Improving Understandability in Consumer Health Information Search: UEVORA @ 2016 FIRE CHIS

Hua Yang
Computer science department
University of Évora
Évora, Portugal
huayangchn@gmail.com

Teresa Gonçalves
Computer science department
University of Évora
Évora, Portugal
tcg@uevora.pt

ABSTRACT

This paper presents our work at 2016 FIRE CHIS. Given a CHIS query and a document associated with that query, the task is to classify the sentences in the document as relevant to the query or not; and further classify the relevant sentences to be supporting, neutral or opposing to the claim made in the query. In this paper, we present two different approaches to do the classification. With the first approach, we implement two models to satisfy the task. We first implement an information retrieval model to retrieve the sentences that are relevant to the query; and then we use supervised learning method to train a classification model to classify the relevant sentences into support, oppose or neutral. With the second approach, we only use machine learning techniques to learn a model and classify the sentences into four classes (relevant & support, relevant & neutral, relevant & oppose, irrelevant & neutral). Our submission for CHIS uses the first approach.

CCS Concepts

• **Information systems→Data management system engines**

Keywords

Health information search; machine learning; IR

1. INTRODUCTION

Online search engines have become a common way for obtaining health information; a life project report shows that about 69% of U.S. adults have the experience of using Internet as a tool for health information such as weight, diet, symptoms and so on [4]. In the meanwhile, research interest in health information retrieval (HIR) has also grown in the past years. As a matter of fact, health information is of interest to a variety of users, from physicians to specialists, from practitioners to nurses, from patients to patients family, and from biomedical researchers to consumers (general public). Also, health information may be available in diverse sources, like electronic health record, personal health records, general web, social media, journal articles, and wearable devices and sensors [5].

While factual health information search has matured considerably, complex health information searching with more than just one single correct answer still remains elusive. Consumer Health Information Search (CHIS) for FIRE 2016 is proposed for investigating complex health information search by laypeople. In this scenario, laypeople search for health information with multiple perspectives from diverse sources both from medical research and from real world patient narratives.

There are two sets of tasks:

A) Given a CHIS query, and a document/set of documents associated with that query, the task is to classify the sentences in the document as relevant to the query or not. The relevant sentences are those from that document, which are useful in providing the answer to the query.

B) These relevant sentences need to be further classified as supporting the claim made in the query, or opposing the claim made in the query.

The five queries proposed in the task are showed in figure 1. Figure 2 gives an example of the output of the system. Annotated data set is provided to participants.

This paper is divided into 4 sections. In the first section, we briefly introduced the background and the 2016 FIRE CHIS task. We then talk about the methods we use in the second section. Two different approaches are experimented to accomplish the task and each approach will be discussed. Experiments and the results are presented in the third section. Finally, the conclusions are made.

Q1: Does sun exposure cause skin cancer?
Q2: Are e-cigarettes safer than normal cigarettes?
Q3: Can Harmone Replacement Therapy(HRT) cause cancer?
Q4: Can MMR Vaccine lead to children developing autism?
Q5:Should I take vitamin C for common cold?

Figure 1. 2016 FIRE CHIS queries

Example Query:
Are e-cigarettes safer than normal cigarettes?

S1:
Because some research has suggested that the levels of most toxicants in vapor are lower than the levels in smoke, e-cigarettes have been deemed to be safer than regular cigarettes
.A) Relevant, B)Support

S2:
David Peyton, a chemistry professor at Portland State University who helped conduct the research, says that the type of formaldehyde generated by e-cigarettes could increase the likelihood it would get deposited in the lung, leading to lung cancer.
A) Relevant, B) oppose

S3:
Harvey Simon, MD, Harvard Health Editor, expressed concern that the nicotine amounts in e-cigarettes can vary significantly.
A)Irrelevant, B) Neutral

Figure 2. 2016 FIRE CHIS task description

2. METHODS

We propose two different approaches to accomplish the task. In order to make it easier to explain, we name them program A and program B. In program A, two different models are trained by using both state of the art in information retrieval and machine learning techniques. In program B, we take the task as a whole and only use machine learning techniques. One single classification model is trained in program B. We will discuss each approach in detail in the following part.

