clinical Trials are shown in Table 3. The cbnuCT1 and cbnuCT3 shows improvement over the median.

RunID	P@5	P@10	P@15
cbnuCT1	**0.3931**	**0.3379**	**0.2897**
cbnuCT2	0.2207	0.2241	0.2046
cbnuCT3	0.2759	0.2586	0.2391
cbnuCT4	**0.4003**	**0.3574**	**0.3026**
Median	0.2929	0.2536	0.2262
Best	0.7714	0.6750	0.5905

Table 3. Experimental results for Clinical Trials

Figure 3 shows a comparison result on each topic for cbnuSA1 and median. The 'Genetic' filed information was used for the topic 11, 13, 14, and 17 ("Gastric cancer", "Cholangiocarcinoma", and "Prostate cancer"). However, for the topic 25, 27, and 28 ("Lung adenocarcinoma", "Pancreatic adenocarcinoma"), the genetic information from Wikipedia was not used.

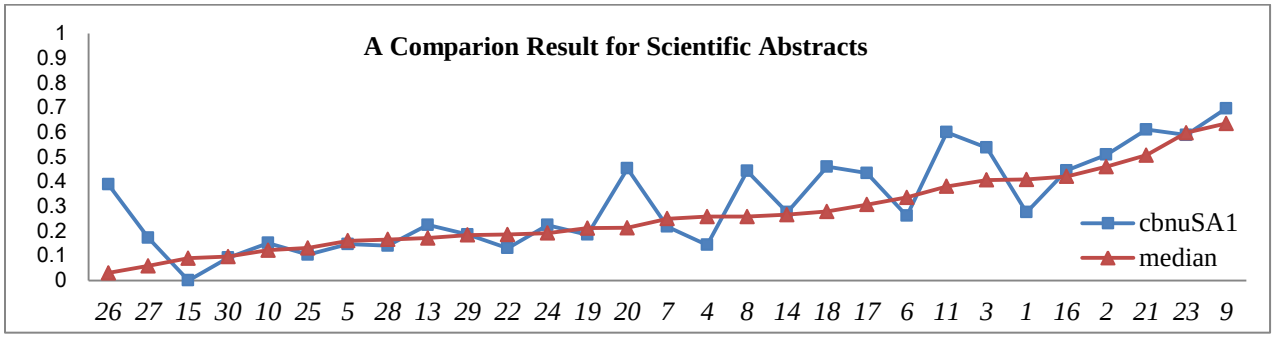

Figure 3. Result of each topic for Scientific Abstract on infNDCG

4. CONCLUSIONS

In this paper, we proposed cancer document clusters based on cancer-gene and causal relation for TREC Precision Medicine track. Experimental results show that the cancer disease clusters based on causal relations are effective for retrieval, and the delicate construction of cancer-gene relation contributes retrieval performance.

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 2016 Clinical Decision Support Track", In Proceedings of the 25th Text Retrieval Conference, 2016.

[2] 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.

[3] 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.

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

[5] J. M. Coffin, S. H. Hughes, and H. E. Varmus, "Retroviruses", Cold Spring Harbor Press, Cold Spring Harbor, 1997.

[6] 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.

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

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

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

[10] J. Xu, P. Wang, G. Tian, B. Xu, J. Zhao, F. Wang, and H. Hao. "Short Text Clustering via Convolutional Neural Networks", In Proceedings of the North American Chapter of the Association for Computational Linguistics – Human Language Technologies (NAACL HLT), 2015.

[11] T. Strohman, D. Metzler, H. Turtle, and W. B. Croft, "Indri: A language model-based search engine for complex queries". In Proceedings of International Conference on Intelligence Analysis, 2005. http://www.lemurproject.org/indri