MayoNLPTeam at the 2016 CLEF eHealth Information Retrieval Task 1

Yanshan Wang[1], Stephen Wu[2] and Hongfang Liu[1]

[1] Department of Health Sciences Research, Mayo Clinic, Rochester, Minnesota, USA
{Wang.Yanshan, Liu.Hongfang}@mayo.edu
[2] Department of Medical Informatics & Clinical Epidemiology, Oregon Health and Science University, Portland, Oregon, USA
wst@ohsu.edu

Abstract. This paper presents the participation of MayoNLPTeam in the 2016 CLEF eHealth Information Retrieval Task (IR Task 1: ad-hoc search). We explored a Part-of-Speech (POS) based query term weighting approach which assigns different weights to the query terms according to their POS categories. The weights are learned by defining an objective function based on the mean average precision. We applied the proposed approach with the optimal weights obtained from TREC 2011 and 2012 Medical Records Track into the Query Likelihood model (Run 2) and Markov Random Field (MRF) models (Run 3). The conventional Query Likelihood model was implemented as the baseline (Run 1).

Keywords: information retrieval, Part-of-Speech, language model, Markov Random Field model

1 Introduction

The amount of health information on the web has increased tremendously during the last decades. People access these contents to find information or answers regarding their health concerns. According to a late 2013 survey by the Pew Research Center's Internet and American Life Project, 85% of US adults use the Internet and 72% of them have looked for health information online [1]. However, it is difficult to find the information related precisely to their concerns. To tackle this issue, the 2016 CLEF eHealth Information Retrieval (IR) Task 1 [3,13] focuses on the retrieval of health contents on the web for the health queries generated by exploring real consumer posts from health forums. The goal of this task is to explore possible IR systems that people could use to search for information or answers to their health questions instead of posting those questions on health forums and waiting for answers.

As a participant in IR Task 1, we introduce a Part-of-Speech (POS) based query term weighting approach. The POS property reflects whether the term is informative or not. Intuitively, a noun is more informative than a preposition, and in the medical domain a proper noun is more important than a noun to understand semantics of a query. Therefore, we hypothesize that leveraging POS information to weight the query terms would improve the performance of IR systems. The experiments on the Electric Health Records (EHRs) retrieval provided by the Text REtrieval Conference (TREC) 2011 and 2012 Medical Records tracks [10] have verified our hypothesis. We would like to examine whether this approach is viable for the internet health contents retrieval.

We submitted three official runs to CLEF eHealth. In Run 1, we utilized plain text of the test topics as the input queries and the Query Likelihood model with Dirichlet smoothing [11] as the retrieval model. This run was served as a baseline. In Run 2, we utilized the POS-based query term weighting method to improve the Query Likelihood model. Different from Run 2 where we assumed that the terms were independent, Run 3 applied Markov Random Field (MRF) model [5] and incorporated the POS-based query term weighting method.

The rest of this paper is organized as follows. Section 2 describes the dataset and queries of the 2016 CLEF eHealth IR Task. Section 3 presents the our framework in detail including the system details and how we trained the weights. The experiments and experimental results are shown in Section 4. Section 5 concludes our study.

2 Dataset and Queries of CLEF eHealth IR Task 1

The ClueWeb12 B13 dataset is used in the 2016 CLEF eHealth IR Task. This dataset is a small portion of the crawled web documents in ClueWeb12 dataset as part of the lemur project[1]. It contains over 52 million documents. Those web documents in this corpus include health and non-health contents. Each document includes a *title* field in the "WARC- TREC-ID" field of the document's WARC header and a *title* and a *heading* field. The goal of this challenge is to find those relevant health contents given an input query.

The queries provided by the task are extracted from the posts in the *askDocs* health web forum[2]. Therefore, this set of queries reflects the real information needs of health consumers. Each of six query creators with different medical expertise was given 50 initial posts from the forum to generate the queries with a total of 300 queries created. Each query has an *id* field and a *title* field. The *id* field is used to distinguish the queries while the *title* field represents the queries. Task 1 of the challenge requires us to treat each query individually and submit up to 3 ranked runs with up to 1000 documents per query.