2.1 Program A

Considering the task is divided into sub-tasks, we implement two different models to satisfy the task, with each model processing one task. For task A, we implement an information retrieval (IR) model to retrieve relevant sentences. The retrieved sentences are regarded as relevant to the query, and non-retrieved ones as irrelevant. For task B, we use a supervised learning algorithm to get a classification model. The retrieved sentences from the first part are then classified as support, oppose or neutral to the claim made in the query.

2.1.1 An IR model for Task A

In task A, sentences provided by the organizer should be classified as relevant to the queries or not. We implement an IR model to do this classification. Retrieved sentences are regarded as relevant to the query and non-retrieved as irrelevant. Figure 3 depicts our model for task A. First, we input the original task queries and provided sentences into the IR model. The relevant sentences are retrieved and ranked according to the weighting methods. Top ranked (in our experiments, we choose top 3) relevant sentences are used as the source to expand the original queries. Expanded queries are used as the input. The IR model is used again to retrieve sentences with expanded queries. The relevant sentences are used as the input of a classification model works. We regard all the retrieved sentences from our IR model as relevant to the query and we use them the input of task B.

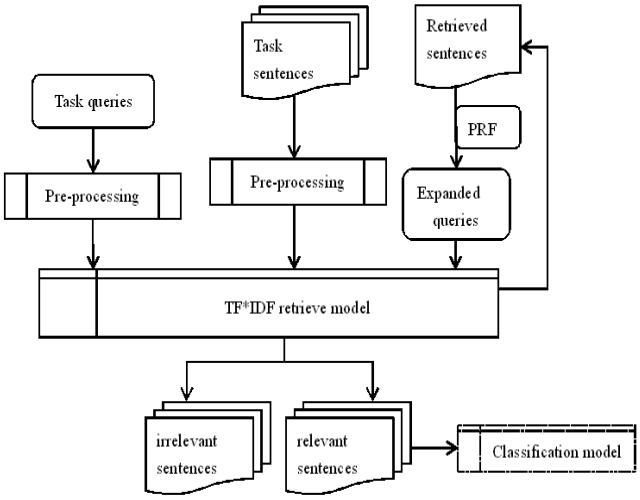

Figure 3. information retrieval model for task A

Terrier[1] is used to implement a baseline IR model. All queries and sentences are pre-processed. Stop-words are removed, stemming and normalization are applied. TF*IDF weighting model is used for the computation of sentence scores with respect to the query. The queries can be retrieved one by one or in batch. We use pseudo relevance feedback as a way to expand the original queries. We set all parameters to Terrier the default ones.

Pseudo relevance feedback (a.k.a. blind relevance feedback) is a way to improve retrieval performance without the user interaction [1]. Previous works showed its effectiveness in improving the performance [2] [3]. Figure 4 depicts how this technique can be used in an IR model to satisfy the user.

This technique is used in our experiments to expand the original query. The most informative terms are extracted from top-returned documents as the expanded query terms, as shown in Figure 4. We use Bo1 [6] as the expanded term weighting model. A Bo1 model uses the Bose-Einstein statistics and terms are weighted in the top retrieved documents. In our experiments, 10 expansion terms are extracted from the top 3 retrieved documents. No other query expansion techniques are used in our experiments.

[1] Terrier.org.

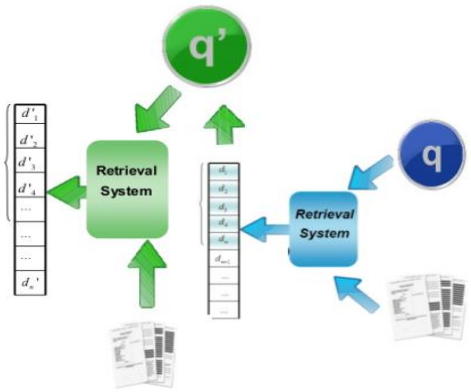

Figure 4. Pseudo relevance feedback[2]

2.1.2 A classification model for task B

For task B, we propose a classification model, presented in Figure 5. With a classification model, we further classify the retrieved sentences into different classes.

