LIMSI @ CLEF eHealth 2015 - task 2

Eva D'hondt, Brigitte Grau, and Pierre Zweigenbaum

Laboratoire d'Informatique pour la Mécanique et les Sciences de l'Ingénieur
(LIMSI-CNRS 3251),
Rue John von Neumann , 91400 Orsay, France
eva.dhondt@limsi.fr, bg@limsi.fr, pz@limsi.fr

Abstract. This paper presents LIMSI's participation in the User-Centered Health Information Retrieval task (task 2) at the CLEF eHealth 2015 workshop[5]. In our contribution we explored two different strategies to query expansion, i.e. one based on entity recognition using MetaMap[1] and the UMLS[3], and a second strategy based on disease hypothesis generation using self-constructed external resources such a corpus of Wikipedia pages describing diseases and conditions, and webpages from the MedlinePlus health portal. Our best-scoring run was a weighed UMLS-based run which put emphasis on incorporating signs and symptoms recognized in the topic text by MetaMap. This run achieved a P@10 score of 0.262 and nDCG@10 of 0.196, respectively.

Keywords: information retrieval, Metamap, information extraction

1 Introduction

The 2015 retrieval task in CLEF eHealth competition[5] focused on the retrieval of health information by lay people. This is an interesting research problem which has gathered a lot of interest from the research community over the last five years. A survey conducted by [4] shows that 80% of internet users in the U.S. look online for health information, most frequently for information about a specific disease or medical problem. General purpose search engines are ill-equipped to deal with this particular type of search[8], however, which may lead to erroneous self-diagnosis and self-treatment[2].

The main problem lies in a vocabulary mismatch between the user and the writers in the corpus: Lay people generally do not know the terminology to accurately name or describe the symptoms or conditions they want to get information on. Instead they use long, descriptive and often highly ambiguous queries, in the hope of finding a relevant website that may show them additional keywords which can be used to further refine the search. Even if such information is found, however, the user does not necessarily recognize its importance. He or she may also add erroneous terms which will lead to inaccurate retrieval results. An additional problem is that of the 'informativeness' of the retrieved documents: A retrieval system may return medical documents that are highly pertinent to the user's information need but since he does not master the medical vocabulary, they are nevertheless useless to him in practice. A retrieval system that is adapted to this task should have the following properties: (a) resolve ambiguities in the circumlocutory queries, (b) bridge the terminology gap between query and documents and (c) incorporate a user-centric mesure of informativeness when presenting the retrieved documents to the user.

In this paper we present our participation to the 2015 health information retrieval track. As this is our first contribution to this type of track, we only explored baseline approaches for content retrieval and did not incorporate a user-centric measure in our retrieval system.

2 Task and Corpus Description of CLEF eHealth Task 2

As part of the CLEF eHealth 2015 track [5] task 2 focused on retrieval of web pages containing health information by non-expert users, i.e. lay man. The organizer made a corpus of 1.102.120 web pages of medical and health-related websites available to the participants. These pages were crawled on the web as part of the khresmoi project[1]. The types of web pages in the corpus are quite varied: They include forum threads from health fora, pages from web shops for health-related products as well as high-quality health information sites such as MedlinePlus, ... The crawled documents were provided in their raw HTML (Hyper Text Markup Language) format along with their uniform resource locators (URL).

Given a circumlocutory query, i.e. long, ambiguous topic query written by a lay person, the retrieval system had to return a maximum of 1000 web pages that were relevant to the users information need. Topics have been created using the process described in [7]: Lay people were asked to write a descriptive query of a sign and/or symptom of a medical condition that was shown in a picture or video. Participants were provided with 5 training topics and 66 topics for test phase. The queries were available in multiple languages, such as Arabic, Czech, French, German, Farsi and Portuguese. We opted to use only the English queries.

3 System Components

3.1 Preprocessing

As the web pages in the corpus came from various sources, their quality varied greatly. We built our own HTML parser to extract the following information from a given web page:

- `title`: We extracted the title from the title tag in the head block of the web page. Even though this is a required element (according to the W3C rules), a small subset of web pages in the corpus had no titles.

[1] http://www.khresmoi.eu/

- description: We extracted the description from the description meta tag (<meta name="description" content="...">) in the head block of the web page. The description of a website is the little text excerpt which is shown to users in search engine rankings. While it is an optional element in a web page, good quality websites will always have a short description text.
- keywords: Like the description tag, the content of this meta tag is useful for search engine rankings and may list a few keywords to summarize the content of a page. It is now generally considered obsolete and mostly used to detect spam so this field is often left unfilled.
- content: We also extracted the body of the web page and performed boilerplate and duplicate removal using the Justext Python package developed by [6]. The resulting strings contains free running text as well as title strings and the contents of tables from the original web page.

