Improving Personalized Consumer Health Search:
Notebook for eHealth at CLEF 2018

Hua Yang[1,2] and Teresa Gonçalves[1]

[1]Computer Science Department, University of Évora
[2]ZhongYuan University of Technology, Zhengzhou, China
huayangchn@gmail.com,tcg@uevora.pt

Abstract. CLEF 2018 eHealth Consumer Health Search task aims to investigate the effectiveness of the information retrieval systems in providing health information to common health consumers. Compared to previous years, this year's task includes five subtasks and adopts new data corpus and set of queries. This paper presents the work of University of Evora participating in two subtasks: IRtask-1 and IRtask-2. It explores the use of learning to rank techniques as well as query expansion approaches. A number of field based features are used for training a learning to rank model and a medical concept model proposed in previous work is re-employed for this year's new task. Word vectors and UMLS are used as query expansion sources. Four runs were submitted to each task accordingly.

Keywords: health information search, learning to rank, query expansion, UMLS, word vectors

1 Introduction

CLEF 2018 eHealth Consumer Health Search (CHS) task is a continuation of the previous CLEF eHealth information retrieval (IR) tasks that started on 2013 [8, 3]. Search engines are commonly used by health consumers seeking better understanding about health problems or medical conditions. This task aims to research on the problem of retrieving web pages to a health consumer for his information needs. The 2018 CHS task includes 5 subtasks and uses a new web corpus and a new set of queries [3].

This paper describes the University of Évora (UEvora) approach to CLEF 2018 eHealth CHS subtasks IRTask-1 and IRtask-2. While IRTask-1 is a standard ad-hoc search that aims at retrieving relevant information to people seeking health advice on the web, IRTask-2 is a personalized search task. This task develops on top of the IRTask-1 and aims to personalize the retrieved list of search results to match user expertise.

The following questions were investigated by conducting experiments on CLEF 2018 eHealth CHS Task:

1. How does a model learned from data based on 2016 and 2017 CLEF eHealth IR task [6, 12] perform on this year's new data collection and new set of queries (learning to rank features exploring)?
2. When applying query expansion techniques, as an expansion source, will domain specific word embeddings (built from a medical related training corpus) outperform a domain specific thesaurus?
3. How does the medical concepts model proposed in previous work [9] perform on a new task?

2 Methods

To answer the questions proposed in the previous section, different approaches were employed in this work. To answer the first question, learning to rank techniques were used: a number of features were explored and assessments results from 2016 and 2017 CLEF eHealth IR task were used for training a model. For the second question, a pre-trained word vectors model was used as a source of query expansion and the result is compared to the one retrieved with query expansion using the domain specific United Medical Language System (UMLS) thesaurus. Finally, to tackle the third question, the model proposed in previous work [9] was re-employed and tested with this year's new task (composed of a a new data corpus and new set of queries).

2.1 Pre-processing

All queries were pre-processed with characters lower-casing, stop words removing and Porter Stemmer stemming. The default stop words list available in the IR platform Terrier 4.2[1] was used.

2.2 Learning to rank

The assessment results from 2016 and 2017 CLEF eHealth IR task were employed and used to train a learning to rank model where a number of fields based features were explored in this work.

Features extracted. In this work, a simple group of features on different fields were extracted for training a learning to rank model. Three information fields were considered: *title*, *H1* and *else*. One kind of the features were using a normal weighting model on a single field. BM25 and PL2 were used as the weighting model, with each weighting model for every field. Query independent features and field length were also taken into account. DL weighting model implementing a simple document length was used [5, 4].

[1] `http://terrier.org/`

Training a model. Different learning to rank algorithms were previously explored (logistic regression, random forests, LambdaMART, AdaRank and ListNet) and among all, LabmdaMART algorithm presented the best performance [1, 7]. As such, this algorithm was employed in this work.

The assessment results from 2016 and 2017 CLEF eHealth IRtask [12, 6] were used as the training data. For IRtask-1, the topical relevance results were used; the result documents were scored with 0, 1 or 2 representing not relevant, relevant or highly relevant, respectively. For IRtask-2, the understandability scores were used; the scores ranged from -1 to 100, with a higher score representing higher understandability. These results were used directly for training and no extra or further processing was performed.

