SINAI at CLEF eHealth 2018 Task 3. Using cTAKES to remove noise from expanding queries with Google

Manuel Carlos Díaz-Galiano, Pilar López-Úbeda, Maria-Teresa Martin-Valdivia, and L. Alfonso Ureña-López

Department of Computer Science, Advanced Studies Center in ICT (CEATIC)
Universidad de Jaén, Campus Las Lagunillas, 23071, Jaén, Spain
{mcdiaz,plubeda,maite,laurena}@ujaen.es

Abstract. In this paper we present our participation as SINAI research group from the Universidad de Jaén at Task 3 "Consumer Health Search" specifically in sub-task 1 "Ad-hoc Search". The main objective of the task is to provide relevant information to people seeking health advice on the web. We apply the query expansion technique using the most famous search engine at the moment: Google. We search additional information related to the query using the search engine. We identify the medical concepts in Google results using cTAKES. This recognizer provides UMLS concepts from a given text. In this way, we avoid introducing noise with words that are not related to the user query. Our system improves NDCG@10 measurement by 48% over the previous year. We also significantly reduces the response time of the Information Retrieval System (IRS) by 90% compared to previous years.

Keywords: Retrieval Information, Google expander, Named entity recognition, cTAKES, UMLS

1 Introduction

CLEF 2018 [15] consists of an independent peer-reviewed conference on a broad range of issues in the fields of multilingual and multimodal information access evaluation, and a set of labs and workshops designed to test different aspects of mono and cross-language Information retrieval system.

This year we have participated in the Information Retrieval (IR) task and we will continue exploring the same problems and issues identified in 2014 [9], 2015 [5], 2016 [8] and 2017 CLEF eHealth information retrieval challenges [6]. However, the 2018 task [7] uses a new web corpus and a new set of queries compared to previous years.

This task deals the problem of retrieving relevant information in the biomedicine domain. People are constantly looking for information by querying the most important search engines. With these searches, the user tries to obtain more information about diseases, illnesses, treatments, etc.

Our research group SINAI has a large experience participating in several tasks of other editions of CLEF eHealth. In previous years, we have participated in this competition making use the Google search engine[2,11,10]. We consider that Google provides extra knowledge as it is the most widely used search engine in the world.

For our approach, we use the clinical Text Analysis and Knowledge Extraction System (cTAKES)[1]. cTAKES is an open-source Natural Language Processing (NLP) system for information extraction from electronic medical record clinical free-text [13] including types of clinical named entities mapped to various biomedical terminologies/ontologies such as the Unified Medical Language System (UMLS).

This paper is organized as follows: In the next section, we introduce the resources provided by the organizers. Our approach is described in Section 3. In Section 4 we include the results obtained and finally, we expose the conclusions and future work.

2 Resources

2.1 Dataset

For task 3 of CLEF 2018 the corpus of documents consists of web passages acquired from CommonCrawl[2]. For the creation, an initial list of websites was selected, this list was constructed by sending the queries to Microsoft Bing APIs and discarding the unreliable websites.

The collection consists of 1,903 domains including some of the most famous such as *facebook.com*, *answer.com* and *health.com*. Each web page is in the original format as crawled from CommonCrawl, thus it may be html, xhtml, xml, etc. It is important to have a global vision for this type of task, as each website can provide relevant information from every point of view.

2.2 Indri index

The organization also provided several indexes created due to the large size of the collection because there are groups that do not have the possibility of dealing with a corpus of this size. This makes it easier to compare the systems between the participants.

After uncompressed, each document of the collection was extended with the traditional TREC format following the Figure 1.

The collection was indexed as shown in the Figure 2.

[1] http://ctakes.apache.org/ (last visited: May 31, 2018)

[2] http://commoncrawl.org/2018/03/february-2018-crawl-archive-now-available/ (last visited: May 31, 2018)

```
<doc> <docno> DOC ID </docno> ORIGINAL CONTENT <doc>
```

Fig. 1. TREC format.

```
<parameters>
    <indexCount>3</indexCount>
    <indexes>
        <index>2</index>
    </indexes>
    <injectURL>true</injectURL>
    <normalize>true</normalize>
    <stemmer>
        <name>krovetz</name>
    </stemmer>
</parameters>
```

Fig. 2. Configuration index.

2.3 Queries

The query set for 2018 consists of 50 queries issued by the general public to the HON[3] (Health On the Net) and TRIP[4] search services. The queries and the process to obtain them are described in [4]. Queries are formatted one per line in the tab-separated query file, with the first string being the query id, and the second string being the query text. The queries for the English language follow the structure of Figure 3.

```
<query>
    <id> 195001 </id>
    <en> affective treatments for chronic lyme disease </en>
</query>
```

Fig. 3. Example of a original query.

3 Methodology

For this task of the CLEF, the SINAI group has wanted to continue in the same line of the previous year. We got our best results using Google query expansion.

