Bridging Layperson's Queries with Medical Concepts-GRIUM@CLEF2015 eHealth Task 2

Xiao Jie Liu, Jian-Yun Nie

RALI, DIRO, University of Montreal
C.P.6128, succursale Centre-ville, Montreal, Quebec Canada H3C 3J7

{xiaojiex, nie}@iro.umontreal.ca

Abstract.

Concepts are often used in Medical Information Retrieval. In any concept-based method one has to extract concepts from texts (query or document). MetaMap is often used for this task. However, if the query is issued by a layperson, the query may not contain the appropriate concept expressions and MetaMap will fail to extract correct concepts. In this situation we need to explore other resources to help extract concepts from such query. In our participation in CLEF2015 eHealth task 2, we investigated the utilization of two resources - UMLS Concept Definition and Wikipedia Articles - to bridge user queries and the intended medical concepts. Our results suggest that such resources could be useful to some extent. In this report, we describe the methods tested as well as their results.

Keywords: UMLS, Wikipedia, MetaMap, Language Model, Query Expansion

1 Introduction

The CLEF2015 eHealth Task 2 aims to evaluate the effectiveness of information retrieval systems when searching for healthcare information on the web. This task is a continuation of the previous CLEF eHealth Task 3 run in 2013 and 2014. The document collection is the same as CLEF2014, but this year the queries look more like queries formulated by laypeople (i.e. non medical experts) who attempt to find out more about the condition they may have from signs, symptoms, etc. For example, when confronted with signs of jaundice, non experts may use queries like "white part of eye turned green" to search for information. Different from the previous CLEF eHealth experiments, these queries use general language wording in place of the accurate specialized expressions to refer to a condition or disease.

In CLEF2014, our team took part in task 3a and obtained the best result [Shen et al. 2014]. The method was based on concept expressions extracted from a query: once a concept is identified with MetaMap, we use all the synonym expressions stored in UMLS Metathesaurus to expand the original query. Each concept expression was used as a phrase, which was matched in strict order or at proximity without order. This strategy turned out to be more appropriate than matching the concept IDs.

For the task of this year, our method is built upon the method of last year. The basic approach is the same. However, as the queries are formulated in a different way, we focus on the problem of identifying appropriate concepts from user queries. As MetaMap relies on concept expressions stored in UMLS Metathesaurus (or another thesaurus) to recognize concepts, it is unlikely that appropriate underlying concepts could be extracted from one of this year's queries. In order to extract concepts from a user query, we leverage external resources that may possibly bridge a layperson's expression with medical concepts. In our experiments we use UMLS Concept Definition and Wikipedia articles to help to do that. In both cases, we hope that the definitions of the concepts and the descriptions in Wikipedia texts correspond better to user's query formulations. Our assumption is that both concept definitions and Wikipedia texts try to explain the medical concepts, and the explanation target less specialized people. Therefore, these definitions and descriptions could use a language more similar to layperson's queries. By matching the definitions and Wikipedia texts with a user's query, it is hopeful that we could identify the underlying medical concepts. In our experiments, in addition to the same basic strategy used in our CLEF 2014 experiments, we try to incorporate the concepts identified in this way and test whether this approach can effectively help solve the vocabulary mismatch problem.

This report is organized as follows. In section 2, we will describe the method used to bridge a user's query and medical concepts. In section 3, we describe the retrieval methods used in our participation. In section 4, we report the experimental results. Preliminary conclusions will be drawn in section 5.

2 Bridging a User Query and Medical Concepts

As we stated, a key problem with a layperson's query is that the underlying medical concepts are not expressed using its expressions stored in a medical thesaurus (e.g. UMLS Metathesaurus), because the user may not know what the right concept expression is. Instead, the user may use common language to describe indirectly the concepts. It may often be the case that the underlying concept could not be identified from such a description.

To identify concepts from a user's query, we believe that some external resources are required. These external resources should play the role of describing medical concepts for laypersons. It is expected that the descriptions in these resources may better correspond to user's query formulations.

In our experiments in CLEF eHealth task 2, we use two such resources: UMLS Concept Definition and Wikipedia articles.

In UMLS, in addition to concepts themselves, we also have definitions for many concepts. For example, the concept is "Alopecia Areata" (C0002171) and its definition is "Loss of scalp and body hair involving microscopically inflammatory patchy areas". In this year task, there is one training query: "loss of hair on scalp in an inch width round". So this concept definition looks more like a user query. In this example, we can observe that the definition uses less specialized vocabulary, which may better correspond to the queries formulated by a layperson. Therefore, we use these definitions as an intermediary to find the corresponding concepts. For example, if a user query contains words similar to those used in a definition, then we assume that the corresponding concept is the one that the user wants to express. This is implemented as follows: the original user query is used to match definitions, which are indexed using an IR system (Indri). We select the concepts of the top 5 definitions retrieved as the most likely concepts underlying the user's query. These concepts are used to expand the original query.