2.3 Query expansion with a medical concepts model

A medical concepts model employed [9] as a source for query expansion. First, cTAKES[2], a Natural Language Processing tool, was used to identify the medical concepts present in a query. Next, the following techniques were applied: medical phrase concepts processing, medical term concepts processing and query expansion. Finally, the new terms were added building the new expanded query.

Two different expansion sources were used: UMLS and word vectors model. For UMLS based expansion, selected terms were added to the original query and the approach employed our previous work [9]. The word vector model was trained using 2011 and 2012 TREC Medical Records Track collection and Word2vec with a skipgram architecture was used as the training tool [11]. The vector dimension was set to 1000 and a total of 25,469 vectors were included in the model.

2.4 Pseudo Relevance Feedback

Besides the query expansion techniques, the pseudo relevance feedback was also tested for automatic expansion during retrieval process. The number of words was set to 10 and the number of top-ranked documents from which those words were extracted was set to 3 in Terrier 4.2 platform.

3 Experiments and Results

This section first briefly presents the IR platform employed in this work, the dataset and queries for the task, as well as the evaluation measures used for the assessments. Next is the description of the techniques employed in our experiments.

3.1 IR model

Terrier platform version 4.2 was chosen as IR model of the system. The Okapi BM25 weighting model was used with all the parameters set to default values.

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

3.2 Dataset

The dataset used in CLEF 2018 CHS task consisted of web pages acquired from the CommonCrawl. By submitting the task queries to the Microsoft Bing APIs repeatedly over a period of time, an initial list of websites used for acquisition was returned. Some URLs domains were excluded and a number of know reliable health websites were added[3]. Totally, a number of 1,903 sites were included in the list.

3.3 Queries

The basic query set used for CLEF 2018 CHS task consisted of 50 queries written in English and were issued by the general public to the search service [2].

For IRtask-1, this basic set of 50 queries was used as the input to participating systems. An example of a query is shown in Figure 1.

```
<query>
        <id> 151001 </id>
        <en> anemia diet therapy </en>
</query>
```

Fig. 1. An example query for IRtask-1

IRtask-2 was based on IRtask-1 with 7 variations for each query. The first 4 variations were issued by people with no medical backgrounds while the remaining were issued by medical experts. An example for IRtask-2 is shown in Figure 2.

3.4 Evaluation Measures

Different evaluation measures were used for IRtask-1 and IRtask-2. For IRtask-1 they were: Normalized Discounted Cumulative Gain at depth 10 (NDCG@10), binary preference-based measure (Bpref) and Rank Biased Precision (RBP) [10].

For IRtask-2, uRBPgr with alpha was be used for the assessment. Based on RBP, uRBPgr is calculated as

$$uRBP = (1 - \rho) \sum_{k=1}^{K} \rho^{k-1} r(k) u(k) \tag{1}$$

where the $u(k)$ function is a graded gain function for the understandability dimension. The parameter ρ attempts to model user behaviour and was set to

[3] `https://sites.google.com/view/clef-ehealth-2018/task-3-consumer-health-search/`

```
<query>
    <id>151001</id> <en>anemia diet therapy</en>
</query>
<query>
    <id>151002</id> <en>anemia change in diet</en>
</query>
<query>
    <id>151003</id> <en>diet for anemia</en>
</query>
<query>
    <id>151004</id> <en>diet for anemia</en>
</query>
<query>
    <id>151005</id> <en>anemia diet therapy</en>
</query>
<query>
    <id>151006</id> <en>anemia diet</en>
</query>
<query>
    <id>151007</id> <en>Diet for anemia</en>
</query>
```

Fig. 2. An example query for IRtask-2

0.8. The $r(k)$ function is the standard RBP gain function: the value is 1 if the document at rank k is relevant and 0 if it is irrelevant [10].

In IRTask 2, each topic has 7 query variations. A parameter *alpha* capturing user expertise is used when evaluating results for query variations. Setting *alpha* to increasing values, an increasing level of medical expertise across the query variations is modeled [6, 10]. For this task the value was set to $\{0.0, 0.2, 0.4, 0.5, 0.6, 0.8,$
$1.0\}$ to query variations 1 to 7, respectively.

3.5 Experiments

Four experiments were conducted for each sub-task. Next paragraphs discuss the techniques used.

