User intent behind medical queries: An evaluation of entity mapping approaches with Metamap and Freebase

João R. M. Palotti
Vienna University of Technology
Favoritenstrasse 9-11/188,
1040 Vienna, Austria
palotti@ifs.tuwien.ac.at

Veronika Stefanov
Vienna University of Technology
Favoritenstrasse 9-11/188,
1040 Vienna, Austria
stefanov@ifs.tuwien.ac.at

Allan Hanbury
Vienna University of Technology
Favoritenstrasse 9-11/188,
1040 Vienna, Austria
hanbury@ifs.tuwien.ac.at

ABSTRACT

This work focuses on understanding the user intent in the medical domain. The combination of Semantic Web and information retrieval technologies promises a better comprehension of user intents. Mapping queries to entities using Freebase is not novel, but so far only one entity per query could be identified. We overcome this limitation using annotations provided by Metamap. Also, different approaches to map queries to Freebase are explored and evaluated. We propose an indirect evaluation of the mappings, through user intent defined by classes such as Symptoms, Diseases or Treatments. Our experiments show that by using the concepts annotated by Metamap it is possible to improve the accuracy and F1 performances of mappings from queries to Freebase entities.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval - *Search process*; J.3 [**Computer Applications**]: Life and Medical Sciences - *Medical Information Systems*

General Terms

Algorithms, Experimentation

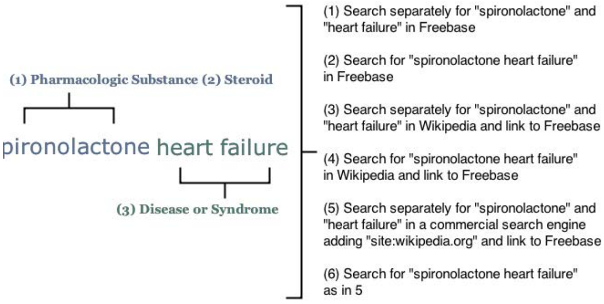
Keywords

Query log analysis, health search, semantic analysis

1. INTRODUCTION AND RELATED WORK

Understanding the user intent is the holy grail of information retrieval research. The use of the Semantic Web to support users in their search activity is relatively new and a promising way to decipher user intent. There are multiple benefits of the fusion of information retrieval (IR) and Semantic Web, both from the user's and the system's perspective ([11], [8]). For example, an IR system could (1) help users to acquire contextual information, (2) suggest related concepts or associated terms, (3) provide navigational suggestion or even (4) improve support for QA queries.

A practical example is shown on the left side of Figure 1. Apart from the already cited benefits, after discovering that a user is searching for a disease and a treatment, an IR system could retrieve the information of a package insert for a lay user, or the main results of clinical trials for an expert. In order to have a system like this, it is necessary to correctly assign the user query to entities or concepts.



(1) Pharmacologic Substance (2) Steroid
spironolactone heart failure
(3) Disease or Syndrome
(1) Search separately for "spironolactone" and "heart failure" in Freebase
(2) Search for "spironolactone heart failure" in Freebase
(3) Search separately for "spironolactone" and "heart failure" in Wikipedia and link to Freebase
(4) Search for "spironolactone heart failure" in Wikipedia and link to Freebase
(5) Search separately for "spironolactone" and "heart failure" in a commercial search engine adding "site:wikipedia.org" and link to Freebase
(6) Search for "spironolactone heart failure" as in 5

Figure 1: Metamap semantic type annotations and the six approaches to match this query to Freebase entities

An approach to map user queries to entities in the Linked Open Data (LOD) cloud using DBpedia was presented by Meij et al. [8]. They proposed a two-step method. In the first step, they generated n-grams for a query (spironolactone heart failure, spironolactone heart, heart failure, spironolactone, heart, failure) and mapped each of these possibilities to Wikipedia concepts ranked by a Language Modeling approach. In the second step, they applied Machine Learning techniques to decide whether each possibility generated in the first step was relevant or not.

A more recent work is Hollink et at. [6]. The authors mapped queries from the movie domain to concepts in DBpedia (and Freebase), but re-issuing the user query into a commercial search engine and adding "site:wikipedia.org" to find the most relevant entity for each query. Hollink's approach has the great advantage of not requiring manual user input (the Machine Learning part of Meij's work), but it is limited to finding only one single entity for a query.