3.2 Indexing

After preprocessing the corpus was indexed using the Apache Solr toolkit[2]. We indexed the following fields separately:

- title: Text extracted from corresponding tag
- description: Text extracted from corresponding tag
- keywords: Words extracted from corresponding tag
- content: Text extracted from body of the web page
- text: A catchall field that combines the information in the four previous fields.
- URL: URL of the web page, both in full form and its base name
- UID: Unique identifier of the web page which serves as the primary key in the index

The text in the title, description and content fields was indexed with a customized Solr fieldType which incorporated intermediary steps such as tokenization, stopword removal, normalization of English possesives and stemming using the Hunspell stemmer [3]. The latter was to English, and we did not employ stemmers for different languages.

3.3 Annotating with MetaMap

Our first strategy was to experiment with direct translations of circumlocutory terms into their medical variants in order to have a well-controlled query expansion based on the UMLS. By allowing both medical terms, e.g. 'Epidermis' and their common language counterparts, e.g. 'skin' we aim to increase the coverage of the query in the corpus. Recognition of the medical terms and entities was

[2] http://lucene.apache.org/solr/

[3] Can be found at https://www.elastic.co/guide/en/elasticsearch/guide/master/hunspell.html

done using MetaMap [1], a state-of-the-art Named Entity Recognizer for medical English. After visual inspection of the training data, we configured MetaMap to return only entities from a limited set of Semantic Types (such as body part, age group, etc.). This filtering set was constructed manually after analysis of the entities found for the training queries.

The identified entities are likely to be more important for the retrieval process than other information in the topic query. We experimented with several weighing schemes in Solr in which information from MetaMap was combined with the original query terms in different set-ups. We used the two weighing schemes that scored the highest on training data for the official runs, i.e. Run2 and Run3. Table 1 shows the different query construction schemes for the final runs.

3.4 Generation of Disease Hypotheses

For our second strategy we aimed to go directly from the circumlocutory query to the name of the disease or condition that the user wanted to find, and use these terms to expand the original query to search in the Khresmoi corpus. The rationale behind this strategy was to search for diseases or conditions in a smaller, more controlled and consistent corpus (rather than the more varied Khresmoi web corpus). To this end we gathered two small corpora of (web) pages describing the most frequent diseases in clear and common language (so as to match the language use in the query topics). We selected a corpus of 2496 articles from Wikipedia, and another corpus of 1796 web pages from MedlinePlus[4]. Each document describes exactly one disease or condition, and the two corpora overlapped, i.e. all diseases in the MedlinePlus corpus has a corresponding page in the Wikipedia corpus. These two corpora were chosen with an eye on their language use: The webpages in MedlinePlus are written in a very accessible style and contain a lot of descriptions of signs and symptoms. Wikipedia articles are more varied in style and are more likely to contain medical terms with (additional) explanations in lay language.

For each corpus we built a separate index in Solr, and then used the topics to query the wikipedia index (run 4) and both the wikipedia and MedlinePlus indices (run5). For each topic we extracted the top 3 documents, and used their title i.e. disease name as query expansion terms for the original query with which we then searched the Khresmoi corpus. Table 1 shows how the original query terms and the terms for the disease hypotheses were combined.

4 Submitted Runs

We submitted the following 5 runs for official evaluation. Table 1 summarizes the query construction methods and resources used for each run.

[4] http://www.nlm.nih.gov/medlineplus/

- **LIMSI_EN_Run1**: A baseline bag-of-words run. In this run the processed words of the query are searched for in the title, description. keywords and content fields of the documents in the Khresmoi corpus. All fields have equal weight.
- **LIMSI_EN_Run2**: A weighed run with emphasis on terms that have been recognized as agegroup and bodypart Semantic Types. The topics were processed with MetaMap to identify a body part terms (UMLS Semantic Types 'bdsy', 'blor', 'bpoc', 'bsoj') and agegroup terms (UMLS Semantic Type 'aggp') in the topic query texts. If terms with these Semantic Types were recognized, the recognized string and preferredString (from UMLS) were added to the query and given extra weight (ˆ1.5) in the keyword field.
- **LIMSI_EN_Run3**: A weighed run with emphasis on terms of agegroup, body parts, symptoms and diseases. The topics were processed with MetaMap to identify a body part terms (UMLS Semantic Types 'bdsy', 'blor', 'bpoc', 'bsoj'), terms that denote an age group (UMLS Semantic Type 'aggp') and terms that denote symptoms (UMLS Semantic Type 'sosy') and disorders (UMLS Semantic Type 'dsyn'). For the recognized body part and age group terms, the recognized string and preferredString (extracted from UMLS) are added to the query and given extra weight (ˆ1.5) in the keywords field. The recognized string and preferredStrings from recognized signs and disorders were also added to the query and given a higher weight in the keywords field (ˆ1.5) and title field (ˆ2).
- **LIMSI_EN_Run4**: In this run we used the Wikipedia corpus as an external source to generate disease hypotheses for a given topic. These disease names were then added to the query with a weight relative to their retrieval score. The terms were added to all fields but given extra weight (ˆ2) when found in the keywords, description and/or title fields.
- **LIMSI_EN_Run5**: This run is similar to run 4 but the disease hypotheses were retrieved on a corpus that consisted of 2496 Wikipedia pages and 1796 MedlinePlus webpages.

