SINAI at CLEF eHealth 2017 Task 3

Manuel Carlos Díaz-Galiano, M. Teresa Martín-Valdivia, Salud María Jiménez-Zafra, Alberto Andreu, and L. Alfonso Ureña López

Department of Computer Science, Universidad de Jaén
Campus Las Lagunillas, E-23071, Jaén, Spain
{mcdiaz,maite,sjzafra,aandreu,laurena}@ujaen.es

Abstract. In this paper we present our participation as SINAI research group from the University of Jaén at Task3 Patient-Centred Information Retrieval. Although only two runs are allowed to be submitted, we have tried several strategies using different models and parameters in order to check the effectiveness of our system. The main 3 approaches try to apply query feedback using MeSH expansion, search engine Google and a Word2Vec model over the Wikipedia. Finally, we have sent two runs in the ad-hoc task. The first one uses Google and the second one applies Word2Vec using the pages related with Health extracted from the Wikipedia.

1 Introduction

The Internet is an important source of health information not only for medical professionals but also for patients or traditional users. Everyday more and more, users are searching for medical information. However, the terminology and the understanding of professional and non-professional users are very different. In this paper, we describe our participation in CLEF eHealth 2017 Task 3: Patient-Centered Information Retrieval [6]. The CLEF eHealth lab aims to evaluate the effectiveness of information retrieval systems when searching for health content on the web. From 2013 the share task eHealth organized by the CLEF [7,5,3,8,4] is focused on studying the medical information retrieval but from the patient point of view, assuming that this kind of user has more difficulties understanding documents in the health domain. In 2017, the CLEF eHealth task 3 Patient-Centred Information Retrieval, continues to focus on evaluating the effectiveness of information retrieval system on the Web [4], the topics provided by the organizers are also the same as those of 2016, with the aim of improving the relevance assessment pool and the collection reusability.

The 2016 topics were developed by mining health web forums where users were seeking advice about specific symptoms, diagnosis, conditions or treatments. For each forum post a set of 6 query variants were generated, representing different ways to express the same information need.

2 Method

In this section, we present the different strategies that we have followed in our participation in CLEF eHealth 2017 Task 3 Patient- Centred Information Retrieval: IRTask 1 Ad-hoc search.

2.1 System description

Although our research group SINAI has a large experience participating in several tasks of other editions of CLEF, mainly in ImageCLEFmed [2], this is the first time that we participate in CLEF eHealth. We have tried 3 main approaches, all of them focused on the integration of external knowledge in order to enrich the query:

- Including terms extracted from MeSH.
- Including information retrieved from Google.
- Including terms extracted from the Wikipedia.

2.2 Preprocessing and indexing

We have used ClueWeb12 B13 corpus[1] and the Lemur IR System[2]. Specifically, we have used Indri search engine for indexing with several default parameters: preprocessing deleting stopwords and stemming words with krovezt algorithm. In addition, we have used Dirichlet prior retrieval method with $\mu = 2500$.

For the queries, we have also applied stopword removal and krovezt stemmer.

3 Experiments

3.1 MeSH approach

Our first approach was to apply the query expansion strategy using MeSH that we have used in other CLEF task in previous years [1]. The main goal is to integrate medical knowledge in order to semantically enrich the query. However, when we tested the result with the assessments of 2016, the results were very poor, even worse than the baseline. We think that the main reason for this is because the collection and the queries are written by non-professional of medicine. Thus, we need to integrate other kind of information with more informal writing instead of using the technical terms extracted from MeSH.

[1] http://lemurproject.org/clueweb12/
[2] http://lemurproject.org/

3.2 Google approach

Since the collection and queries are designed to simulate a typical searching on the web, we have tried to integrate the knowledge from the most popular web search engine, i.e., Google. Thus, we first launch a query on Google and then, we have accomplished experiments with different parameters:

- Replace the query by the titles of the top X retrieved documents, with X ={1,2,3,4,5,10}
- Replace the query by the snippets of the top X retrieved documents, with X ={1,2,3,4,5,10}
- Replace the query by the titles and snippets of the top X retrieved documents, with X ={1,2,3,4,5,10}
- Include in the query the titles of the top X retrieved documents, with X ={1,2,3,4,5,10} plus the original query
- Include in the query the snippets of the top X retrieved documents, with X ={1,2,3,4,5,10} plus the original query
- Include in the query the titles and snippets of the top X retrieved documents, with X ={1,2,3,4,5,10} plus the original query

We have evaluated the experiments using the 2016 relevance assessments and the results are a bit better than the MeSH approach although only the experiments including the information of 5 and 10 documents overcome the baseline. It is worth to mention that the inclusion of the original query always improves the results. Of course, run time is increasing as the number of documents increases, being the experiment with 10 documents the slowest one. Anyway, we have selected the experiment including the titles and snippets of the top 10 retrieved documents plus the original query because is the one with higher precision.

