UESTC at ImageCLEF 2011 Medical Retrieval Task

Hong Wu, Chengbo Tian,

School of Computer Science and Engineering,
University of Electronic Science and Technology of China,
611731 Chengdu, P. R. China
hwu@uestc.edu.cn, tianchengbo@gmail.com

Abstract. This paper describes methods and results archived by our research group at the ImageCLEF 2011 medical retrieval task. We performed two sub-tasks, ad-hoc retrieval and case-based retrieval, and only used text information for retrieval. In our work, a phrase-based retrieval model was adopted, and UMLS metathasaurus was used to expand query. The phrase-based model was implemented based on Indri search engine and their structured query language. For query expansion, the detected concepts and their direct children were used to append the structured query. Both phrases and medical concepts were identified with the help of the MetaMap program. The parameters of our approach were trained on the data of ImageCLEFmed 2010 ad-hoc retrieval subtask.

Keywords: Medical Retrieval, Phrase-Based Model, Indri, Query Expansion, MetaMap, UMLS

1 Introduction

This paper describes the second participation of the UESTC group at the ImageCLEF medical retrieval task. In previous years, we tested a phrase-based approach for medical retrieval. Phrases and subphrases were extracted with the help of MetaMap, and the individual words in each phrase or subphrase were concatenated and used as an indexing term in vector space model, together with single word terms. In this year, we adopted a more principled and efficient phrase-based retrieval model, which was implemented based on Indri search engine [1] and their structured query language. For query expansion, the concepts detected from the original text query and their direct children were used to append the structured query. Our approach had gotten promising results on the data of ImageCLEFmed 2010 ad-hoc retrieval subtask [2].

ImageCLEFmed 2011 [3] includes three types of tasks, ad-hoc retrieval, case-based retrieval and modality classification. For the retrieval tasks, the dataset contains 230,088 images from more than 55,000 articles published in online medical journals. In the ad-hoc retrieval task, a set of 30 textual queries, each of which with several sample images, are given, and the goal is to retrieve the images most relevant to each topic. In the case-based task, a set of 10 case-based information requests are given, and the goal is to retrieve the articles most relevant to the topic case.

The remainder of this paper is organized as follows. Our approach is described in section 2. And our submitted runs and results are presented in section 3, followed by the conclusions in section 4.

2 Phrase-Based Retrieval Model and Query Expansion

To utilize phrase in information retrieval, there're three steps: (1) Identify phrases in query text, (2) Identify these phrases in document, (3) Combine phrase with individual word in ranking function. In this work, the first step was performed with the help of MetaMap, and the last two steps were implemented based on Indri search engine [1] and its structured query language. Indri is a scalable search engine that inherits the inference net framework from InQuery and combines it with language modeling approach to retrieval.

2.1 Phrase Identification

In our approach, phrase identification was conducted with the help of MetaMap [4] which is a tool to map biomedical text to concepts in the UMLS Metathesaurus [5]. MetaMap first parses the text into phrases, and then performs intensive variant generation on each phrase. After that, candidates are retrieved from the Metathesaurus to match the variants. Finally, the candidates are evaluated by a mapping algorithm, and the best candidates are returned as the mapped concepts.

In this work, the concept mapping was restricted to three source vocabularies: MeSH, SNOMED-CT, and FMA. And two phrase identification strategies were explored. One is to use the phrases produced by the early step in MetaMap program, and filter out the unwanted words in them, such as preposition, determiner etc. Another is to consider the individual word or sequence words which mapped to UMLS concept as a phrase.

2.2 Phrase Representation

After identifying phrase in query text, the query phrases would be recognized again within documents. Many phrase identification techniques only look at contiguous sequences of words. But, the constituent words of a query phrase might be several words apart, and even with different order when used within a document. Fortunately, this problem can be easily solved by using operators in indri query language. The query can be reformulated with the special operators to provide more exact information about the relationship of terms in the original text query. Here, we introduce some operators which are commonly used for representing phrase.

- Ordered Window Operator : #N(T1...Tn) or #odN(T1...Tn)

The terms within an ordered window operator must appear ordered with at most *N*-1 terms between adjacent terms in the document in order to contribute to the document's belief score.

- Unordered Window Operator: #uwN(T1 ... Tn)

The terms contained in an unordered window operator must be found in any order within a window of *N* words in order to contribute to the belief score of the document. For example, the phrase "congestive heart failure" can be represented as #1(congestive heart failure), which means that the phrase is recognized in document only if the three constituent words are found in the right order and no other words between them. It can also be represented as #uw6(congestive heart failure), which means that the phrase is recognized in document only if the three constituent words are found in any order within a window of 6 words.

There are also some other operators related to our works. They are introduced as follow.

- Combine Operator: #combine (T1 ... Tn)

The terms or nodes contained in the combine operator are treated as having equal influence on the final result. The belief scores provided by the arguments of the combine operator are averaged to produce the belief score of the #combine node.