In this paper, we focus on mapping queries in the medical domain, which has received special attention from researchers [12, 3, 5, 15]. We only explain Cartright et al.'s

IIiX '14 August 26 - 29 2014, Regensburg, Germany

http://dx.doi.org/10.1145/2637002.2637043 ...$15.00.

work, which is the most closely related to our work. Cartright et al. [3] presented a log-based study of users' behavior searching for health information online. The authors used handcoded rules to classify user queries into three classes: symptoms, causes and remedies. Thus it is possible to analyze the change of search focus along a session. We expand the three original classes created by Cartright et al. and present various ways to automatically map queries in the medical domain to concepts in the LOD cloud.

The two main contributions of this paper are (1) a comparison of different methods to automatically map user queries to classes in the medical domain and (2) a dataset of 1000 real user queries from different search engines, classified according to the best matching user intent. Our evaluation shows that the use of annotations provided by Metamap improves the performance of mappings created between queries and Freebase entities, as more than one entity can be detected per query.

2. MAPPING QUERIES TO USER INTENTS

In this section, we explain how we employed Metamap and Freebase to map user queries into the user intents.

2.1 Metamap

The US National Library of Medicine's Metamap tool is a well-established tool for mapping biomedical text to concepts in the Unified Medical Language System (UMLS) [1]. The left part of Figure 1 shows an example of annotation provided by Metamap for a query issued by a physician. In this short text, two concepts are detected and annotated with three different semantic types: (1) Pharmacologic Substance, and (2) Steroid for 'spironolactone' and (3) Disease or Syndrome for 'heart failure'. There are 133 possible semantic types grouped into 14 high-level groups such as Anatomy, Disorders, Chemicals & Drugs or Physiology.

Unfortunately, the 14 high-level groups provided by Metamap are not in direct accordance with the existing literature. To be in accordance with Cartright et al.[3], we have to classify queries into the classes: Symptoms, Diseases and Treatment. However, the high-level Metamap group 'Disorders' encompasses both Cartright's 'Symptoms' and 'Diseases' classes. Therefore, a mapping from the Metamap types to the classes is necessary. We decide to expand the original three classes to include two further common classes of queries: Diagnosis and Anatomy. We explain below how we mapped the semantic types provided by Metamap into the five classes (also some query examples are shown):

- **Symptom:** only the type Sign or Symptom (cough; sore; headache; red eyes)
- **Diseases:** all types belonging to the high-level group Disorders, except for Sign or Symptom (Disease or Syndrome (diabetes; heath failure), Mental or Behavioral Dysfunction (addiction, bipolar disorder), Neoplastic Process (lung cancer, tumor), etc.)
- **Treatment:** all types belonging to the high-level group Chemicals & Drugs (Clinical Drug (cough syrup), Antibiotic (penicillin), Pharma. Substance (tylenol), etc.)
- **Anatomy:** all types belonging to the high-level group Anatomy (Body Part, Organ, or Organ Component (head, skin), Body Substance (blood), etc.)
- **Diagnosis and Tests:** only the types Diagnostic Procedure (endoscopy, biopsy), Laboratory Procedure (Blood Test) and Laboratory or Test Result (fsh level)

2.2 Freebase

Metamap is capable of annotating the queries, but there is no link between the annotations and the Linked Open Data, therefore Metamap is unable to provide relations between concepts. In turn, large semantic knowledge bases, such as Wikipedia (DBpedia) or Freebase[1], are designed to provide these relations. In this work, we opt to use Freebase rather than Wikipedia, as preliminary results showed Freebase to be more complete than Wikipedia. Also one important recent work, on which our work is based, used it to map queries from the movie domain to entities [6].

Hollink's [6] approach consists of running each query from their dataset on a large commercial search engine adding a 'site:wikipedia.org' operator to find the Wikipedia page that best matches each query. Then, it takes advantage of the linked web between Wikipedia and Freebase (dbpedia.org provides this information) and indirectly finds the entity represented in the query. This solution has the advantage of a large commercial search engine, such as fixing typos and disambiguation, however up to one entity can be recognized per query. We explore the same approach, as well as two more direct ways to get to the Freebase entity: (1) searching directly in Freebase and (2) searching in Wikipedia. To deal with the limitation of one entity per query, we take advantage of the multiple Metamap annotations for a query, and use them as well. Therefore, instead of searching for the whole query in each one of the systems, we search for the annotations. It results in a total of six Freebase variations, as shown in the right part of Figure 1.