The annotated dataset provided by the organizer is first pre-processed. Then TF*IDF scheme is used to extract data features from the text. These features will be used as the input of the learning system to train a classification model. This model is able to further classify the relevant sentences retrieved from the IR model into support, oppose or neutral to the claim stated in the query.

TextBlob[3] tool is used for text processing. Naïve Bayes and decision tree classifiers are used as learning methods. Only TF*IDF features are extracted, no other data features are used in our experiments.

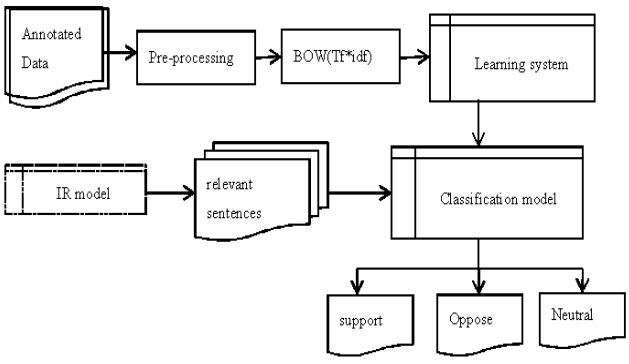

Figure5. classification model for task B

[2]Image from http://www.slideshare.net/LironZighelnic/querydrift-prevention-for-robust-query-expansion-presentation-43186077

[3] https://textblob.readthedocs.io/en/dev/

2.1.3 Integration

The retrieved sentences by an IR model are regarded as relevant to the query and they are further labeled as 'neutral', 'support', or 'oppose' to the query by the classification model. The non-retrieved sentences from the IR model are regarded as irrelevant to the query, and we assign 'neutral' label to all the irrelevant sentences.

2.2 Program B

As another approach to figure out the problem and provide multi-perspective for the users, we look on the task as a whole and re-organize the annotated data with four different labels:

-irrelevant & neutral

-relevant & support

-relevant & oppose

-relevant & neutral

Using the annotated data with the labels above, we get a classification model and this model is used to classify the test sentences into those four classes. The approach is the same as the one described in sub-section 2.1.3, but here we are using all the sentences and instead of having three classes, we have four, as figure 6 shows. The output is a sentence with one label from the fours that we list above. For example:

Sentence: Harvey Simon, MD, Harvard Health Editor, expressed concern that the nicotine amounts in e-cigarettes can vary significantly.
Output: Irrelevant & Neutral

All the sentences provided are pre-processed data and used to train a classification model with supervised machine learning techniques. We extract features with TF*IDF scheme. Test data needs to be pre-processed before classification.

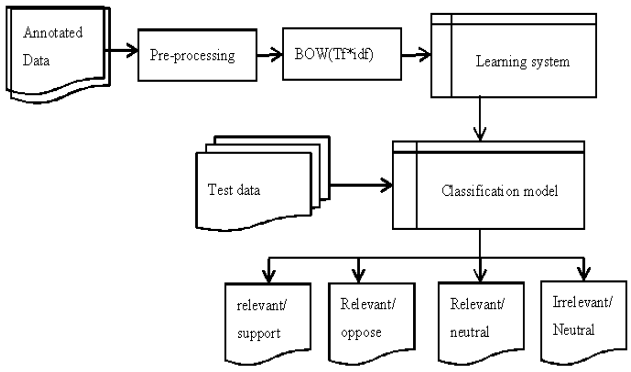

Figure 6. classification model for program B

3. EXPERIMENTS AND RESULTS

In this part, we give the results in our experiments. We will present our experiments separately according to each program we proposed in the previous part.

3.1 Experiments of Program A

3.1.1 Runs for task A

The results for different runs are shown in table1. TrecEval[4] program is used to evaluate the performance. We produce different runs to compare the performance using F1 score as the evaluation method.