Unfortunately, not all the concepts in UMLS have a definition. Only 213,844 of about 3 million concepts in UMLS Metathesaurus have definitions. So the above approach will likely suffer from poor coverage – the intended concepts of a user query may not have a definition, thus we will fail to find the concept through definition.

As an attempt to increase coverage, we leverage Wikipedia. Wikipedia contains many articles describing medical concepts. In most cases, the article's title corresponds to a medical concept, and the article's text explains the concept. For example, one Wikipedia article title is "Wikipedia: Abadie's sign of exophthalmic goiter" and its article text is "Abadie's sign of exophthalmic goiter is Spasm of the Levator Palpebrae Superioris muscle with retraction of the upper lid (so that sclera is visible above cornea) seen in Graves-Basedow disease which, together with exophthalmos causes the bulging eyes appearance". The title is about the concept exophthalmos (C0015300). The text is the description of this concept. Again, we assume that a layperson's query may better correspond to a Wikipedia text. In that case, the article's title is used to identify the corresponding concept (through MetaMap). In our experiment, we only use abstracts of Wikipedia articles, which are more concise and may contain less noise (words that are less directly related to the concept). Our implemen-

tation is similar to the use of UMLS concept definition: a user query is used to find a few Wikipedia articles, and the corresponding titles are submitted to MetaMap to identify concepts. In order to select only Wikipedia articles related to medical area, we use Wikipedia categories.

3 Retrieval Methods

In this section, we describe the methods we used in our participation.

3.1 Baseline

As baseline, we use a traditional approach based on language modeling, with Dirichlet smoothing. In this method the score of a document D given a query Q is determined as follows:

$$S(Q,D) = \frac{1}{n}\sum_{i=1}^{n} \log P(q_i|D) \quad \textbf{(1)}$$

where Q is the query, D is the document, n is the length of query and $P(q_i|D)$ is the probability of document language model to create query term q_i, which is adjusted by Dirichlet smoothing below:

$$P(q_i|D) = \frac{tf_{q_i,D} + \mu \frac{tf_{q_i,C}}{|C|}}{|D| + \mu} \quad \textbf{(2)}$$

where tf is the term frequency, q_i is query term in query Q, D is the document, C is the whole collection, |C| is its size and μ is the smoothing parameter, which is set at 2000. This method is named bag-of-words (BOW) method in this report.

3.2 Query Expansion

In this report, we use query expansion to be another basic method besides the baseline method. Query expansion is generally implemented with the following formula:

$$S(Q|D) = \lambda S(Q, D) + (1-\lambda) S(Q', D) \quad \textbf{(3)}$$

where Q is the original query, Q' is the query expansion composed of terms (concepts) related to Q, λ is an interpolation parameter. In our experiment, we exact concepts from the original query and then use all the synonym expressions of these concepts to expand the query in different ways [Shen et al. 2014]: (1) use the exact concept phrase matching – a concept expression is matched in exactly the same form (#1 operator in Indri); (2) proximity concept phrase matching (#uwN operator in Indri, where N is the size of the phrase +1); (3) matching the bag of words in concept ex-

pressions (i.e. the words in concept expressions form a bag of words). This method is the same as the one we used in last year task. So the score S(Q|D) is defined as follows:

$$S(Q|D) = \lambda_1 S_{BOW} + \lambda_2 S_{Exact\text{-}Phrase} + \lambda_3 S_{Prox\text{-}Phrase} + \lambda_4 S_{BOW\text{-}Concepts} \quad \textbf{(4)}$$

where S_{BOW} is the score from BOW method with the original query, $S_{Exact\text{-}Phrase}$ is the score from exact concept phrase matching, $S_{Prox\text{-}Phrase}$ is the score from proximity concept phrase matching, $S_{BOW\text{-}concepts}$ is the score from the bag-of-words of concept expressions matching. λ_1, λ_2, λ_3 and λ_4 are the parameters and $\lambda_1 + \lambda_2 + \lambda_3 + \lambda_4 = 1$. These matching strategies are summarized in the following table.

Table 1. Matching Strategy and Expression in Indri

Matching Strategy	**Expression in Indri**
Exact-Phrase matching	#1()
Prox-Phrase matching	#uwN()
BOW-Concepts matching	#combine()

We tested three ways to identify concepts from a user query:
(1) Using MetaMap directly (the same method as last year);
(2) Using UMLS Concept Definition;
(3) Using Wikipedia Articles.

3.3 Using MetaMap

In this method, we use MetaMap to extract concepts from queries directly. Because the queries often do not contain specialized terms, MetaMap fails to identify the underlying concepts from them. In our preliminary tests with the training queries, we found that the concepts identified by MetaMap did not improve retrieval effectiveness over the basic BOW method. This observation was quite different from the results of last year on more specialized queries: we found that the concepts identified by MetaMap could improve retrieval effectiveness last year. Therefore, we did not submit this run this year.