It is important to mention that when a query is issued to the Freebase API, a result list is presented which contains the concept name and type. We used only the first result for which the type belonged to */medicine/*. A more robust approach is left as future work.

Once the Freebase entity is identified, we use the */common/topic/notable_for* property as the type assigned by Freebase to the entity. Finally, a mapping from the Freebase type to the five classes explained in Section 2.1 is easily traced: the class **Symptoms** is made of entities of type */medicine/symptom*, class **Diseases** uses the types related to diseases such as */medicine/disease*, */medicine/disease_cause* or */medicine/infectious_disease*, etc.

3. QUERY LOG DATASETS

We use a variety of search logs from different search engines taking free text queries: two datasets focused on queries by laypeople, one made of queries from medical professionals and one of queries not related to health or medical information. This way we can analyse different scenarios where annotated queries would be desirable.

The query logs assumed to consist almost completely of queries submitted by laypeople were obtained from health-related searches in America Online's search service [10][2] and from the Health of the Net Foundation website (HON[3]).

[1] Freebase is a large collaborative knowledge base containing information of more than 40 million entities and 2 billion facts relating them

[2] Obtained from `http://www.gregsadetsky.com/aol-data/`

[3] `http://www.hon.ch/HONsearch/Patients/index.html`

The AOL logs were obtained from March to May of 2006. We divided them into two non-overlapping sets: AOL-Health and AOL-NotHealth. For this purpose, the click-through information available in the AOL data was used (only 53% of the AOL log entries have this information). For every clicked URL, we checked if that URL was listed in the Open Directory Project (ODP)[4]. The URL could: (1) be found in the Health category[5] (2%); (2) be found in any other category: News, Arts, Games, Health/Animals, etc (68%); (3) not be found (30%). We formed the AOL-Health set using the queries in case (1), AOL-NotHealth using the queries in situation (2), and we ignore the queries in case (3) and when the click-through was not available. Very similar approaches to separating queries are present in the literature [14, 4, 15, 9]. Although ethical concerns have arisen, we opt to use this dataset in the way it was intended to be used, which also allows other researchers to replicate our experiments. All the code used is available online[6].

The HON dataset is composed of anonymous logs ranging from December 2011 to August 2013. HON is a non-governmental organization responsible for the HONcode [2]. They provide a search engine to facilitate the access to the certified sites. The majority of the queries are issued in English, however the use of French or Spanish is very frequent. Aiming to reduce noise, only queries consistent with Unicode block Latin 1 (iso-8859-1) were kept[7].

For medical professionals, we use the logs from the Turning Research Into Practice (TRIP) database[8]. TRIP is a search engine indexing more than 80,000 documents and covering 150 manually selected health resources such as MEDLINE and the Cochrane Library. Its intent is to allow easy access to online evidence-based material for physicians [7]. The logs contain queries of 279,280 anonymous users from January 2011 to August 2012.

4. EVALUATION AND DISCUSSION

From each dataset described in Section 3 we take two samples of queries: the 125 most frequent and 125 random queries, comprising a total evaluation corpus of 1000 user queries. In this corpus, there are 750 medical queries: ranging from very technical ones to barely medical ones, such as laypeople searching for aesthetic surgery; and 250 non-medical queries, which we keep to analyse the number of false positives of each system.

The evaluation process itself is very challenging. For each mapping created from a query to a Freebase entity (or UMLS concept for Metamap), we could evaluate the pair <query, entity>, but there are some issues with this approach. As discussed by Hollink et al. [6], we cannot expect a rater to reliably judge whether or not a better entity exists for each pair. Therefore there is a positive bias, as several different entities could be rated as correct for the same query. As Hollink's dataset was small (only 50 queries) and in a less complex domain (movies), the solution adopted consisted of asking raters to manually create links to LOD concepts, which is considerably more time consuming, but feasible for a small number of queries.