[3] https://hon.ch/en/ (last visited: May 31, 2018)

[4] https://www.tripdatabase.com/ (last visited: May 31, 2018)

We found out that entering everything provided by Google, it is included a lot of noise to the final query and for this reason, we decided to use a biomedical entity recognizer.

The Figure 4 detail the procedure followed. The IRS (explained in section 2.2) uses an information retrieval model described in [14].

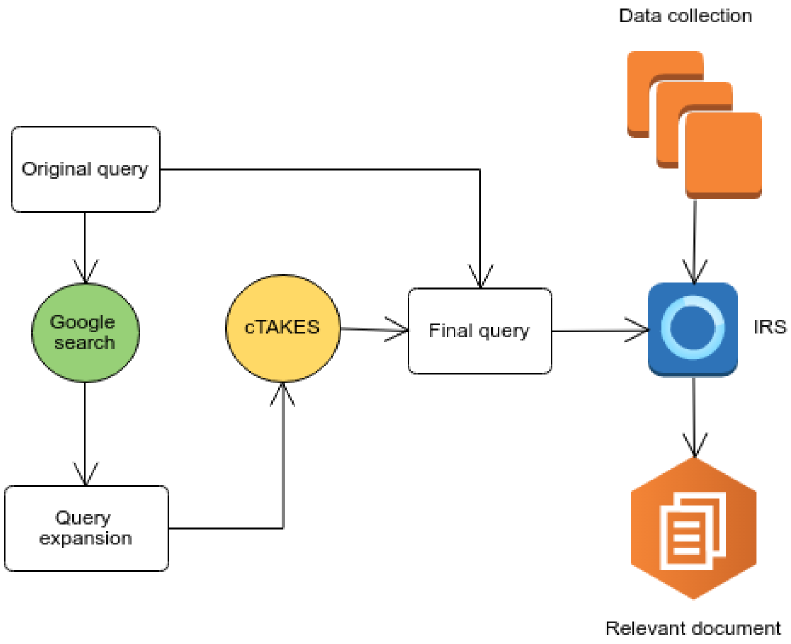

Fig. 4. Developed system architecture.

3.1 Google expansion

The new document collection contains web pages from different websites. We consider a good idea to simulated a typical web search and so, we have tried to integrate the knowledge from the most popular web search engine: Google.

We have performed a search for each query using the Google API for the extraction of titles and snippets. The query was cleared of punctuation marks. A total of 10 results have been taken into account and added to the query the titles and snippets of those results.

The average word count of the expanded query with 10 titles and 10 snippets is 330 words, out of which 29% are stop-words.

3.2 cTAKES: recogniser of medical entities

To select the terms used to expand the query, we use cTAKES, a tool for the recognition of medical entities. This system identifies UMLS concepts in the query expanded by Google. We create a new query only with terms detected with cTAKES.

To build the final query, the concepts returned by cTAKES will be assigned a weight according to the number of occurrences obtained in Google, and finally, we add those words existing in the original query and that have not been detected with cTAKES with a fixed weight W, thus informing the system that these words are also relevant to the query. In this case, the stop-words of original query are not included.

To determine the value of W, several experiments have been performed using the 2017 relevant judgments and queries. In these experiments we observe that the most optimal value is equal to the number of documents selected in Google. In our case $W = 10$.

Figure 5 shows the result of expanding the query following the process in Figure 3. We have used the Indri query language[5], where `#combine` allows us to use multiple words as a unique term and the `#weight` operator assign varying weights to each expressions.

```
<query>
  <type>indri</type>
  <number>195001</number>
  <text>#weight(
    1 #combine(bacteria) 15 #combine(treatment)
    4 #combine(treatments) 1 #combine(microscopy)
    1 #combine(direct microscopy) 1 #combine(test)
    1 #combine(tips) 1 #combine(prevention)
    3 #combine(therapy) 3 #combine(oxygen therapy)
    3 #combine(hyperbaric oxygen therapy) 1 #combine(sick)
    1 #combine(diagnosis) 1 #combine(illness)
    20 #combine(disease) 20 #combine(lyme disease)
    1 #combine(infection) 1 #combine(fibromyalgia)
    1 #combine(conditions) 5 #combine(antibiotics)
    6 #combine(antibiotic) 2 #combine(rise)
    3 #combine(oxygen)  10 #combine(affective)
    10 #combine(chronic) 10 #combine(lyme)
    )
  </text>
</query>
```

Fig. 5. Example of a expanded query.

[5] https://www.lemurproject.org/lemur/IndriQueryLanguage.php

4 Results

This year, we have a new corpus for this task, therefore, we do not have the results obtained for our system. But in the Table 4 we can see the results of the experiments carried out in this year using the 2017 collection and the 2017 relevant judgments.