3 Method

In this section, we present the IR system and the proposed approach, and then detail the submitted runs.

[1] http://lemurproject.org/

[2] https://www.reddit.com/r/AskDocs/

3.1 System Overview

In our previous submission of the 2013 ShARe/CLEF eHealth Evaluation Lab [12], we focused on semantics and utilized multiple external sources, such as Mayo Clinic clinical notes collection and the Unified Medical Language System (UMLS), for query expansion. This year we focus more on syntactics and leverage the POS property to assign different weights to query terms.

Fig. 1 gives an overview of our system. The whole system consists of two modules: Query module and Retrieval module. In the Query module, an input query is annotated by a POS annotation engine and each query term is then weighted based on its POS category. In the Retrieval module, the relevant documents are then retrieved based on the retrieval models. In our system, we separately utilize the Query Likelihood model and MRF model as the retrieval models to compare the performance. The overall system is a very typical IR system except adding the query term weighting component.

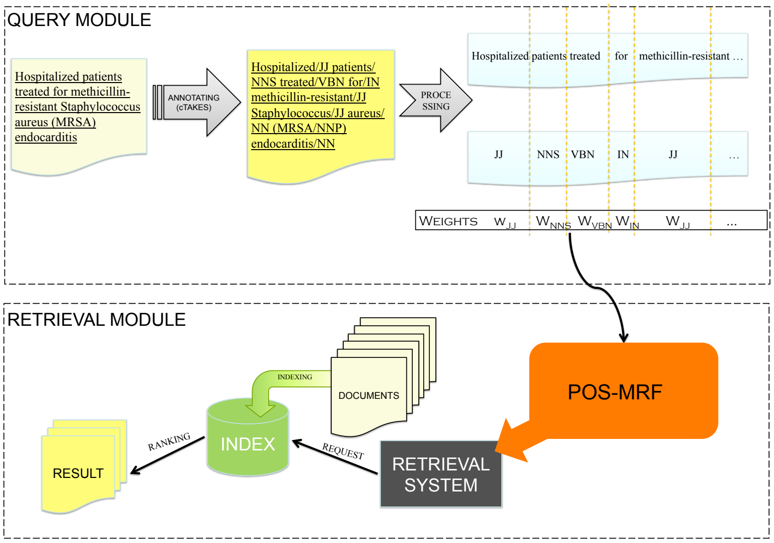
QUERY MODULE
Hospitalized patients treated for methicillin-resistant Staphylococcus aureus (MRSA) endocarditis
ANNOTATING (cTAKES)
Hospitalized/JJ patients/NNS treated/VBN for/IN methicillin-resistant/JJ Staphylococcus/JJ aureus/NN (MRSA/NNP) endocarditis/NN
PROCESSING
Hospitalized patients treated for methicillin-resistant ...
JJ NNS VBN IN JJ ...
WEIGHTS W_{JJ} W_{NNS} W_{VBN} W_{IN} W_{JJ} ...
RETRIEVAL MODULE
INDEXING
DOCUMENTS
POS-MRF
INDEX
RANKING
REQUEST
RESULT
RETRIEVAL SYSTEM

Fig. 1. System Overview

3.2 Preprocessing and Indexing

We used the computing infrastructure, i.e., Microsoft Azure, provided by the organizers for the task. The dataset and standard indexes were available in the Azure instance. Specifically, the preprocessing included stopwords removal and Krovetz stemming. Then the *title* and *heading* fields were indexed using Indri [8].

The POS categories for queries were obtained using Apache cTAKES [3], an open source software package. The POS model in Apache cTAKES was derived using multiple POS annotated corpora including a corpus of a collection of clinical notes, GENIA corpus [9] and Penn Treebank [4], and thus favorable for medical text annotation.