- Weight Operator: #weight (W1 T1 ... Wn Tn)

The terms or nodes contained in the weight operator contribute unequally to the final result according to the weight associated with each (W*i*). The belief scores provided by the arguments of the weight operator are weighted averaged to produce the belief score of the #weight node. Taking #weight(1.0 dog 0.5 train) for example, its belief score is 0.67 log(b(dog)) + 0.33 log(b(train)).

2.3 Phrase-Based Retrieval Model

When utilizing phrase in information retrieval, phrasal term should be combined with word term in ranking function. Simply, we can use one ranking function for word term, another for phrasal term, and the final ranking function is weight sum of them.

$$F(Q,D) = w_1 f_1(Q,D) + w_2 f_2(Q,D) \quad (1)$$

where Q and D stand for query and document respectively. $f_1(Q,D)$ is the ranking function for word term, and $f_2(Q,D)$ is the ranking function for phrasal term. The weights w_1 and w_2 can be tuned by experiment.

With Indri search engine, this ranking function can be implemented easily with a structured query and the inference network model. For example, the topic 8 in ad-hoc track of imageCLEFmed 2011 is:

"x-ray images of a hip joint with prosthesis"

With the second phrase identification strategy, the phrases are "x-ray images", "hip joint" and "prosthesis". The query can be formulated as following,

#weight(0.9 #combine(x ray images of a hip joint with prosthesis) 0.1 #combine(#uw8(x ray images) #uw8(hip joint) prosthesis))

The first #combine() in the structured query corresponds to $f_1(Q,D)$, and the second one corresponds to $f_2(Q,D)$, while $w_1$is 0.9 and w_2 is 0.1.

2.4 Thesaurus-Assistant Query Expansion

After the original query text is mapped to concepts in UMLS, the query can be expanded with the mapped concept terms, their synonym, hierarchical or related term information. Since the query terms of ordinary users tend to be general, we used the preferred names of the mapped concepts and their direct children to expand query. The adding terms are not necessarily important as the original ones, so weight can be introduced. And the new ranking function is,

$$F(Q,D) = w_1 f_1(Q,D) + w_2 f_2(Q,D) + w_3 f_3(Q,D) \quad (2)$$

and $f_3(Q,D)$ is the ranking function for the concepts, and w_3 is the corresponding weight. The concepts can be represented in the same way as the phrases.

To expand query with the mapped concept and their direct children, the preferred names of these concepts are normalized and redundant names were erased. Taking the topic 8 for example, the query expanded only with concepts can be as following.

#weight(0.7 #combine(x-ray images of a hip joint with prosthesis) 0.1 #combine(#uw8(x-ray images) #uw8(hip joint) prosthesis) 0.2 #combine(radiography #uw12(entire hip joint) prosthesis))

3 Experiments and Results

The parameters w_i of our approach were trained on the data of ImageCLEFmed 2010 ad-hoc retrieval subtask to maximize MAP with the constraint that sum of w_i equals to one. And the optimal parameters were used for both ad-hoc retrieval and case-based retrieval.

3.1 Ad-hoc Retrieval

For ad-hoc retrieval, the caption of each image was used as document representation. We tested tow phrase identification strategies. The first is represented as “p1” which using the filtered phrase identified in the early step of MetaMap program, and the second is represented as “p2” which using the word sequence as phrase which corresponding to the mapped concept. We used #uwN() for representing phrase and concept in the structured query, and chose $N=k*n$, where n is the number of terms within the operator, and k is a free parameter. We tested two settings of k, 2 and 4, and denoted $k=2$ as *sw*, which standing for small window. Table 1 shows the list of all 10 runs for ad-hoc retrieval. Table 2 shows the results of all runs. The results indicate that the phrase-based retrieval model can improve the retrieval performance, and query expansion can retrieve more relevant images but get lower MAP. After all, the results were not competitive, the best of our runs only ranked 30th among all 64 automatic text runs.

Table 1. Descriptions of ad-hoc retrieval experiments

Runid	Description
UESTC_adhoc_indri	Baseline, directly use indri search engine
UESTC_adhoc_p1	with p1 phrase identification strategy
UESTC_adhoc_p2	with p2 phrase identification strategy
UESTC_adhoc_p1QE	p1 + query expansion with concepts
UESTC_adhoc_p2QE	p2 + query expansion with concepts
UESTC_adhoc_p1_sw	p1 (k = 2)
UESTC_adhoc_p1QE_sw	p1 + query expansion with concepts (k = 2)
UESTC_adhoc_p2QE_sw	p2 + query expansion with concepts (k = 2)
UESTC_adhoc_p1QE_sw_chd	p1 + query expansion with concepts and children (k = 2)
UESTC_adhoc_p2QE_sw_chd	p2 + query expansion with concepts and children (k = 2)

