CBNU at TREC 2016 Clinical Decision Support Track

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

ABSTRACT

This paper describes the participation of the CBNU team at the TREC Clinical Decision Support track 2016. We propose construction of disease-centered document clusters and semantic word vectors using word embeddings. Hierarchical disease-centered document clusters are constructed based on clinical causal relationships such as disease-symptom, disease-test, and disease-treatment relationships. Semantic word vectors for medical terms are constructed by using word2vec. Documents are retrieved by expanding disease terms and semantic words for a clinical query, and by re-ranking using disease document clusters.

Keywords

clinical decision support, clinical causal knowledge, disease-centered document cluster, word embeddings, re-ranking, UMLS, Wikipedia

1. INTRODUCTION

The goal of the Clinical Decision Support (CDS) track is to retrieve biomedical articles relevant for answering generic clinical questions about medical records [1].

Based on our observation that generic clinical questions about patient cases are related to diseases, we assume that adding the disease terms for the given a list of symptoms can improve retrieval effectiveness. Also, based on our observation that a clinical document is described with at least one disease, we assume that focusing disease-centered document clusters which are related to a disease can be helpful to clinical document retrieval.

In our participation to TREC 2016 CDS, we propose construction of clinical causal knowledge for clinical document retrieval. A biomedical document about patient cases typically describes a challenging medical case such as a patient's medical problem and a physician's action. Diseases can be detected using clinical causal knowledge to a clinical query which is given a list of symptoms to describe a patient's situation. Disease-centered document clusters are constructed based on clinical causal relationships and fine-grained by MeSH categories.

Semantic word vectors are constructed based on word embeddings. The detected clusters and semantic word vectors to a query are used for query expansion, pseudo-relevance feedback, and re-ranking for improving retrieval effectiveness.

2. ESTABLISHING A CLINICAL CAUSAL KNOWLEDGE

2.1 Clinical Causal Relationships Using UMLS and Wikipedia

Clinical causal relationships are constructed using Unified Medical Language System (UMLS) and Wikipedia articles [11]. The relationship forms are 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: < $disease_n$: $treatment_{n1}$, $treatment_{n2}$ … >

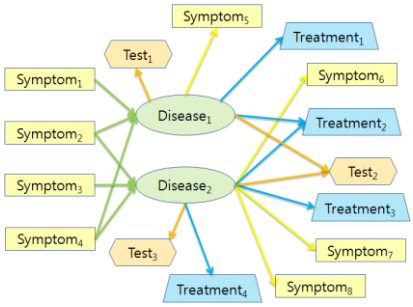

Figure 1. Construction of clinical causal relationships

2.2 Disease-Centered Document Clusters Based on Clinical Causal Relationships

Disease-centered document clusters based on clinical causal relationships can be effectively used for detecting the disease documents for a given list of symptoms or a given situation since medical documents are mainly

described with symptom, test, and treatment terms for a specific disease.

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

- 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 2 shows the example of initial disease document clustering. Documents are examined using three types of relationships for each disease. For the group A, each document certainly contains disease-symptom and disease-test relation terms. For the group B, a document contains disease-symptom and disease-treatment relation terms. For the group C, a document contains disease-test and disease-treatment relation terms. For the group D, all the documents contain disease-symptom, disease-test and disease-treatment relation terms. The number of initial disease document clusters is 18,442, which is the same number of diseases extracted from Wikipedia titles.

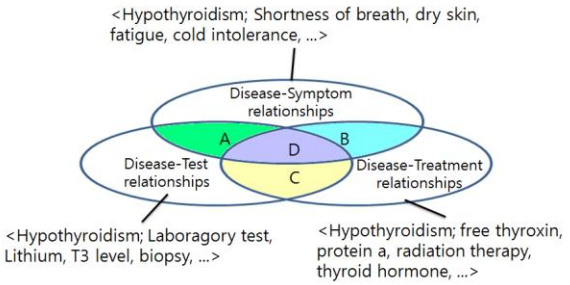

Figure 2. The Initial disease document cluster for a disease "Hypothyroidism"

Medical Subject Headings (MeSH) is a comprehensive controlled vocabulary for the purpose of indexing journal articles and books in the life sciences; it serves as a thesaurus that facilitates searching.

If 'disease A' and 'disease B' are similar, terms for these diseases such as symptom, test, and treatment are likely to be similar. Similar disease documents are clustered based on MeSH categories.

The top level of MeSH hierarchy has disease category. The MeSH disease hierarchy consists of 23, 271 and 10543 categories at the level 1, level 2, and level 4, respectively. We use the MeSH level 2 for 10543 disease categories. The 6327 MeSH disease categories are associated to the TREC2016 CDS documents.

