CBNU at TREC 2015 Clinical Decision Support Track

Seung-Hyeon Jo
Division of Computer Science and Engineering, CAIIT
Chonbuk National University
Jeonju, Republic of Korea
jackaa@jbnu.ac.kr

Jae-Wook Seol
Korea Institute of Science and Technology Information
Daejeon, Republic of Korea
wodnr754@kisti.re.kr

Kyung-Soon Lee
Division of Computer Science and Engineering, CAIIT
Chonbuk National University
Jeonju, Republic of Korea
selfsolee@jbnu.ac.kr

ABSTRACT

This paper describes the participation of the CBNU team at the TREC Clinical Decision Support track 2015. We propose a query expansion method based on a clinical semantic knowledge and a topic model. The clinical semantic knowledge is constructed by using medical terms extracted from Unified Medical Language System (UMLS) and Wikipedia articles. The word and document topics are generated by using a topic model for a document collection. The proposed methods achieved 0.2327 and 0.3033 in the inferred NDCG on Task A and Task B, respectively.

Keywords

clinical decision support, clinical semantic knowledge, query expansion, UMLS, Wikipedia, topic model

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].

Basically, we assume that disease terms are helpful for query expansion for all kind of query types. Besides, the different kinds of expansion terms would be effective according to the query types such as diagnosis, treatment, and test.

In our participation to TREC 2015 CDS, we propose a query expansion method based on a clinical semantic knowledge and a topic model. The first step is to establish a clinical semantic knowledge by using medical terms extracted from Unified Medical Language System (UMLS) [3] concepts and Wikipedia articles. The second step is to generate document topics and word topics for the CDS document collection using Latent Dirichlet Allocation(*LDA*). Then the query expansion step is to select expansion terms from clinical semantic knowledge, and word topics associated with the document topics of an initial retrieval set.

2. ESTABLISHING A CLINICAL SEMANTIC KNOWLEDGE BASED ON UMLS AND WIKIPEDIA

2.1 Defining clinical categories for UMLS semantic types

The UMLS semantic types represent intensional knowledge, while the UMLS concepts that are assigned to those types represent extensional knowledge.

We have defined four clinical categories for 27 semantic types which are related to medical terms among a total of 133 semantic types. The clinical categories defined are Disease, Symptom, Test, and Treatment. Table 1 shows the clinical categories and the associated UMLS semantic types.

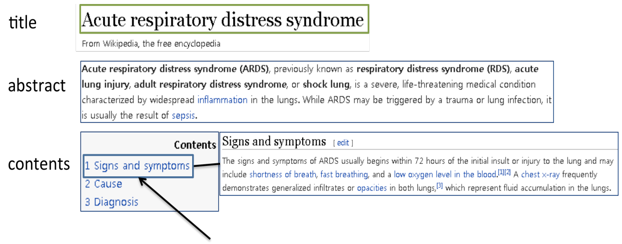
Table 1. Four clinical categories for a query expansion

Clinical categories	UMLS semantic types	# of concepts (terms)
Disease	“Anatomical Abnormality”, “Congenital Abnormality”, “Acquired Abnormality”, “Pathologic Function”, “Disease or Syndrome”, “Mental or Behavior Dysfunction”, “Neoplastic Process”	610,356
Symptom	“Signs and Symptoms”, “Pathologic Function”, “Disease or Syndrome”, “Mental or Behavior Dysfunction”, “Neoplastic Process”, “Finding”	1,224,254
Test	“Medical Device”, “Drug Delivery Device”, “Research Device”, “Indicator, Reagent, or Diagnostic Acid”, “Laboratory Procedure”, “Diagnostic Procedure”	296,161
Treatment	“Biomedical or Dental Material”, “Therapeutic or Preventive Procedure”, “Clinical Drug”, “Pharmacologic Substance”, “Antibiotic”, “Biologically Active Substance”, “Neuroreactive Substance or Biogenic Amine”, “Hormone”, “Enzyme”, “Vitamin”, “Immunologic Factor”, “Receptor”, “Amino Acid, Peptide, or Protein”	609,675

2.2 Establishing a clinical semantic knowledge from Wikipedia and UMLS

Wikipedia articles are used for enriching a Clinical semantic knowledge. The articles are extracted by using UMLS terms belong to the Disease and Symptom clinical category. The number of Wikipedia articles extracted is 60,982.