Table 2. Retrieval performance of ad-hoc retrieval runs

Runid	MAP	P10	RelRet
UESTC_adhoc_p1	**0.1672**	0.2667	1373
UESTC_adhoc_p2	**0.1672**	0.2733	1348
UESTC_adhoc_p1QE_sw	0.1669	**0.2833**	1384
UESTC_adhoc_p2QE_sw_chd	0.1666	0.2700	1362
UESTC_adhoc_p1_sw	0.1635	0.2733	1368
UESTC_adhoc_p1QE	0.1632	0.2533	**1385**
UESTC_adhoc_p2QE_sw	0.1590	0.2567	1286
UESTC_adhoc_indri	0.1588	0.2600	1377
UESTC_adhoc_p2QE	0.1583	0.2500	1350
UESTC_adhoc_p1QE_sw_chd	0.1471	0.2100	1346

3.2 Case-Based Retrieval

For case-based retrieval, we investigated two different document representations, the first representation is named *full* which contains the full text, title and mesh terms of an article, and the second one is called *ac* which contains the abstract, all image captions, title and mesh terms of an article. We only experimented with p2 phrase identification strategy and set k=4. Besides our phrase-based approach, we also tested okapi retrieval model, and a pseudo relevance feedback method which add the abstracts and mesh terms of the first two returned article to the original query. Table 3 shows the list of all 9 runs for case-based retrieval. Table 4 shows the results of all runs for case-based retrieval. Our approach did not make success for case-based

retrieval, but the *full* document representation with indri search engine got the top rank among all submissions in automatic text runs.

Table 3**.** Descriptions of case-based retrieval experiments

Runid	Description
UESTC_full_indri	*full* doc representation
UESTC_full_p2	*full* doc representation + p2
UESTC_full_p2QE	*full* doc representation + p2 + query expansion with concepts
UESTC_ac_indri	*ac* doc representation
UESTC_ac_p2	*ac* doc representation + p2
UESTC_ac_p2QE	*ac* doc representation + p2 + query expansion with concepts
UESTC_full_okapi	*full* doc representation + okapi
UESTC_full_okapi_fb	*full* doc representation + okapi + pseudo relevance feedback
UESTC_ac_okapi	*ac* doc representation + okapi
UESTC_ac_okapi_fb	*ac* doc representation + okapi + pseudo relevance feedback

Table 4**.** Retrieval performance of case-based retrieval runs

Runid	MAP	P10	RelRet
UESTC_full_indri	**0.1297**	**0.1889**	144
UESTC_full_p2QE	0.1199	0.1556	143
UESTC_full_p2	0.1179	**0.1889**	145
UESTC_full_okapi	0.0907	0.1444	**148**
UESTC_ac_okapi	0.0835	0.1222	128
UESTC_ac_indri	0.0767	0.1111	127
UESTC_full_okapi_fb	0.0762	0.1333	113
UESTC_ac_p2	0.0722	0.1222	128
UESTC_ac_p2QE	0.0677	0.1000	127
UESTC_ac_okapi_fb	0.0500	0.0778	103

4 Conclusions

This paper describes our contribution to the ImageCLEF 2011 medical retrieval task. We adopted a phrase-based retrieval model, and an UMLS-based query expansion. For ad-hoc retrieval, we submitted 10 runs. The results were not competitive, but indicate that phrase-based model and query expansion can improve the retrieval performance. For ad-hoc retrieval, we submitted 9 runs. Our approach did not make

success for case-based retrieval, but the full document represent with indri search engine got the top rank among all automatic text runs. We conjecture that the unsatisfactory results of our approach at ImageCLEFmed 2011 are because the weight parameters were trained on data of ImageCLEFmed 2010's ad-hoc retrieval subtask, and the two datasets are quite different.

Acknowledgments. This research is partly supported by the National Science Foundation of China under grants 60873185 and the Open Project Program of the National Laboratory of Pattern Recognition (NLPR).

Reference

1. Strohman, T., Metzler, D., Turtle, H., Croft, W. B.: Indri: A language model-based search engine for complex queries. In: Proceedings of the International Conference on Intelligence Analysis (2005)
2. Wu, H., Tian C.B.: Thesaurus-Assistant Query Expansion for Context-based Medical Image Retrieval, submitted to Pacific-Rim Conference on Multimedia (2011)
3. Kalpathy-Cramer, J., Müller, H., Bedrick, S., Eggel, I., Garcia Seco de Herrera, A., Tsikrika, T.: The CLEF 2011 medical image retrieval and classification tasks. In: CLEF 2011 working notes (2011)
4. Aronson, A.R.: Effective Mapping of Biomedical text to the UMLS Metathesaurus: the MetaMap Program. In: Proceedings of the AMIA Symposium (2001)
5. Bodenreider, O.: The Unified Medical Language System (UMLS): integrating biomedical terminology. In: Nucleic Acids Research 32, pp. 267--270 (2004)