3.4 Using UMLS Concept Definitions and Wikipedia Texts

Instead of using MetaMap to extract concepts from the original queries, we use UMLS concept definitions and Wikipedia texts to do it. The top 5 concepts and Wikipedia titles are identified separately. This method is described in Section 2.

In addition, intuitively UMLS is the specific medical ontology and it should have better precision, but lower recall. On the contrary, Wikipedia articles are not only about the medical field and it should have better recall, but lower precision. So we try

to combine these two kinds of resources. In our experiment, from the training queries we observe that exact concept phrase matching in the method which uses UMLS concept definitions contributes much, but BOW-concepts matching in the method of using Wikipedia articles contributes much. So in our combination of these two resources, in formula 4, besides the instantiation of S_{BOW} used, the instantiation of $S_{Exact\text{-}Phrase}$ used is from the UMLS concept definition method, and the instantiation of $S_{BOW\text{-}Concepts}$ used is from the Wikipedia articles method. We will ignore $S_{Prox\text{-}Phrase}$ in this combination.

4 Experiments

The document collection for CLEF2015 task 2 consists of a set of documents in the medical domain, provided by the Khresmoi project. Each document contains four fields: #uid, #date, #url and #content. We convert the collection into TREC style format. In #content part, we eliminate all commend, css and JavaScript part and all HTML tags. Only the pure text contents are left.

There are in total 66 queries in this year's task. Below is one query example.

CLEF2015 Query Example:

<query>
<number>clef20115.test.60</number>
<text>baby white dot in iris</text>
</query>

We use Indri as the basic experimental platform for all the methods. Usual processings are used: Porter stemming and removal of stopwords (Lemur stoplist). The following 7 methods (runs) have been submitted:

Table 2. Experiments setup for our 7 submitted runs

Run	Experiment Method	Submission
1	Baseline (language model with Dirichlet smoothing, μ=2000)	GRIUM_EN_Run1
2	Combination of UMLS concept definition and Wikipedia articles λ_1=0.95, λ_2=0.025, λ_3=0, λ_4=0.025	GRIUM_EN_Run2
3	Using UMLS concept definition λ_1=0.98, λ_2=0.02, λ_3=0, λ_4=0	GRIUM_EN_Run3
4	Using Wikipedia articles λ_1=0.8, λ_2=0, λ_3=0, λ_4=0.2	GRIUM_EN_Run4
5	Using Run2 with different parameters λ_1=0.9, λ_2=0.05, λ_3=0, λ_4=0.05	GRIUM_EN_Run5

6	Using Run3 with different parameters λ_1=0.95, λ_2=0.05,λ_3=0, λ_4=0	GRIUM_EN_Run6
7	Using Run4 with different parameters λ_1=0.95, λ_2=0, λ_3=0, λ_4=0.05	GRIUM_EN_Run7

For evaluation, P@10 and NDCG@10 are the main performance indicators for this year task. In this year, in addition to (topical) relevance assessments (qrels), we also have readability assessments (qread). We use RBP (0.8) to indicate readability performance. The table below summarizes the results of our 7 runs.

Table 3. Results of our 7 runs

Submission	**p@10**	**NDCG@10**	**RBP (0.8)**
GRIUM_EN_Run1	0.3136	0.2875	0.3249
GRIUM_EN_Run2	0.3091	0.2850	0.3305
GRIUM_EN_Run3	0.3167	0.2887	0.3296
GRIUM_EN_Run4	0.3030	0.2853	0.3244
GRIUM_EN_Run5	0.3045	0.2841	0.3278
GRIUM_EN_Run6	0.3182	0.2868	0.3306
GRIUM_EN_Run7	0.3061	0.2803	0.3272

It can be observed that P@10 and NDCG@10 for all the 7 runs are quite similar. We show one plot of P@10 with the run GRIUM_EN_Run3 below.