Runs for IRTask-1. *UEvoraIRTask1Run1* is based on the Medical Concepts Model presented in previous work [9]. First, cTAKES is used to identify the medical concepts in a pre-processed query. Next, the identified concepts are expanded with UMLS; an extra weight of 2.0 and 1.5 were set for words expanded from a phrase concept or from a term concept, respectively. Then, a phrase is reconstructed into a loose phrase with the maximum interval words set to 2; the loose phrase is regarded as a must check item during the retrieval process.

Finally, these processed phrases and terms with extra weights are added to the query.

For *UEvoraIRTask1Run2* the same techniques are used but the expansion is performed with the pre-trained word vectors model.

UEvoraIRTask1Run3 uses a ranking model trained with the topical assessments from 2016 and 2017 CLEF eHealth IR task (see sub-section 2.2).

UEvoraIRTask1Run4 uses the similar techniques to *UEvoraIRTask1Run3*, yet using five folders cross validation to obtain a learning to rank model.

Runs for IRTask-2. Similar techniques and parameters were used for IRTask-2. *UEvoraIRTask2Run1* uses the same approach employed *UEvoraIRTask1Run1*; the queries and their variations were processed and issued to the retrieval system. *UEvoraIRTask2Run2* uses a similar approach to *UEvoraIRTask2Run1* with queries expanded using a pre-trained word vectors model and for *UEvoraIRTask2Run3* the understandability assessments from 2016 and 2017 CLEF eHealth IR task were used for training a learning to rank model. Finally and for *UEvoraIRTask2Run4*, the same techniques of *UEvoraTask2run3* were employed and a five cross validation was done when training the learning to rank model.

3.6 Results

The assessments for the CLEF eHealth 2018 IR tasks are still being conducted, so they are not available at this time.

4 Conclusion and future work

This working note reports the UÉvora team participation in two different tasks of CLEF 2018 eHealth CHS. A number of field based features were explored while applying learning to rank techniques. Based on previous work, both UMLS and a word vector model were applied for performing query expansion.

As the future work, the methods proposed in this paper will be further analyzed: different learning to rank features will be explored and an ensemble algorithm will be investigated.

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.

References

[1] Olivier Chapelle and Yi Chang. “Yahoo! learning to rank challenge overview”. In: *Proceedings of the Learning to Rank Challenge*. 2011, pp. 1–24.

[2] Lorraine Goeuriot et al. “Meta-analysis of the second phase of empirical and user-centered evaluations”. In: *Public Technical Report, Khresmoi Project*. 2014.

[3] Jimmy et al. “Overview of the CLEF 2018 Consumer Health Search Task.” In: *CLEF 2018 Evaluation Labs and Workshop: Online Working Notes, CEUR-WS,September,* 2018.

[4] Craig Macdonald, Rodrygo LT Santos, and Iadh Ounis. “The whens and hows of learning to rank for web search”. In: *Information Retrieval* 16.5 (2013), pp. 584–628.

[5] Craig Macdonald et al. “About learning models with multiple query dependent features”. In: *ACM Transactions on Information Systems (TOIS)* 31.3 (2013), p. 11.

[6] Joao Palotti et al. “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.

[7] Luca Soldaini and Nazli Goharian. “Learning to rank for consumer health search: a semantic approach”. In: *European Conference on Information Retrieval*. Springer. 2017, pp. 640–646.

[8] Hanna Suominen et al. “Overview of the CLEF eHealth Evaluation Lab 2018.” In: *CLEF 2018 - 8th Conference and Labs of the Evaluation Forum, Lecture Notes in Computer Science (LNCS), Springer,September,* 2018.

[9] Hua Yang and Teresa Gonçalves. “UEvora at CLEF eHealth 2017 Task 3”. In: *Working Notes of Conference and Labs of the Evaluation (CLEF) Forum. CEUR Workshop Proceedings*. 2017.

[10] Guido Zuccon. “Understandability biased evaluation for information retrieval”. In: *European Conference on Information Retrieval*. Springer. 2016, pp. 280–292.

[11] Guido Zuccon et al. “Integrating and evaluating neural word embeddings in information retrieval”. In: *Proceedings of the 20th Australasian document computing symposium*. ACM. 2015, p. 12.

[12] Guido Zuccon et al. “The IR Task at the CLEF eHealth evaluation lab 2016: user-centred health information retrieval”. In: *CLEF 2016-Conference and Labs of the Evaluation Forum*. Vol. 1609. 2016, pp. 15–27.