The Wikipedia article [2] consists of title, abstract, and contents. Here, disease terms are extracted from the “title” part.



Acute respiratory distress syndrome

From Wikipedia, the free encyclopedia

Acute respiratory distress syndrome (ARDS), previously known as **respiratory distress syndrome (RDS)**, **acute lung injury**, **adult respiratory distress syndrome**, or **shock lung**, is a severe, life-threatening medical condition characterized by widespread inflammation in the lungs. While ARDS may be triggered by a trauma or lung infection, it is usually the result of sepsis.

Contents

1 Signs and symptoms
2 Cause
3 Diagnosis

Signs and symptoms [edit]

The signs and symptoms of ARDS usually begins within 72 hours of the initial insult or injury to the lung and may include shortness of breath, fast breathing, and a low oxygen level in the blood.[1][2] A chest x-ray frequently demonstrates generalized infiltrates or opacities in both lungs,[3] which represent fluid accumulation in the lungs.

Figure 1. The example of a Wikipedia article

The seven fields of the “contents” part are used for extraction of medical information: “Signs and symptoms”, “Diagnosis”, “Characteristics”, “Complications”, “Screening”, “Treatment”,

Table 2 shows that the fields of contents corresponding to the clinical categories.

Table 2. Fields in a Wikipedia article for the clinical categories

Clinical categories	Fields of Contents	# of UMLS concepts
Disease	title	18,442
Symptom	“Signs and symptoms”, “Diagnosis”, “Characteristics”, “Complications” field	70,312
Test	“Diagnosis”, “Screening” field	36,048
Treatment	“Treatment”, “Management” field	57,625

The terms for Symptom, Test, and Treatment are extracted from the corresponding fields of contents in a Wikipedia page, as shown in Table 2. The terms are extracted from “Signs and symptoms”, “Diagnosis”, “Characteristics”, and “Complications” fields for the Symptom category, the terms are extracted from “Diagnosis”, and “Screening” fields for the Test category, and the terms are extracted from “Treatment”, “Management” fields for the Treatment category.

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.

A clinical semantic knowledge is established from these terms extracted by matching UMLS.

The clinical semantic knowledge forms are as follow:

- SYMPTOM-DISEASE relation: < **symptom**: disease$_1$, disease$_2$ ... >
- DISEASE-SYMPTOM relation: < **disease**: symptom$_1$, symptom$_2$ … >
- TEST-DISEASE relation: < **test**: disease$_1$, disease$_2$ ... >
- DISEASE-TEST relation: < **disease**: test$_1$, test$_2$ … >
- TREATMENT-DISEASE relation: < **treatment**: disease$_1$, disease$_2$ ... >
- DISEASE-TREATMENT relation: < **disease**: treatment$_1$, treatment$_2$ … >

2.3 Generating documents and words topics

On the other hand, we have used documents topics and words topics for a query expansion. The topics are generated for the document collection using LDA.

Latent Dirichlet allocation (LDA) is a generative probabilistic model. The basic idea is that documents are represented as random mixtures over latent topics, where each topic is characterized by a distribution over words [4]. Our system runs a query expansion using the words topics corresponding to the documents topics for the initial retrieval set.

3. SELECTING EXPANSION TERMS FROM THE CLINICAL SEMANTIC KNOWLEDGE AND TOPICS

3.1 Expansion terms from the clinical semantic knowledge

For a query expansion, the medical terms are selected based on our clinical semantic knowledge.

The expansion terms are selected as follows:

1) Extracting symptom terms from a query.
2) Selecting disease terms based on the symptom terms in the query and SYMPTOM-DISEASE relations.
3) Selecting symptom, test, and treatment terms based on DISEASE-SYMPTOM, DISEASE-TEST, and “DISEASE-TREATMENT relations.
4) Selecting expansion terms according to the highest frequency.