The terminology in Table 1 is as follows: 'origQuery' refers to the (processed) words in the original query; 'bodypartEntities', 'agegroupEntities', ... refer to the recognized strings and preferredNames of UMLS concepts that were recognized by MetaMap (per UMLS Semantic Type); 'diseaseHypotheses' refers to the strings of the diseases and conditions that were retrieved in the Wikipedia (and MedlinePlus) corpora.

5 Results

Table 2 shows the official results we achieved for the five different runs on the test set. Please note that only the top 10 retrieved documents from the run 1 to 3 were used in the evaluation pool. Overall, we find that the best-scoring system was that of run3, which combined term weighing and query expansion based on MetaMap and the UMLS. Run 4 and 5 which both used external sources to

Table 1. Query construction and resources used for submitted runs.

Run name	Query construction	Resources used
LIMSI_EN_Run1	title:{origQuery} + content:{origQuery} + description:{origQuery} + keywords:{origQuery}	-
LIMSI_EN_Run2	title:{origQuery} + content:{origQuery} + description:{origQuery} + keywords:{origQuery}+ keywords^1.5:{bodypartEntities & agegroupEntities}	UMLS
LIMSI_EN_Run3	title:{origQuery} + content:{origQuery} + description:{origQuery} + keywords:{origQuery}+ keywords^1.5:{bodypartEntities & agegroupEntities} + keywords^1.5:{symptomEntities & disorderEntities} + title^2:{symptomEntities & disorderEntities}	UMLS
LIMSI_EN_Run4	title:{origQuery} + content:{origQuery} + description:{origQuery} + keywords:{origQuery} + title:{diseaseHypotheses^retrievalScore} + content:{diseaseHypotheses^retrievalScore} + description:{diseaseHypotheses^retrievalScore} + keywords^2:{diseaseHypotheses^retrievalScore} + title^2:{diseaseHypotheses^retrievalScore} + description^2:{diseaseHypotheses^retrievalScore}	Wikipedia
LIMSI_EN_Run5	title:{origQuery} + content:{origQuery} + description:{origQuery} + keywords:{origQuery} + title:{diseaseHypotheses^retrievalScore} + content:{diseaseHypotheses^retrievalScore} + description:{diseaseHypotheses^retrievalScore} + keywords^2:{diseaseHypotheses^retrievalScore} + title^2:{diseaseHypotheses^retrievalScore} + description^2:{diseaseHypotheses^retrievalScore}	Wikipedia MedlinePlus

generate disease hypotheses performed considerably worse. We show the results in terms of P@10 (primary evaluation criteria), nDCG@10 and readability score.

Table 2. Results of submitted runs

Run name	P@10	nDCG@10	RBP
LIMSI_EN_Run1	0.232	0.180	0.229
LIMSI_EN_Run2	0.230	0.168	0.216
LIMSI_EN_Run3	0.262	0.196	0.242
LIMSI_EN_Run4	0.056	0.038	0.056
LIMSI_EN_Run5	0.056	0.038	0.056

6 Discussion

The scores presented in table 2 are generally quite low: Of all submitted runs, only Run 3 outperformed the baseline (Run 1). Overall, we see that the UMLS-based runs did better than those that incorporated disease hypotheses. Close analysis of the disease hypotheses runs shows that the submitted runs contained a bug in the query construction process which lead to lower performance. We reran Runs 4 and 5 and evaluated them with the relevance assessments released by the track organizers. While these scores are higher, they remain well below the scores attained by Run3.