Figure 3 shows that disease document clusters based on MeSH hierarchy. The initial disease document cluster of "*Thyroid diseases*" contains all the documents from "Hypothyroidism", "Goiter" and "Thyroid dysgenesis" since the disease "Hypothyroidism", "Goiter", and "Thyroid dysgenesis" belong to "Thyroid diseases".

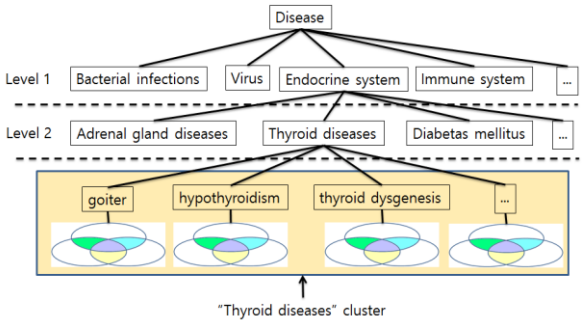

Figure 3. Document clusters based on MeSH hiearchy

However, when a disease does not exist in MeSH categories, the documents for the disease can not belong to any disease cluster. To deal these disease documents, document similarity measure is used to group these documents to the existing diseases in MeSH categories. If clinical terms which describe a disease are similar, these diseases are likely to be similar.

In order to calculate similarity, clinical terms are represented as a centroid vector for a disease using disease, symptom, test, and treatment terms extracted from the abstract part in a document. If the cosine similarity is above the threshold, two documents can be grouped as the same cluster. In our experiment, the threshold is set to 0.7, which is learned from the training set.

Figure 4 shows disease-centered document clustering based on similarity. For example, the disease "Hashitoxicosis" which does not exist in MeSH categories belong to "Thyroid disease" cluster by similarity measure.

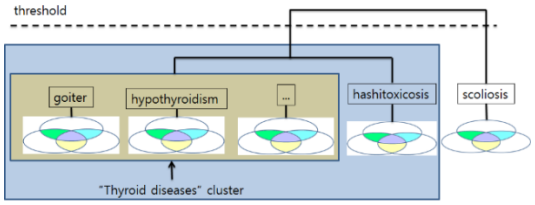

Figure 4. Document clusters based on similarity

2.3 Constructing Semantic Word Vectors Based on Word Embeddings

Artificial neural networks are a family of models inspired by biological neural networks which are used for classification and clustering of data. Word2vec algorithm [5] is artificial neural network for processing text which is used for learning word embeddings.

Semantic word vectors of medical terms are extracted by using word2vec algorithm. Medical terms are extracted from Wikipedia. Figure 5 shows the method of constructing

semantic word vectors of "hypothyroidism" and "mild symptom" medical terms.

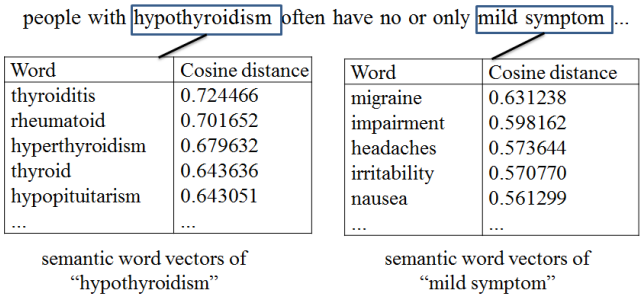

Figure 5. Constructing word vectors of medical terms using neural word embeddings

3. DOCUMENT RETREIVAL USING CLINICAL CAUSAL KNOWLEDGE

3.1 Query Expansion Based on Word Embeddings

Using the symptom-disease relationships, disease terms are detected for given symptom terms in a query. At least three symptom terms should be matched to detect a disease.

Expansion terms are selected from semantic word vectors of detected diseases. Expansion terms are weighted adding the semantic word vectors expansion terms are selected having high weight. The number of expansion terms is e.

$$Weight(t) = \sum_{i=1}^{|C|} cos(V_i, t) \qquad (1)$$

where t is a medical term. $cos(V_i, t)$ is the cosine similarity for semantic word vector V_i. and |C| is the number of detected diseases. Figure 6 shows the method of selecting expansion terms using word embeddings. In figure 6, $cos(V_{Hypothyroidism}, rheumatoid)$ is 0.7016 and $cos(V_{systemic\ lupus\ erythematosus}, rheumatoid)$ is 0.6234. Thus, $Weight(rheumatoid)$ is 0.7016+0.6234 = 1.3250.