Run	NDCG@10	BPref	RBP 0.8
Baseline	0.1343	0.0926	0.0070
Google	0.1730	0.1627	0.1471
Google + cTAKES	0.2577	0.1445	0.1139

Table 1. NDCG@10, BPref and RBP with 2017 relevance judgments

Compared to the experiment of the last year (*Google*) where we didn't use cTAKES to filter, this new experiment (*Google + cTAKES*) improves NDCG@10 value by almost 48%, whereas the BPref and RBP values only decreases almost 11% and 22% respectively. We are analyzing the obtain results to understand why these values are lower.

In addition, the time required by the IRS to process the query is significantly reduced, we improve in 90% of the total time invested. A query with Google expansion without cTAKES filter contains 390 words on average, of which 23% are stop-words. However, a filtered query with cTAKES contains about 30 weighted terms.

5 Conclusion

In this paper, we have presented a method to expand queries with the most popular search engine at the moment, adding more information to the user's query. We have observed that this method introduces a lot of noise into the final query, so we have tried to filter and keep the words most related to the biomedical domain.

For this reason, we use the medical entity recognizer cTAKES to filter Google results and we obtain a final query focused on the biomedical domain, and we also use the weighted words, giving more importance to some identified terms.

In the future, we will continue this work to improve our systems by adding knowledge to them. We will study different medical ontologies for the expansion of queries [3,1]. Through these ontologies we will be able to extract knowledge to get closer to the user needs. Besides, we will use disambiguation algorithms to select the most appropriate biomedical terms for the query, using UMLS graph similar to the work presented in [12].

Acknowledgments

This work has been partially supported by a grant from Fondo Europeo de Desarrollo Regional (FEDER) and REDES project (TIN2015-65136-C2-1-R) from the Spanish Government.

References

1. Chen, P., Verma, R.: A query-based medical information summarization system using ontology knowledge. In: Computer-Based Medical Systems, 2006. CBMS 2006. 19th IEEE International Symposium on. pp. 37–42. IEEE (2006)
2. Díaz-Galiano, M.C., Martín-Valdivia, M.T., Jiménez-Zafra, S.M., Andreu, A., Ureña-López, L.A.: SINAI at CLEF eHealth 2017 Task 3 (2017)
3. Díaz-Galiano, M.C., Martín-Valdivia, M.T., Ureña-López, L.: Query expansion with a medical ontology to improve a multimodal information retrieval system. Computers in biology and medicine 39(4), 396–403 (2009)
4. Goeuriot, L., Hanbury, A., Hegarty, B., Hodmon, J., Kelly, L., Kriewel, S., Lupu, M., Markonis, D., Pecina, P., Schneller, P.: Meta-analysis of the second phase of empirical and user-centered evaluations (2014)
5. 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)
6. 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: International Conference of the Cross-Language Evaluation Forum for European Languages. pp. 291–303. Springer (2017)
7. Jimmy, Zuccon, G., Palotti, J., Goeuriot, L., Kelly, L.: Overview of the clef 2018 consumer health search task. In: CLEF 2018 Evaluation Labs and Workshop: Online Working Notes, CEUR-WS (2018)
8. Kelly, L., Goeuriot, L., Suominen, H., Névéol, A., Palotti, J., Zuccon, G.: Overview of the clef ehealth evaluation lab 2016. In: International Conference of the Cross-Language Evaluation Forum for European Languages. pp. 255–266. Springer (2016)
9. 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)
10. Martínez-Santiago, F., Montejo-Ráez, A., García-Cumbreras, M.: Sinai at clef adhoc robust track 2007: applying google search engine for robust cross-lingual retrieval. In: Workshop of the Cross-Language Evaluation Forum for European Languages. pp. 137–142. Springer (2007)
11. Martinez-Santiago, F., Montejo-Ráez, A., García-Cumbreras, M.Á., Urena-López, L.A.: Sinai at clef 2006 ad hoc robust multilingual track: query expansion using the google search engine. In: Workshop of the Cross-Language Evaluation Forum for European Languages. pp. 119–126. Springer (2006)
12. Montejo-Ráez, A., Martínez-Cámara, E., Martín-Valdivia, M.T., Ureña-López, L.A.: Ranked wordnet graph for sentiment polarity classification in twitter. Computer Speech & Language 28(1), 93 – 107 (2014), http://www.sciencedirect.com/science/article/pii/S0885230813000284

13. Savova, G.K., Masanz, J.J., Ogren, P.V., Zheng, J., Sohn, S., Kipper-Schuler, K.C., Chute, C.G.: Mayo clinical text analysis and knowledge extraction system (ctakes): architecture, component evaluation and applications. Journal of the American Medical Informatics Association 17(5), 507–513 (2010)
14. 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)
15. Suominen, H., Kelly, L., Goeuriot, L., Kanoulas, E., Azzopardi, L., Spijker, R., Li, D., Névéol, A., Ramadier, L., Robert, A., Palotti, J., Jimmy, Zuccon, G.: 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 (2018)