-taskA.run1: process all the queries without bath pseudo relevance feedback

-taskA.run2: process all the queries in batch with pseudo relevance feedback

-taskA.run3: process the queries individually without pseudo relevance feedback

-taskA.run4: process the queries individually with pseudo relevance feedback

We got our best results with run4 and the average F1 score is 0.73. The results present that our IR model works well on query3, query4 and query5.

Considering the way of processing, we can see that processing the queries one by one is much better than all the queries in batch.

As a way to do the query expansion, PRF technique does improve the recall obviously, which means it can get more relevant documents returned. Also, this technique reacts differently depending on the processing way. If all the queries are processed in batch, using PFR decreases the performance in F1 score compared with the results without using PFR,. If the query is processed one by one, PRF increases the performance totally; but some queries show a lit bit down score compared with non-PRF using. We can also see that for query1 and query2, the score is improved sharply when using PRF. Combining the task and our system, we adopt PRF as a way to improve the system performance.

3.1.2 Run for task B

For task B, we use the traditional TF*IDF scheme to extract data features and Naïve Bayes is used as the learning method. Table 2 present our experiment results for this part.

From the results, we can see that the average score for this classification is 0.28, which is very low.

The classification is based on the results from the IR model. Some sentences may be irrelevant to the query indeed, but is classified as relevant to query, this kind of sentences are regarded as relevant and be classified by the classification model. This will affect the performance of the system.

[4] http://trec.nist.gov/trec_eval/

Table 1 results comparison of taskA runs (F1 score)

	taskA.run1	taskA.run2	taskA.run3	taskA.run4
Query1	0.46	0.27	0.62	0.65
Query2	0	0.02	0.18	0.55
Query3	0.52	0.43	0.88	0.86
Query4	0.59	0.37	0.86	0.84
Query5	0.62	0.40	0.77	0.75
Average	0.44	0.30	0.66	0.73

Table 2 results of taskB (F1 score)

	Task B
Query1	0.33
Query2	0.30
Query3	0.24
Query4	0.27
Query5	0.26
Average	0.28

3.2 Experiments of Program B

Table 3 gives the final results for this program. In this program, we regard the task as a whole and only one classification model is trained. We evaluate the final output of the program and the score is used for measuring both task A and task B as an integral.

The average score for this model is 0.64. We get highest score for query 3 and the lowest one for query 5.

Table 3 results of program B (F1 score)

	Task B
Query1	0.56
Query2	0.56
Query3	0.8
Query4	0.73
Query5	0.54
Average	0.64

4. CONCLUSION

In this paper, we present our two different approaches to accomplish 2016 FIRE CHIS task.

With the first approach, we implement both an IR model and a classification model. The results show that our IR model works well generally except on query2. The classification model shows a low performance for all.

With the second approach, we take the task as a whole and using machine learning techniques only to do the classification.

Although we figure out different approaches to the task, we have different output form for two approaches; we do not compare the performance of both approaches. The second approach presented in our paper is just another possible way to solve the problem proposed by the organizer. Program A is used as the final submission to the challenge.

5. REFERENCES

[1] Christopher D Manning and Hinrich Schütze. Foundations of statistical natural language processing, volume 999. MIT Press, 1999.

[2] Yang Song, Yun He, Qinmin Hu, Liang He, and E Mark Haacke. Ecnu at 2015 ehealth task 2: User-centred health information retrieval. Proceedings of the ShARe/CLEF eHealth Evaluation Lab, 2015.

[3] Ellen M Voorhees, Donna K Harman, et al. TREC: Experiment and evaluation in information retrieval, volume 1. MIT press Cambridge, 2005.

[4] Susannah Fox and Maeve Duggan. Tracking for health. Pew Research Center's Internet & American Life Project, 2013.

[5] Lorraine Goeuriot, Gareth JF Jones, Liadh Kelly, Henning Müller, and Justin Zobel. Medical information retrieval: introduction to the special issue. Information Retrieval Journal, 1(19):1–5, 2016.

[6] Giambattista Amati. Probability models for information retrieval based on divergence from randomness. PhD thesis, University of Glasgow, 2003.