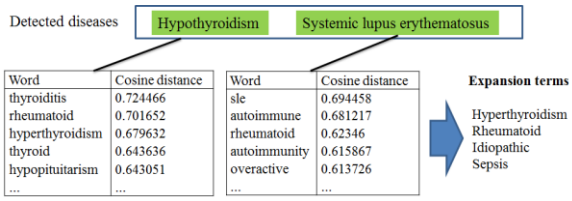

Figure 6. Selecting expansion terms using word embeddings

3.2 Re-ranking Documents Using Disease-Centered Document Clusters

Disease-centered document clusters can be used to improve retrieval effectiveness by giving preference to the document clusters which contain diseases related to a query.

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.

$$QL'(Q, D) = \lambda \cdot QL(Q, D) + (1 - \lambda) \frac{1}{|C|} \sum_{i=1}^{|C|} CL(Q', C_i) \quad (2)$$

where Q is an original query and Q' represents three types of relationships for initial clusters. QL(Q, D) is the initial document result. CL(Q', C_i) represents the retrieval result for a disease-centered document cluster for 'disease i'. |C| is the number of disease-centered document clusters. The parameter λ is set to 0.6, which is learned from training set.

4. EXPERIMENTS

4.1 Run Description

Our methods are described as follows:

- baseline_summ: baseline using solr search engine [6] (using "summary" part)
- cbnu_s1: pseudo-relevance feedback based on the clinical causal knowledge (using "summary" part)
- cbnu_s2: re-ranking based on the clinical causal knowledge (using "summary" part)
- baseline_note: baseline using solr search engine (using "EHR note" part)
- cbnu_n1: pseudo-relevance feedback based on the clinical causal knowledge (using "EHR note" part)
- cbnu_n2 (non-submitted run): re-ranking based on the clinical causal knowledge (using "EHR note" part)

4.2 Experimental Results

The experimental results are shown in Table 1. The proposed method shows significant improvement over the median.

Table 1. Experimental results

RunID	infNDCG	R-prec	P@10
median (using summary)	0.1859	0.1219	0.2633
baseline_summ	0.1927	**0.1384**	0.2800
cbnu_s1	**0.2382**	0.1153	**0.3400**
cbnu_s2	0.2365	0.1225	0.3367
median (using EHR note)	0.1228	0.0792	0.1833
baseline_note	0.1157	0.0816	0.1967
cbnu_n1	**0.1723**	**0.0935**	**0.2467**
cbnu_n2	0.1738	0.0916	0.2533

5. CONCLUSIONS

Using causal knowledge for retrieving biomedical documents is helpful. The disease-centered clustering using clinical causal relationships could be effectively used for re-ranking documents. And word embeddings could be effectively used for query expansion.

6. ACKNOWLEDGEMENTS

This research was supported by the MSIP(Ministry of Science, ICT and Future Planning), Korea, under the ITRC(Information Technology Research Center) support program (IITP-2016-R0992-15-1023) supervised by the IITP(Institute for Information & communications Technology Promotion). This work was partially supported by the Brain Korea 21 PLUS Project, National Research Foundation of Korea.

REFERENCES

[1] K. Roberts and M. S. Simpson. “Overview of the TREC 2015 Clinical Decision Support Track”. In Proceedings of the 24th Text Retrieval Conference, 2015.

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

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

[4] C. E. Lipscomb. “Medical Subject Headings (MeSH)”. Bulletin of the Medical Library Association, vol. 88, pp. 265-266, 2000.

[5] T. Mikolov, K. Chen, G. Corrado, and J. Dean. “Efficient Estimation of Word Representations in Vector Space”. In proceedings of the 1st International Conference on Learning Representations (ICLR’13), 2013.

[6] http://lucene.apache.org/solr/

[7] S. Logeswari and K. Premalatha. "Biomedical Document Clustering Using Ontology based Concept Weight". In proceedings of the 2013 International Conference on Computer Communication and Informatics (ICCCI′13), 2013.

[8] R. Prasath and P. O′Reilly. ″Exploring Clustering Based Knowledge Discovery towards Improved Medical Diagnosis″. In proceedings of the Medical Information Retrieval Workshop at SIGIR2014 (MedIR′14), pp 12-15, 2014.

[9] W. Yonghui, X. Jun, J. Min, Z. Yaoyun and X. Hua. ″A Study of Neural Word Embeddings for Named Entity Recognition in Clinical Text″, In Proceedings of the AMIA Annual Symposium Proceedings, pp.1326-1333, 2015.

[10] L. Yue, G. Tao, S. M. Kusum, J. Heng and L. M. Deborah. ″Exploiting Task-Oriented Resources to Learn Word Embeddings for Clinical Abbreviation Expansion″, In Proceedings of the 2015 Workshop on Biomedical Natural Language Processing (BioNLP`15), pp.92–97, 2015.

[11] 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.