3.3 Part-of-Speech based Query Term Weighting

In this subsection, we describe the POS-based query term weighting approach and how the Indri queries are generated.

Given the bag-of-words assumption, the conventional Query Likelihood model ranks the documents according to the following ranking function:

$$r(Q, D) = \sum_{q_i} f(q_i, D), \tag{1}$$

where Q represents the query, D the document, q_i the ith query term, and $f(q_i, D)$ the potential function over q_i and D. $f(q_i, D)$ could be defined by various methods, such as *tf-idf* [7], BM25 [6], Jelinck-Mercer method [2], or Dirichlet smoothing method [11]. The conventional Query Likelihood model treats each query term equally. However, in the medical domain we observe that the query terms are not always equally important. For example, in the query *Alzheimer's disease*, the term *Alzheimer* is more important and specific than *disease* to understand the query semantics where the semantics can be partially inferred from syntactics. As in the previous example, *Alzheimer* is a proper noun and *disease* is a noun. The observation motivates us to assign different weights to query terms according to the POS categories. By doing so, the ranking function can be written as:

$$r(Q, D) = \lambda_{q_1} f(q_1, D) + \lambda_{q_2} f(q_2, D) + \cdots + \lambda_{q_n} f(q_n, D), \tag{2}$$

where λ_{q_i} is the weight corresponding to the POS category of query term q_i.

Beyond the bag-of-words assumption, the MRF model considers the relations between terms by using the Markov property [5]. The ranking function is defined as:

$$r(Q, D) = \sum_{c \in T} f(c, q_i, D) + \sum_{c \in O} f(c, q_i, D) + \sum_{c \in U} f(c, q_i, D), \tag{3}$$

where c is the clique set, q_i is the ith query term in that clique, and T, O, U denote dependency types *full independence*, *sequential dependence*, *full dependence*, respectively. We can incorporate the POS-based query term weighting approach and define the new ranking function as:

$$\begin{aligned} r(Q, D) = \sum_{c \in T} \left\{ \sum_{q_i} \lambda_{q_i} f(c, q_i, D) \right\} + \sum_{c \in O} \left\{ \sum_{q_i} \lambda_{q_i} f(c, q_i, D) \right\} \\ + \sum_{c \in U} \left\{ \sum_{q_i} \lambda_{q_i} f(c, q_i, D) \right\}. \end{aligned} \tag{4}$$

[3] http://ctakes.apache.org/

3.4 Weight Training

The weights can be trained by defining an objective function based on retrieval performance metrics, e.g., mean average precision (MAP). Our goal is to maximize the performance metrics. Since there are over 30 POS categories according to the Penn Treebank Project [4], this is a multidimensional optimization problem.

We adopted MAP as the performance metric for weight training and considered seven POS categories: (*singular or mass nouns (NN), plural nouns (NNS), past participle verbs (VBN), past tense verbs (VBD), adjectives (JJ), adverbs (RB), singular proper nouns (NNP)*) and marked all other categories as *others*. Then we utilized a cyclic coordinate method to solve this optimization problem. The dataset of the TREC 2011 and 2012 Medical Records tracks [10] was used to train the weights. It contained over 93 thousand de-identified clinical reports and 34 test queries for the TREC 2011 and 47 for the TREC 2012. We first trained on the TREC 2011 data and tested on the TREC 2012 data, and then trained on the TREC 2012 data and tested on the TREC 2011 data. Finally the average weight for each POS category was used to generate the Indri queries for the 2016 CLEF eHealth IR Task.

4 Experiments and Results

Table 1 shows the average results of the experiments on the TREC 2011 and 2012 Medical Records tracks. We can observe that the proposed approach enhanced the conventional Query Likelihood model and the POS+MRF model performs better than POS+Query Likelihood model. Table 2 lists the optimal weights learned from this experiment. We use those weights in the 2016 CLEF task. We submitted three runs including one baseline run (Run 1) and two runs incorporating POS information. For each of the submitted runs, we set the Dirichlet smoothing parameter μ to 1000.