Table 3. Results of corrected versions of Run 4 and Run 5

Run name	P@10	nDCG@10
LIMSI_EN_Run4_Rerun	0.083	0.061
LIMSI_EN_Run5_Rerun	0.085	0.061

The scores for both disease hypotheses runs are very similar. Analysis of the disease hypotheses that were generated by retrieval in the Wikipedia corpus versus the combination of the Wikipedia and the MedlinePlus corpora, showed that the difference between the lists of retrieved diseases from both corpora were very small. In fact, different hypotheses were retrieved for only 19 out of the 66 queries. These queries did achieve higher scores in run 5 which shows that the added information from the MedlinePlus corpus leads to better hypotheses. As a follow-up to this task, it would be interesting to compare our hypotheses with those that were used to generate the test queries.

When we turn our attention to the UMLS-based runs we see that these remain fairly close to the baseline scores. We believe that the relatively small impact of additional fields could be attributed to uneven distribution of information in the corpus: We found that around 57.4% and 59.2% of documents in the corpus had empty fields for the description and keywords fields, respectively. The other fields were better filled: Only around 1% of the documents had an empty title field, but we do see that about 11% of the pages in the corpus did not have any information in the content field. The latter is mostly due to the different (file) formats that were included in the crawl. Our parser could only deal with HTML documents and did not process other file types such as PDF or PowerPoint files. In total, however, only 7,949 files were completely empty (i.e. no text information in either title, description, keywords and content fields, and therefore unretrievable for any of the topics). None of these files featured in the list of reference files though.

The decreased score for Run2 shows that terms for recognized age group and body part are not informative query expansion terms and including them actually harms performance. We believe that certain age group terms (such as 'child', 'toddler', ...) are too general and their relative higher importance in the query leads to the retrieval of non-relevant articles.

Recognized symptoms and disorders seem much more informative (Run3) query expansion terms, especially in the title field. Intuitively this makes sense: If a (correctly recognized) disease or symptom is referenced in the title of the web page, it will be topically relevant.

7 Conclusion

In this paper we presented our participation to the User-Centered Health Information Retrieval task (task 2) at the CLEF eHealth 2015 workshop. We explored two different baseline strategies to query expansion: One based on Entity Recognition of concepts in the UMLS, and another based on disease hypothesis generation using external sources such as Wikipedia and MedlinePlus. We found that the disease hypotheses approach was not adequate and lead to a serious decrease in performance, which is most likely caused by the relatively small coverage of the corpora. For the UMLS-based runs, we found that performing query expansion with terms describing recognized Disease and Symptom entities, lead to improvements over our baseline. Our best-scoring run was a weighed UMLS-based run which put emphasis on finding terms for signs and symptoms recognized in the topic text in the title and keywords field (with relative weights of 2 and 1.5, respectively). This run achieved a P@10 score of 0.262 and nDCG@10 of 0.196. In this year's participation we limited our systems to a baseline content-based retrieval. In a future participation, we would like to incorporate a method of reranking based on readability, and profit from the multilingual topics that were offered in this years track.

Acknowledgments

This work was supported by the French National Agency for Research under grant Accordys[5] ANR-12-CORD-0007.

References

1. Aronson, A. R. (2001). Effective Mapping of Biomedical Text to the UMLS Metathesaurus: the MetaMap program. In Proceedings of the AMIA Symposium, p. 17. American Medical Informatics Association.
2. Benigeri, M., and Pluye, P. (2003). Shortcomings of Health Information on the Internet. Health promotion international, 18(4), 381-386.
3. Bodenreider, O.(2004). The Unified Medical Language System (UMLS): Integrating Biomedical Terminology. Nucleic acids research 32, no. suppl 1: D267-D270.
4. Fox, S. (2011). Health topics: 80% of Internet Users look for Health Information online. Pew Internet & American Life Project.
5. Goeuriot, L., Kelly, L., Suominen, H., Hanlen, L., Névéol, A., Grouin, C., Palotti, J., and G. Zuccon (2015). Overview of the CLEF eHealth Evaluation Lab 2015" in CLEF 2015 - 6th Conference and Labs of the Evaluation Forum, Lecture Notes in Computer Science (LNCS), Springer.
6. Pomikálek, J. (2011). Removing boilerplate and duplicate content from web corpora. Disertacnı práce, Masarykova univerzita, Fakulta informatiky.
7. Stanton, I., Ieong, S., and Mishra, N. (2014). Circumlocution in Diagnostic Medical Queries. In Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval, pp. 133-142. ACM.
8. Zuccon, G., Koopman, B., and Palotti, J. (2015). Diagnose this if you can: On the effectiveness of search Engines in finding medical self-diagnosis information.

[5] Agrégation de Contenus et de COnnaissances pour Raisonner à partir de cas dans la DYSmorphologie foetale