UEvora at CLEF eHealth 2017 Task 3

Hua Yang and Teresa Gonçalves

Computer Science Department, University of Évora
Évora, Portugal
huayangchn@gmail.com,tcg@uevora.pt

Abstract. This paper describes the methods we used for our participation to CLEF eHealth 2017 Task 3 IRTask 1: ad-hoc search. This task aims at retrieving information relevant to people seeking health advice on the web. We present our work of using query reformulation techniques in this paper. We use cTAKES, a clinical natural processing system, to identify UMLS concepts in the original query. Query expansion techniques are then applied to the identified medical concepts. Query expansion based on UMLS meta-thesaurus or a Word2vec model trained with domain data is used in our work. We also use other techniques, like increasing the weight of the terms that are considered to catch the users' need much more compared to other terms.

Keywords: UMLS; word2vec; query reformulation; query expansion; cTAKES;

1 Introduction

CLEF eHealth 2017 information retrieval (IR) tasks 3 is a continuation of the previous tasks that ran in 2013, 2014, 2015, and 2016 and embraces the TREC-style evaluation process, with a shared collection of documents and queries, the contribution of runs from participants and the subsequent formation of relevance assessments and evaluation of the participants submissions [5, 6].

CLEF eHealth 2017 Task 3 includes four sub tasks this year. Our team participates in Task 3 IRTask 1, which is a standard ad-hoc search task, aiming at retrieving information relevant information to people seeking health advice on the web.

Data corpus. ClueWeb12-B13 is used as the corpus of the CLEF eHealth 2017 Task 3. We use the indexes provided by the organizers from Microsoft Azure, which are available with Terrier and Indir formats.

Queries. All the queries used in the task are extracted from public health web forums where users were seeking advice about specific symptoms, diagnosis, conditions or treatments [6]. The queries are considered as the real health information needs expressed by the general public. For each forum post a set of 6 query variants are generated, representing different ways to express the same information need. A total of 300 queries are created for the task.

Evaluation. Evaluation measures for IRTask1 are NDCG@10, BPref and RBP.

The rest of this paper is organized as follows. The methods we used for participating in the task are presented in section 2. The experiments and submission runs are described in section 3.

2 Methods

In this work, we use query reformulation techniques to reform the original queries. Figure 1 illustrates the framework we used.

We first use natural language processing tools to identify the medical concepts in the original query. For the identified medical concept, we then use query expansion techniques to find its related terms or synonyms with the same concept. The expanded queries are then issued to the retrieval platform. With an weighting model in the IR platform, a ranked list of documents is returned.

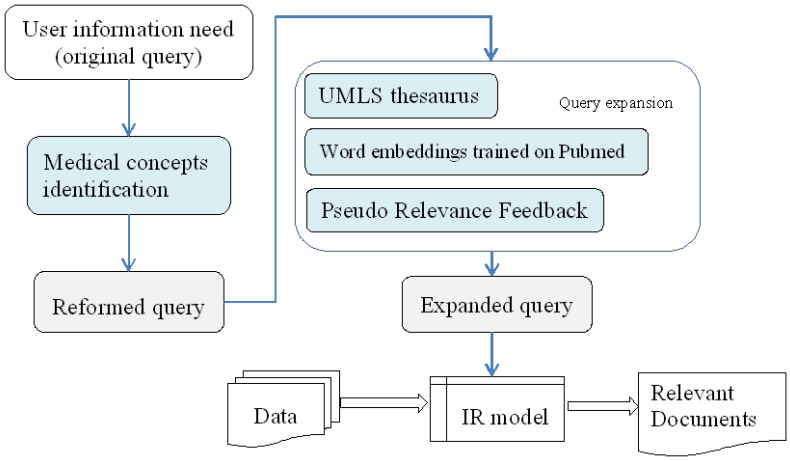

Fig. 1. Framework of query reformulation

2.1 Medical concepts identification

Complex semantic relationships exists in health articles, like term dependency and vocabulary mismatch [1]. Natural language processing tools applied in clinical area can extracts concepts from free text and normalises them with respect to a gold standard ontology to alleviate issues of vocabulary mismatch [2].

In our system, we use clinical NLP tool to identify the medical concepts in the original queries. For medical concepts which are identified, we regard them as important information reflecting the users' needs. We increase the weight of these terms or phrases. We denote them as reformed query in our system.