Table 1. Experimental results on the TREC 2011 and 2012 Medical Records tracks

Method	Data	MAP	P@10
Query Likelihood model	TREC 2011	0.30	0.48
Query Likelihood model	TREC 2012	0.21	0.37
POS + Query Likelihood model	TREC 2011	0.33	0.47
POS + Query Likelihood model	TREC 2012	0.23	0.37
POS + MRF	TREC 2011	0.36	0.56
POS + MRF	TREC 2012	0.27	0.45

[4] https://www.cis.upenn.edu/ treebank/

Table 2. The optimal weights for POS categories

POS category	"NN"	"NNS"	"VBN"	"VBD"	"JJ"	"RB"	"NNP"	"others"
weight	0.5970	0.2265	0.3065	0.2260	0.3730	0.1040	0.8930	0.0

5 Conclusion

In this paper we presents our participation to the 2016 CLEF eHealth Information Retrieval Task 1. We explored a Part-of-Speech (POS) based query term weighting approach which assigns different weights to the query terms according to their POS categories. In the future work, we would like to explore how to utilize the external resources for query expansion in the proposed method.

Acknowledgments

The authors gratefully acknowledge the support from the National Library of Medicine (NLM) grant R01LM11934.

References

1. Pew Internet and American Life Project: Health Fact Sheet. Tech. rep., Pew Research Center (2013), `http://www.pewinternet.org/fact-sheets/health-fact-sheet/`
2. Jelinek, F.: Interpolated estimation of markov source parameters from sparse data. Pattern recognition in practice pp. 381–402 (1980)
3. Kelly, L., Goeuriot, L., Suominen, H., Nvol, A., Palotti, J., Zuccon, G.: Overview of the clef ehealth evaluation lab 2016. clef 2016 - 7th conference and labs of the evaluation forum. In: Lecture Notes in Computer Science (LNCS). Springer (2016)
4. Marcus, M.P., Marcinkiewicz, M.A., Santorini, B.: Building a large annotated corpus of english: The penn treebank. Computational linguistics 19(2), 313–330 (1993)
5. Metzler, D., Croft, W.B.: A Markov random field model for term dependencies. In: Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval. pp. 472–479. ACM (2005)
6. Robertson, S.E., Walker, S., Jones, S., Hancock-Beaulieu, M.M., Gatford, M., et al.: Okapi at trec-3. NIST SPECIAL PUBLICATION SP pp. 109–109 (1995)
7. Salton, G., McGill, M.J.: Introduction to modern information retrieval. McGraw-Hill, Inc. (1986)
8. 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 Intelligent Analysis. vol. 2, pp. 2–6. Citeseer (2005)
9. Tsuruoka, Y., Tateishi, Y., Kim, J.D., Ohta, T., McNaught, J., Ananiadou, S., Tsujii, J.: Developing a robust part-of-speech tagger for biomedical text. Advances in informatics pp. 382–392 (2005)
10. Voorhees, E.M., Hersh, W.: Overview of the trec 2012 medical records track. In: The Twenty-first Text REtrieval Conference proceedings (TREC) (2012)

11. Zhai, C., Lafferty, J.: A study of smoothing methods for language models applied to ad hoc information retrieval. In: Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval. pp. 334–342. ACM (2001)
12. Zhu, D., Wu, S.T.I., Masanz, J.J., Carterette, B., Liu, H.: Using discharge summaries to improve information retrieval in clinical domain. In: Proceedings of the ShARe/CLEF eHealth Evaluation Lab (2013)
13. Zuccon, G., Palotti, J., Goeuriot, L., Kelly, L., Lupu, M., Pecina, P., Mueller, H., Budaher, J., Deacon, A.: The ir task at the clef ehealth evaluation lab 2016: User-centred health information retrieval. In: CLEF 2016 Evaluation Labs and Workshop: Online Working Notes. CEUR-WS (2016)