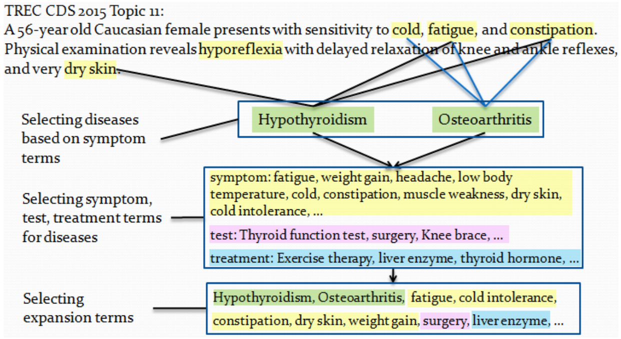

TREC CDS 2015 Topic 11:
A 56-year old Caucasian female presents with sensitivity to cold, fatigue, and constipation. Physical examination reveals hyporeflexia with delayed relaxation of knee and ankle reflexes, and very dry skin.
Selecting diseases based on symptom terms
Hypothyroidism
Osteoarthritis
Selecting symptom, test, treatment terms for diseases
symptom: fatigue, weight gain, headache, low body temperature, cold, constipation, muscle weakness, dry skin, cold intolerance, ...
test: Thyroid function test, surgery, Knee brace, ...
treatment: Exercise therapy, liver enzyme, thyroid hormone, ...
Selecting expansion terms
Hypothyroidism, Osteoarthritis, fatigue, cold intolerance, constipation, dry skin, weight gain, surgery, liver enzyme, ...

Figure 2. Selecting expansion terms based on the clinical semantic knowledge

3.2 Expansion terms from documents and words topics

For the initial retrieval set, document topics are selected when the document topic contains the top-retrieved documents (*R*) in the highest rank (*M*). Then the corresponding word topics are selected when the word topic contains query terms (*S*) in the highest rank (*N*).

If the document topic and the word topic selected are the same topic, then expansion terms are selected from the word topics.

The parameters are learned by using the TREC 2014 CDS collection.

- # of topics (*D*): {50, 100, 150, **200**, 250, 500}
- # of words in the highest rank in a word topic (*N*): {10, 13, **15**, 17, 20, 25}
- # of query terms belong to a word topic (*S*): {**4**, 5, 6, 7, 8, 9, 10}
- # of documents in the highest rank in a document topic (*M*): {**50**, 100, 150, 200, 250, 500}

- # of retrieved documents (*R*): {100, **200**, 500, 1000}
- # of the retrieved documents belong to a document topic (*T*): {10, **15**, 20, 25, 50}

Note that the bold character indicates the final parameter.

3.3 Expansion terms from synonym of disease terms

The synonyms of disease terms are used for a query expansion.

In the UMLS Metathesaurus, each unique medical concept is associated with a "concept unique identifier" or CUI such that all entries in the Metathesaurus which refer to the same concept share the same CUI.

Wikipedia also contains pages called redirects which do not contain content themselves, but rather redirect a user to another article (or section of an article). Redirects typically embody alternate names, spellings, forms, closely related words, alternately punctuated or encoded forms, less specific forms in which the redirected name is the primary topic, or more specific forms of some other page.

The synonyms of disease terms are selected by UMLS CUI and Wikipedia redirects.

3.4 Expansion terms according to the query types (Task B)

The different kinds of expansion terms would be effective according to the query types such as diagnosis, treatment, and test. In our experiments, the expansion terms are selected according to the query types. The query types and expansion term categories are as follow.

Table 3. Expansion terms according to the query types

TREC CDS Query Types	Expansion terms
Diagnosis	Disease and Symptom terms from Disease and Symptom categories
Test	Disease and Test terms from Disease and Test categories
Treatment	Disease and Treatment terms from Disease and Treatment categories

3.5 Query expansion methods

Expansion terms are gained through section 3.1 to 3.4. In our experiment, the Indri search engine [5] is used for the query likelihood (QL) model with Dirichlet prior smoothing as our baseline.

The final score is a combination of scores received from the original query and expanded query. The expanded query consists of disease, symptom, test, and treatment terms. The weight of disease terms is assigned as 1.