2.2 Query expansion

UMLS metathesaurus or word2vec models trained with domain data is used for query expansion in our work. Also, pseudo relevance feedback techniques are used for automatic expansion. We denote the query expanded with UMLS or word2vec models as expanded query in our system. We first use cTAKES[1] to identify medical concepts. The terms identified as 'anatomy' or 'disorder' are expanded using UMLS. We include all the terms with the same CUI number. We use word embeddings to find two terms that are nearest to each other in the original query. The two terms are regarded as a loose phrase and is included in the original query.

2.3 Phrase search

Term dependency is a the characteristic of health articles. For example, "inguinal hernia" means hernia occurs in inguinal part, but not the other parts of the body. In our work, we treat this phrasal medical concept as an integral part and implement phrase search in our system [3].

3 Experiments

In this section, we introduce the platforms and models that are used in our work and then we describe the submission runs for the task.

3.1 Terrier

Terrier[2] retrieval platform version 4.17 was used as the search engine. Terrier is described to be "a highly flexible, efficient, and effective open source search engine, readily deployable on large-scale collections of documents". It implements state-of-the-art indexing and retrieval functionalities, and provides an ideal platform for the rapid development and evaluation of large-scale retrieval applications. In our experiments, we use BM25 as the retrieval model and all the parameters are set to default.

3.2 cTAKES

In our work, we use cTAKES to identify the medical concepts in the query. Apache cTAKES is an open source natural language processing system for extraction of information from electronic medical record clinical free-text [2]. It includes following components:

- Sentence boundary detector
- Tokenizer
- Normalizer

[1] http://ctakes.apache.org/index.html
[2] http://terrier.org/

- Part-of-speech (POS) tagger
- Shallow parser
- Named entity recognition (NER) annotator, including status and negation annotators.

3.3 Word2vec models

In our work, we produce word embeddings using word2vec algorithms [4]. Word2vec uses shallow, two-layer neural networks and includes two model architectures for learning distributed representations of words: Continuous Bag-of- Words model (CBOW) and Continuous Skip-gram Model (Skip-gram). We used the CBOW model, a context window size equal to five and a word vector of size 100 in our experiments. We use the data snapshotted on 16th Feb, 2017 from PMC Open Access Subset [3] and the trained word embeddings contain 25,140,380 words types in the result.

3.4 Runs

We submit 5 runs for IRTask 1. For all runs, the stop words are removed and Porter stemmer are used for word stemming. We use BM25 as the weighting model and the parameters are set to default in Terrier.

UEvora_EN_Run1: We use cTAKES to identify the medical concepts. The medical concept identified as a phrase replaces the single terms in the original query. Meanwhile, we expand the concepts with UMLS synonyms and increase their weight.

UEvora_EN_Run2: Based on run1, for the terms that are not identified by cTAKEs, we use our trained word2vec model to do the expansion.

UEvora_EN_Run3: For identified medical concept terms, we expand them with UMLS synonyms and increase their weight.

UEvora_EN_Run4: The medical concepts are identified with cTAKEs. The concept identified as phrase replaces the single terms in the original query.

UEvora_EN_Run5: For identified medical concept terms, we expand them with our trained word2vec model and increase their weight.

Acknowledgement

This work was supported by EACEA under the Erasmus Mundus Action 2, Strand 1 project LEADER - Links in Europe and Asia for engineering, eDucation, Enterprise and Research exchanges.

[3] https://www.ncbi.nlm.nih.gov/pmc/tools/openftlist/

References

1. Cogley, James. Applying natural language processing to clinical information retrieval. (2014)
2. Savova, Guergana K., James J. Masanz, Philip V. Ogren, Jiaping Zheng, Sunghwan Sohn, Karin C. Kipper-Schuler, and Christopher G. Chute. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. Journal of the American Medical Informatics Association 17, no. 5 (2010)
3. Manning, Christopher D., Prabhakar Raghavan, and Hinrich Schütze. Introduction to information retrieval. Vol. 1, no. 1. Cambridge: Cambridge university press, 2008
4. Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781 (2013)
5. Lorraine Goeuriot, Liadh Kelly, Hanna Suominen, Aurélie Névéol, Aude Robert, Evangelos Kanoulas, Rene Spijker, João Palotti, and Guido Zuccon. CLEF 2017 eHealth Evaluation Lab Overview. CLEF 2017 - 8th Conference and Labs of the Evaluation Forum, Lecture Notes in Computer Science (LNCS), Springer, September, 2017
6. Palotti, J., Zuccon, G., Jimmy, Pecina, P., Lupu, M., Goeuriot, L., Kelly, L., Hanbury, A.: CLEF 2017 Task Overview: The IR Task at the eHealth Evaluation Lab. In: Working Notes of Conference and Labs of the Evaluation (CLEF) Forum. CEUR Workshop Proceedings (2017)