3.3 Wikipedia approach

Finally, we have integrated information by including word vectors obtained from Word2Vec. We have applied two different approaches to create our model, one using the whole Wikipedia (Wikipedia-All) and another one using only pages related to health categories from the Wikipedia (Wikipedia-Health). To create Wikipedia-Health we have obtained all the pages included in the Health category and subcategories. For subcategories we have gone down four levels. We have downloaded a total of 80,765 pages from 13,279 categories.

To expand the original query we calculate a vector for each word in the query. Next, we find the centroid of these vectors calculating the average vector. Finally, we obtain the words whose vectors are near to the centroid. We use the proximity value as weight for this word in the expansion.

Although usually the Word2vec models work better as more documents are included, in this case, the Wikipedia-Health seems that is more efficient than the whole Wikipedia. In our case, evaluating with the 2016 assessments, both approaches slighted overcome the baseline although the Wikipedia-Health works a bit better. For these reasons, we have selected this last experiment to be summited to CLEF eHealth 2017.

3.4 Results

After running all the experiments described in the previous section, we select only two of them in order to be presented in the CLEF eHealth 2017. We have selected the best one from the Google approach and the best one from the Wikipedia approach.

- SINAI-Run1: experiment including the titles and snippets of the top 10 retrieved documents plus the original query.
- SINAI-Run2: experiment including one word got using word2vec model generated from the Health-Wikipedia.

Unfortunately, assessments and official results will be released before the conference, thus we can not include our system evaluation.

Table 1 shows results obtained with 2016 relevance judgments.

Table 1. MAP and R-Prec with 2016 relevance judgments.

Runs	MAP	R-Prec
Base line	0.0862	0.1292
SINAI-Run1	0.1330	0.1786
SINAI-Run2	0.0892	0.1311

Acknowledgments

This work has been partially supported by a grant from the Ministerio de Educación Cultura y Deporte (MECD - scholarship FPU014/00983), Fondo Europeo de Desarrollo Regional (FEDER) and REDES project (TIN2015-65136-C2-1-R) from the Spanish Government.

References

1. Díaz-Galiano, M.C., García-Cumbreras, M., Martín-Valdivia, M.T., Montejo-Ráez, A., Urena-López, L.: Integrating mesh ontology to improve medical information retrieval. In: Workshop of the Cross-Language Evaluation Forum for European Languages. pp. 601–606. Springer (2007)
2. Díaz-Galiano, M.C., García-Cumbreras, M., Martín-Valdivia, M., Urena-López, L., Montejo-Ráez, A.: SINAI at ImageCLEFmed 2008. In: Peters, C., Ferro, N. (eds.) CLEF (Working Notes). CEUR Workshop Proceedings, vol. 1174. CEUR-WS.org (2008)
3. Goeuriot, L., Kelly, L., Suominen, H., Hanlen, L., Névéol, A., Grouin, C., Palotti, J., Zuccon, G.: Overview of the clef ehealth evaluation lab 2015. In: International Conference of the Cross-Language Evaluation Forum for European Languages. pp. 429–443. Springer (2015)

4. Goeuriot, L., Kelly, L., Suominen, H., Névéol, A., Robert, A., Kanoulas, E., Spijker, R., Palotti, J., Zuccon, G.: Clef 2017 ehealth evaluation lab overview. In: CLEF 2017 - 8th Conference and Labs of the Evaluation Forum. Springer (2017)
5. Kelly, L., Goeuriot, L., Suominen, H., Schreck, T., Leroy, G., Mowery, D.L., Velupillai, S., Chapman, W.W., Martinez, D., Zuccon, G., et al.: Overview of the share/clef ehealth evaluation lab 2014. In: International Conference of the Cross-Language Evaluation Forum for European Languages. pp. 172–191. Springer (2014)
6. Palotti, J., Zuccon, G., Jimmy, L., 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)
7. Suominen, H., Salanterä, S., Velupillai, S., Chapman, W.W., Savova, G., Elhadad, N., Pradhan, S., South, B.R., Mowery, D.L., Jones, G.J., et al.: Overview of the share/clef ehealth evaluation lab 2013. In: International Conference of the Cross-Language Evaluation Forum for European Languages. pp. 212–231. Springer (2013)
8. 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-Conference and Labs of the Evaluation Forum. vol. 1609, pp. 15–27 (2016)