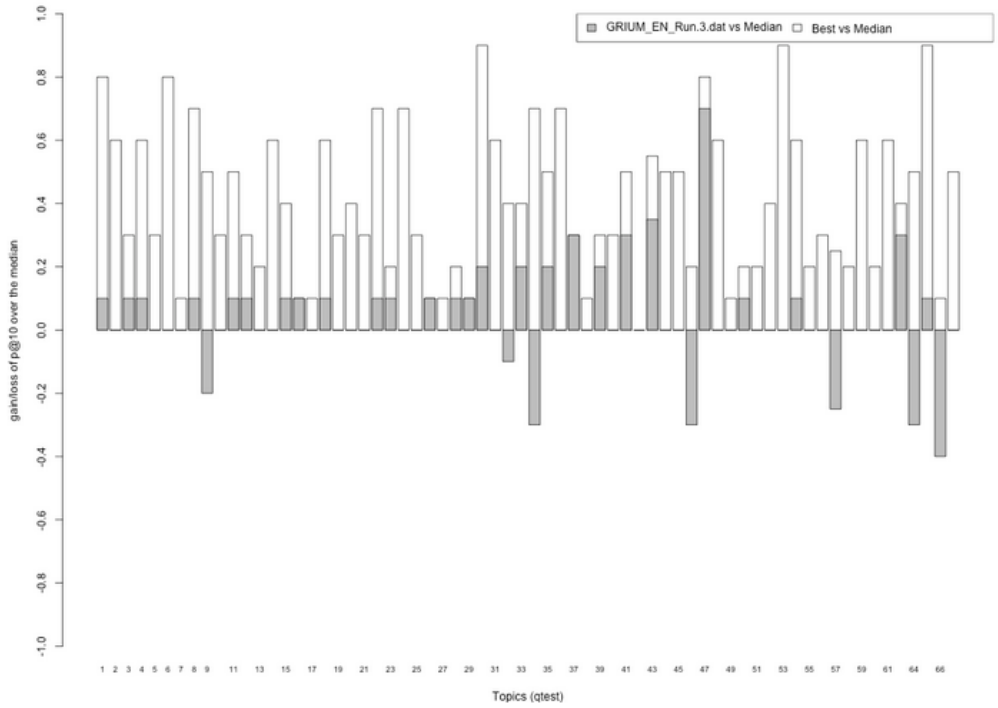

Fig. 1. GRIUM_EN_Run3 VS. Median run VS. Best run at P@10. White bar is best run; grey bar is our run GRIUM_EN_Run3.

For GRIUM_EN_Run3, among the total 66 queries, 7 queries have worse performances than the median; 4 queries have best performance; the other 55 queries have performances equal to or higher than median.

Among all the submitted runs, GRIUM_EN_Run3 produced higher scores on all the measures. With a slightly different parameter setting in GRIUM_EN_Run6, we also observe that P@10 and RBP are the highest. In these two runs, we only used UMLS concept definitions. Despite the small scale of improvements, this result may indicate that concept definitions could help finding the underlying concepts from a layperson's query.

On the other hand, when Wikipedia articles are used, we do not observe improved results. This may indicate that Wikipedia could not be an appropriate resource to bridge user's queries and medical concepts.

All the submitted runs have been produced with parameters that are set without many training queries. It is possible that with different settings, we could obtain different results. This is what we will test in the future.

5 Conclusion

In this year CLEF eHealth task 2, we focus on the problem of bridging user's queries with medical concepts. We use MetaMap, UMLS concept definition and Wikipedia articles. We submit 7 runs. From the results, we observe that our runs are generally better than the median. In particular, when UMLS concept definitions are used, we observe slight improvements over the baseline. This may indicate that such a method could be useful to identify the underlying medical concepts of a layperson's query. The experimental results have been produced without fine-tuning the parameters. In the future, we will further investigate the impact of each resource.

Reference:

1. Palotti, Joao and Zuccon, Guido and Goeuriot, Lorraine and Kelly, Liadh and Hanbury, Allan and Jones, Gareth JF, and Lupu, Mihai and Pecina, Pavel. CLEF eHealth Evaluation Lab 2015, task 2: Retrieving Information about Medical Symptoms. 2015.
2. Lorraine Goeuriot, Liadh Kelly, Hanna Suominen, Leif Hanlen, Aurélie Névéol, Cyril Grouin, Joao Palotti, Guido Zuccon. Overview of the CLEF eHealth Evaluation Lab 2015. 2015.
3. Wei Shen, Jian-Yun Nie, Xiao Hua Liu, Xiao Jie Liu. An Investigation of the Effectiveness of Concept-based approach in Medical Information Retrieval GRIUM@CLEF2014HealthTask3. 2014.
4. Zuccon, Guido and Koopman, Bevan and Palotti, Joao. Diagnose This If You Can: On the effectiveness of search engines in finding medical self-diagnosis information. 2015.
5. Stanton, Isabelle and Ieong, Samuel and Mishra, Nina. Circumlocution in diagnostic medical queries.2014.
6. Zuccon, Guido and Koopman, Bevan. Integrating Understandability in the Evaluation of Consumer Health Search Engines.2014.
7. Park, Laurence AF and Zhang, Yuye. On the distribution of user persistence for rank-biased precision.2007.
8. Goeuriot, Lorraine and Kelly, Liadh and Li, Wei and Palotti, Joao and Pecina, Pavel and Zuccon, Guido and Hanbury, Allan and Jones, Gareth JF and Mueller, Henning. Share/clef ehealth evaluation lab 2014, task 3: User-centred health information retrieval.2014.