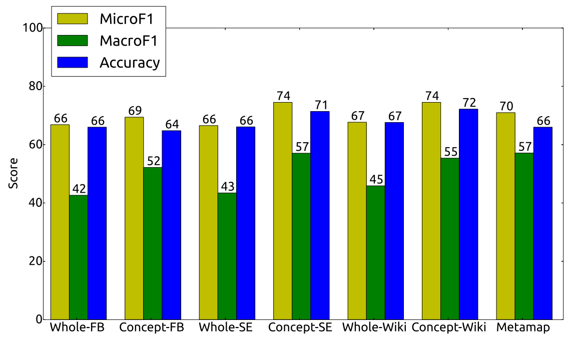


Figure 2: Comparing micro-F1, macro-F1 and accuracy scores of different systems

We opt to indirectly evaluate the mappings, thus using the classes described in Section 2, which represent the user intent when issuing a query. Considering that for some important works in the medical domain ([3, 15], for example), knowing only the classes a query belongs to (symptom, disease or treatment) was sufficient, we evaluate here only the assignment to the classes, using well-known metrics for classification tasks, such as F1 and accuracy.

The two first authors of this paper were responsible for separately classifying the entire evaluation corpus. The divergent cases were argued until a consensus was reached. As a proof step, the third author and two physicians checked the classification output. The raters were instructed to use any resource and assign any number of labels for each query. Following recommendations such as those for the evaluation campaign CLEF eHealth[13], we considered that if an anatomical concept was part of a disease, the query should not be labelled as anatomical. This means that “heart failure” should be assigned only to the class Disease.

Figure 2 compares micro-F1, macro-F1 and accuracy of all system variations: Metamap alone, obtaining Freebase entities through Wikipedia (Wiki), a commercial search engine (SE) or Freebase itself (FB), using the whole query (Whole) or the concepts identified by Metamap (Concept). The best systems in terms of micro-F1 were Concept-SE and Concept-Wiki (both with 74.51%), while the worst system used plain queries to match Freebase concepts through a search engine, Whole-SE (66.54%). This figure also shows that among the three options to match entities to Freebase, using the Freebase API itself was the worst one, confirming what motivated Hollink et al. [6] to use a commercial search engine. To illustrate this, we use the query *uncontrolled nerve pain*. It is annotated by Metamap as *uncontrolled*, a qualifier, and *nerve pain*, a symptom. However, the first two results of Freebase search API are Spillway (ignored as it is not medical) and Cancer. In turn, the search engine approach maps the concepts to the entities Uncontrolled_airspace and Neuropathic_pain, while the Wikipedia API maps them to Uncontrolled and Neuralgia.

One very important finding is that using the concepts annotated by Metamap is highly advisable. It can be seen when comparing the versions Whole and Concept in Figure 2. For some metrics, such as macro-F1, the use of multiple concepts increased the perormance for the search engine system from 43.45% (Whole-SE) to 57.07% (Concept-SE), an improvement of 30%. In another domain, however, an-

[4] dmoz.org

[5] We excluded URLs in the Health/Animals subcategory

[6] `https://sites.google.com/site/joaopalotti/`

[7] The Latin 1 covers the majority of European languages, however it excludes the majority of Asian languages

[8] `http://www.tripdatabase.com/`

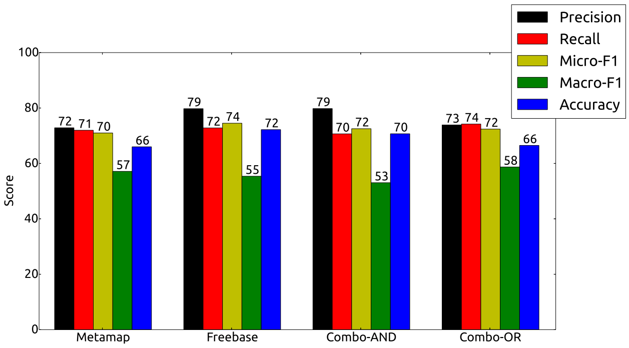


Figure 3: Combining systems: a trade-off between precision and recall

Table 1: Micro-F1 value for each class and system, the number of elements of each class is shown inside the parenthesis

Classes	Metamap	Freebase	Combo AND	Combo OR
Symp.(67)	45.51	50.00	46.32	47.78
Dis.(320)	80.07	77.51	76.72	80.74
Treat.(186)	56.30	59.93	52.31	62.01
Anat.(17)	43.08	41.03	41.03	43.08
Diag.(22)	39.22	22.22	22.22	39.22
None (462)	78.59	81.56	79.65	79.50

other tool would be required to annotate the queries.

