Comparison of Several Word embedding Sources for Medical Information Retrieval

Julie Budaher, Mohannad Almasri, and Lorraine Goeuriot

Laboratoire d'informatique de Grenoble
Université Grenoble Alpes

{*prenom.nom*}*@imag.fr*

No Institute Given

Abstract. This paper describes the participation of MRIM team in Task 3: Patient-Centered Information Retrieval-IRTask 1: Ad-hoc search of CLEF eHealth Evaluation lab 2016. The aim of this task is to evaluate the effectiveness of information retrieval systems when searching for health content on the web. Our submission investigates the effectiveness of word embedding for query expansion in the health domain. We experiment two variants of query expansion method using word embedding. Our first run is a baseline system with default stopping and stemming. The other two runs expand the queries using two different word embedding sources. Our three runs are conducted on Terrier platform using Dirichlet language model. **Keywords:** Query expansion, Word embedding, Language model

1 Introduction

The goal of the eHealth evaluation lab is to evaluate information retrieval systems helping people in understanding their health information [2]. We are describing in this paper our participation to Task 3: Patient-Centered Information Retrieval which aims to evaluate the effectiveness of information retrieval systems when searching for health content on the web, with the objective to foster research and development of search engines tailored to health information seeking [4], this task is divided into three sub-tasks: ad-hoc search which extends the evaluation framework used in 2015 (which considered, along with topical relevance, also the readability of the retrieved documents) to consider further dimensions of relevance such as the reliability of the retrieved information, query variation which explores query variations for an information need and multilingual search which offers parallel queries in several languages (Czech, French, Hungarian, German, Polish and Swedish)[3].

Our team MRIM has submitted three runs, in order to investigate the following research questions:

1. Is the word embedding approach for query expansion effective for consumer health search?

2. What influence has the word embedding source on the results?

This paper is organized as follows. In Section 2, we describe our approaches and in Section 3, we describe our future work and conclusion.

2 Approach: Using Word Embedding for Query Expansion

2.1 Dataset

The dataset contains a document collection, a set of topics, and relevance judgments. The document collection is Clueweb12 B13[1], created by the Lemur Project to support research on information retrieval and related human language technologies. This collection contains more than 50 million documents, on varied topics. The collection has been made available to participants via the Microsoft Azure platform, along with in standard indexes built with the Terrier tool and the Indri tool.

The topics provided explore real health consumer cases, extracted from posts from health web forums. The posts were extracted from the 'askDocs' forum of Reddit, and presented to query generators, who had to create queries based on what they read and think would be queried by the posts author. Different query creators generated different queries for the same post, creating variations of the same information need (forum post).

For IRTask 1, participants had to treat each query individually, submitting the returned documents for each query. Example queries follow:

```
<queries>
<query>
    <id>900001</id>
    <title>medicine for nasal drip</title>
</query>
<query>
    <id>900002</id>
    <title>bottle neck and nasal drip medicine</title>
</query>
....
  <query>
    <id>904001</id>
    <title>omeprazole side effect</title>
  </query>
....
</queries>
```

Example queries were provided, and a final set of 300 queries was distributed for the runs.

[1] http://lemurproject.org/clueweb12/specs.php

2.2 Runs

Three runs were submitted, the mandatory baseline run and two other runs with query expansion based on word embedding with two different training sets. We used Terrier for indexing and retrieval from the Azure platform.

Run1- Baseline In this run, we apply Dirichlet language model with the default Mu value (2500) on the 300 queries. Stop-words are removed using default Terrier TermPipeline interface. We use PorterStemmer for stemming as used in the document index. This model is the simple approximation and is the baseline for comparing more complex models.

Run2 and Run3 The second two runs are using query expansion method based on word embedding. Term embeddings are learned using Neural Networks. Each term is represented by a real-valued vector. The resulting vectors carry relationships between terms. The similarity between two terms is measured with the normalized cosine between their two vectors. The resulting expanded query is the union of the original query terms and the expanded terms which are the k-most similar terms to those of the query.

The tool `word2vec` is used to generate deep learning vectors. The word2vec tool takes as its input a large corpus of text and produces the term vectors as its output.

The training corpus of Run2 is built using three different CLEF medical collections: Image2009, Case2011 and Case2012. The training corpus consists of about 400 million words. The vocabulary size for this training corpus is about 350,000 different terms [1].

The training corpus of the third run is built using CLEF eHealth 2014 medical collection. The training corpus consists of 1,056,629,741 words. The vocabulary size of this training corpus is 1,210,867 terms

3 Conclusion

We have described our participation in CLEF eHealth 2016 for Task3. Our purpose was to investigate the effectiveness of the word embedding for query expansion on consumer health search, as well as the effect of the learning resource for learning on the results. Our system was based on Terrier with Dirichlet language model. We applied query expansion on two different training sets using word embedding sources.

References

1. Mohannad ALMasri, Catherine Berrut, and Jean-Pierre Chevallet. A comparison of deep learning based query expansion with pseudo-relevance feedback and mutual information. In *Advances in Information Retrieval*, pages 709–715. Springer, 2016.

2. Liadh Kelly, Lorraine Goeuriot, Hanna Suominen, Aurélie Névéol, Joao Palotti, and Guido Zuccon. Overview of the clef ehealth evaluation lab 2016. In *CLEF 2016 - 7th Conference and Labs of the Evaluation Forum, Lecture Notes in Computer Science (LNCS)*. Springer, 2016.
3. Joao Palotti, Guido Zuccon, Lorraine Goeuriot, Liadh Kelly, Allan Hanbury, Gareth Jones, Mihai Lupu, and Pavel Pecina. Clef ehealth evaluation lab 2015, task 2: Retrieving information about medical symptoms. In *CLEF 2015 Working notes*, 2015.
4. Guido Zuccon, Joao Palotti, Lorraine Goeuriot, Liadh Kelly, Mihai Lupu, Pavel Pecina, Henning Mueller, Julie Budaher, and Anthony Deacon. 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*, 2016.