The parameters are learned by using TREC 2014 CDS collection:

- the weight of the original query (λ): {0.1, 0.2, **0.3**, …, 0.9}
- # of expansion terms (*e*=e1+e2+e3+e4): {20, 25, 30, 35, 40, 45, **50**…, 150}
- # of disease terms (*e1*): {5, 10, 15, **20**, 25, …, 50}
- # of symptom terms (e2): {5, 10, **15**, …, 50}
- # of test terms (e3): {**5**, 10, 15, 20}
- # of treatment terms (e4) : {5, **10**, 15, 20, 25, 30}
- The weight of symptom terms (α): {0.1, 0.2, …, **0.7**, 0.8 0.9}
- The weight of test terms (β): {0.1, 0.2, …, **0.5**, ..., 0.9}
- The weight of treatment terms (ɣ): {0.1, 0.2, …, **0.7**, 0.8 0.9}

Bold character indicates final parameter.

4. EXPERIMENTS

4.1 Run description

In our experiments for Task A, the query type information is not used for a query expansion. For Task B, the different kinds of expansion terms are selected according to the query types.

Our methods for task A are described as follows:

- **cbnu0:** Baseline Language model
- **cbnu1:** query expansion based on the clinical semantic knowledge (section 3.1)
- **cbnu2:** query expansion based on the clinical semantic knowledge and LDA (section 3.1 and 3.2)

Our methods for task B are described as follows:

- **cbnu0'**: Baseline Language model with terms in TREC Diagnosis field
- **cbnu3:** query expansion based on the clinical semantic knowledge and LDA according to the query types (section 3.1, 3.2, and 3.4)
- **cbnu4:** query expansion using medical terms and synonym disease terms based on the clinical semantic knowledge according to the query types(section 3.1, 3.3, and 3.4)
- **cbnu5**: query expansion using medical terms and synonym disease terms based on the clinical semantic knowledge and LDA according to query types(section 3.1, 3.2, 3.3 and 3.4)

4.2 Experimental Results

The experimental results for Task A are shown in Table 4:

Table 4. Experimental results for task A

RunID	infNDCG	R-prec	P@10
cbnu0	0.2179	0.1863	0.3733
cbnu1	0.2317	**0.1954**	0.3833
cbnu2	**0.2327**	0.1913	0.3967

The experimental results for Task B are shown in Table 5:

Table 5. Experimental results for task B

RunID	infNDCG	R-prec	P@10
cbnu0'	0.2779	0.2265	0.4600
cbnu3	0.2260	0.1836	0.4000
cbnu4	0.2917	0.2456	0.4900

cbnu5	0.3033	0.2503	0.4800

5. CONCLUSIONS

Combining the clinical semantic knowledge extracted from UMLS and Wikipedia with a topic model is effective for TREC Clinical Decision Support Track. In our experiments for Task A, the query type information is not used for a query expansion. For Task B, the different kinds of expansion terms are selected according to the query types. The proposed methods achieved 0.2327 and 0.3033 in the inferred NDCG on Task A and Task B, respectively.

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-2015-R0992-15-1023) supervised by the IITP(Institute for Information & communications Technology Promotion).

7. REFERENCES

[1] http://trec-cds.appspot.com/2015.html

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

[3] Olivier Bodenreider. “The Unified Medical Language System(UMLS): intergrating biomedical terminology”. Nucleic Acids Res. 2004;32:D267–D270.

[4] David M. Blei, Andrew Y. Ng, Michael I. Jordan. "Latent dirichlet allocation". The Journal of Machine Learning Research. Volume 3, 3/1/2003. Pages 993-1022.

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

[6] Lu Liu, Jie Tang, Yu Cheng, Ankit Agrawal, Wei-keng Liao, Alok Choudhary. "Mining diabetes complication and treatment patterns for clinical decision support". CIKM '13 Proceedings of the 22nd ACM international conference on Conference on information & knowledge management. Pages 279-288.

[7] Corey Arnold, William Speier. "A topic model of clinical reports". SIGIR '12 Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval. Pages 1031-1032.

[8] Isabelle Stanton, Samuel Ieong, Nina Mishra. "Circumlocution in diagnostic medical queries". SIGIR '14 Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval. Pages 133-142.

[9] Ryen W. White, Eric Horvitz. "Studies of the onset and persistence of medical concerns in search logs". SIGIR '12 Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval. Pages 265-274.