We highlight that the Wikipedia API is as good as a commercial search engine for this task, thus being a free and compatible option for academic research.

Combining the output of Metamap and any one of the six Freebase systems is also possible. We show only the results of combining Metamap and Concept-Wiki (called only "Freebase" from now on). Two simple combination approaches are made by the intersection (Combo-AND) and the union (Combo-OR) of the result sets. Figure 3 shows the results. As expected, Combo-AND is a good option when the precision is the most relevant metric, while Combo-OR focuses more on recall. We also report the micro-F1 values for each class in Table 1, as well as the number of elements in each class. Among the medical classes, Diseases was the most frequent and most successfully identified one, while Diagnosis was the worst one. A reason for the poor performance of the class Diagnosis is that many diagnostic procedures are redirected to the related disease, for example, the query *serodiagnosis* is mapped to Syphilis.

5. CONCLUSION AND FUTURE WORK

Identifying concepts in user queries is a key task to better understand user intents. Some effort in this direction was made in the movie/entertainment domain, but issues are still present, such as a maximum of one concept per query. In this work, we evaluated different automatic mapping solutions in the medical domain using Metamap and Freebase. We treated the entity mappings as a classification task and considered the types of the entities detected to classify them. The results showed that using annotations provided by Metamap improved the accuracy and F1 performances of mappings created, as more than one entity could be detected per query. We released the dataset containing 1000 labeled real user queries and the code used in this paper, facilitating future work on the problem. As future work, we will compare our automatically generated results with more costly approaches such as Machine Learning and Language Modeling.

ACKNOWLEDGMENTS.

This research was partly funded by the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement n°257528 (KHRESMOI) and partly funded by the Austrian Science Fund (FWF) project number I1094-N23 (MUCKE).

6. REFERENCES

[1] A. R. Aronson and F.-M. Lang. An overview of MetaMap: historical perspective and recent advances. *JAMIA*, 2010.

[2] C. Boyer, V. Baujard, and A. Geissbuhler. Evolution of Health Web certification through the HONcode experience. *Stud Health Tech Inform*, 2011.

[3] M.-A. Cartright, R. W. White, and E. Horvitz. Intentions and attention in exploratory health search. In *Proc of SIGIR*, 2011.

[4] S. Duarte Torres, D. Hiemstra, and P. Serdyukov. Query log analysis in the context of information retrieval for children. In *Proc. of SIGIR*, July 2010.

[5] R. Islamaj Dogan, G. C. Murray, A. Névéol, and Z. Lu. Understanding PubMed user search behavior through log analysis. *Database*, 2009.

[6] R. B. Laura Hollink, Peter Mika. Web usage mining with semantic analysis. In *Proc. of WWW*, 2013.

[7] E. Meats, J. Brassey, C. Heneghan, and P. Glasziou. Using the Turning Research Into Practice (TRIP) database: how do clinicians really search? *J Med Libr Assoc*, 2007.

[8] E. Meij, M. Bron, L. Hollink, B. Huurnink, and M. de Rijke. Mapping queries to the Linking Open Data cloud: A case study using DBpedia. *J Web Sem.*, 2011.

[9] J. Palotti, A. Hanbury, and H. Müller. Exploiting health related features to infer user expertise in the medical domain. WSCD, 2014.

[10] G. Pass, A. Chowdhury, and C. Torgeson. A picture of search. In *Proc. of InfoScale*. ACM, 2006.

[11] U. Shah, T. Finin, A. Joshi, R. S. Cost, and J. Matfield. Information retrieval on the semantic web. In *CIKM*, 2002.

[12] A. Spink, Y. Yang, J. Jansen, P. Nykanen, D. P. Lorence, S. Ozmutlu, and H. C. Ozmutlu. A study of medical and health queries to web search engines. *HILJ*, 21, Mar. 2004.

[13] H. Suominen, S. Salantera, et al. Overview of the ShARe/CLEF eHealth Evaluation Lab 2013. In *CLEF*, 2013.

[14] R. W. White, S. T. Dumais, and J. Teevan. Characterizing the influence of domain expertise on web search behavior. In *Proc. of WSDM*, 2009.

[15] R. W. White and E. Horvitz. Studies of the onset and persistence of medical concerns in search logs. In *Proc. of SIGIR*. ACM, 2012.