Filtering and reranking using MetaMap named entities recognizer

Pilar López-Úbeda and Manuel Carlos Díaz-Galiano and
Maria-Teresa Martín-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
{plubeda, mcdiaz, maite, laurena}@ujaen.es

Abstract

In this paper we present our participation as SINAI research group from the Universidad of Jaén at *Text REtrieval Conference* (TREC), specifically in sub-task *Precision Medicine*. The main objective of the task is to locate relevant information for a patient using information retrieval technologies. Our group applies one of the techniques of Natural Language Processing: Named Entities Recognition. For this task we have used MetaMap. This recognizer provide UMLS concepts from a given text. In addition, we have applied the document ranking technique to sort the final list of relevant documents using the common concepts found in the query and each document. The results obtained are not as expected because not all the concepts detected by MetaMap are relevant in the queries. However, our results are above the average of the runs sent by participants.

1 Introduction

Over time the information stored digitally in the biomedical domain is growing exponentially and this presents a problem for clinicians, as the large literature available for precision medicine can make it difficult to find the most appropriate treatment for the patient.

In the biomedical domain, the application of *Natural Language Processing* (NLP) techniques helps to create computational mechanisms to facilitate man-machine interrelation through natural language (Friedman et al., 1999).

More specifically, *Information Extraction* (IE) techniques process text to detect explicit information of interest. One of the tasks performed by IE is the Recognition and Classification of Named Entities. These tasks focus first on detecting medical concepts in a text and then assigning a category from a predetermined set. For this task, we use the most extensive and popular biomedical terminology: UMLS (Bodenreider, 2004).

The 2018 TREC Precision Medicine track continues the prior 2017 Precision Medicine track (Roberts et al., 2017), which was a specialization of the previous TREC Clinical Decision Support track. These tasks could greatly help clinicians to find the most up-to-date evidence-based treatment for their patients.

This paper is organized as follows: In Section 2 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.

2 Data collection

There are two collections of document for the Precision Medicine track: scientific abstracts and clinical trials.

2.1 Scientific abstracts

Our system uses the collection of scientific abstracts. These documents are taken from PubMed abstracts of January 2017.

This collection is composed by: **MedLine** (U.S. National Library of Medicine, 2019) that contains journal citations and abstracts for biomedical literature, *The American Association for Cancer Research* (**AACR**) (AACR, 2019) and the *American Society of Clinical Oncology* (**ASCO**) (ASCO, 2019) proceedings.

ASCO and AACR were included with the intention of providing potentially relevant cancer related reports and are not included in MedLine. These are more targeted toward cancer therapy,

and likely to include precision medicine studies not in PubMed.

The organizers provide these documents in XML and TXT format.

2.2 Index

We use Lemur (Croft and Callan, 2016) with the Indri search engine. Indri provides a structured query language for text collections. We have accessed each document and stored in the index the fields of each document that most interest us. These fields are described below. We have cleaned all the documents by eliminating/erasing the HTML tags.

For MedLine journals and abstracts we have taken into account the following fields:

- PMID: the PubMed unique identifier, used as the document ID for TREC submissions.
- Article title: contains the entire title of the journal article.
- Abstract text: the full text of the abstract.
- Keyword: contains terms that describe the content of the article.
- Name of substances: is the name of the substance that carries the MeSH unique identifiers.
- Descriptor name: this attribute has the MeSH unique identifiers for descriptors.

For AACR and ASCO proceedings we have stored the following fields:

- ID: name of file that is used as the document ID for TREC submissions.
- Article title: contains the entire title of the journal article.
- Abstract text: the full text of the abstract.

3 Strategies

In this section, we will describe the strategy followed for the task. Our group will use the MetaMap tool to identify UMLS terminology in the medical domain and subsequently, we perform a ranking of relevant documents to be returned.

3.1 UMLS

The *Unified Medical Language System* (National Library of Medicine, 2019b) is a repository of biomedical vocabularies developed by the US *National Library of Medicine* (NLM).

Our work has used "2017AA Full Release UMLS Metathesaurus", it contains approximately 3.47 million concepts and 13.5 million unique concept names from 201 source vocabularies.

Vocabularies integrated in the UMLS Metathesaurus include the NCBI taxonomy, Gene Ontology, the *Medical Subject Headings* (MeSH), MedDRA, RxNorm, or SNOMED-CT.

In UMLS when a concept is added to the Metathesaurus, it receives a unique identifier named *Concept Unique Identifiers* (CUI). This identifier will be very useful in our system.

3.2 MetaMap

MetaMap (National Library of Medicine, 2019a) is a highly configurable application developed to map biomedical text to the UMLS Metathesaurus (Aronson, 2001).

MetaMap employs NLP and computational linguistic techniques(Aronson and Lang, 2010): tokenization, part-of-speech tagging, syntactic analysis, word sense disambiguation, and others.

This tool first breaks the text into phrases and then, for each phrase, it returns the concepts detected and several other information. Concepts are ranked according to a relevance value. Researchers have used MetaMap for a variety of tasks such as information retrieval (Aronson and Rindflesch, 1997) and molecular binding from biomedical text (Rindflesch et al., 1999).

3.3 Our approach

In Figure 1 we describe the system architecture used for this task. Below, we will detail the steps followed:

1. **Fields of query**: we first treat the query by concatenating three fields: disease, gene and demography.
2. **Normalize**: the queries were converted to lower case and we removed the special characters. Then, the query is rewritten in the format that Lemur/Indri needs.
   Figure 2 shows an example of the original query 1 and its normalized.
   The result of entering these queries in Indri

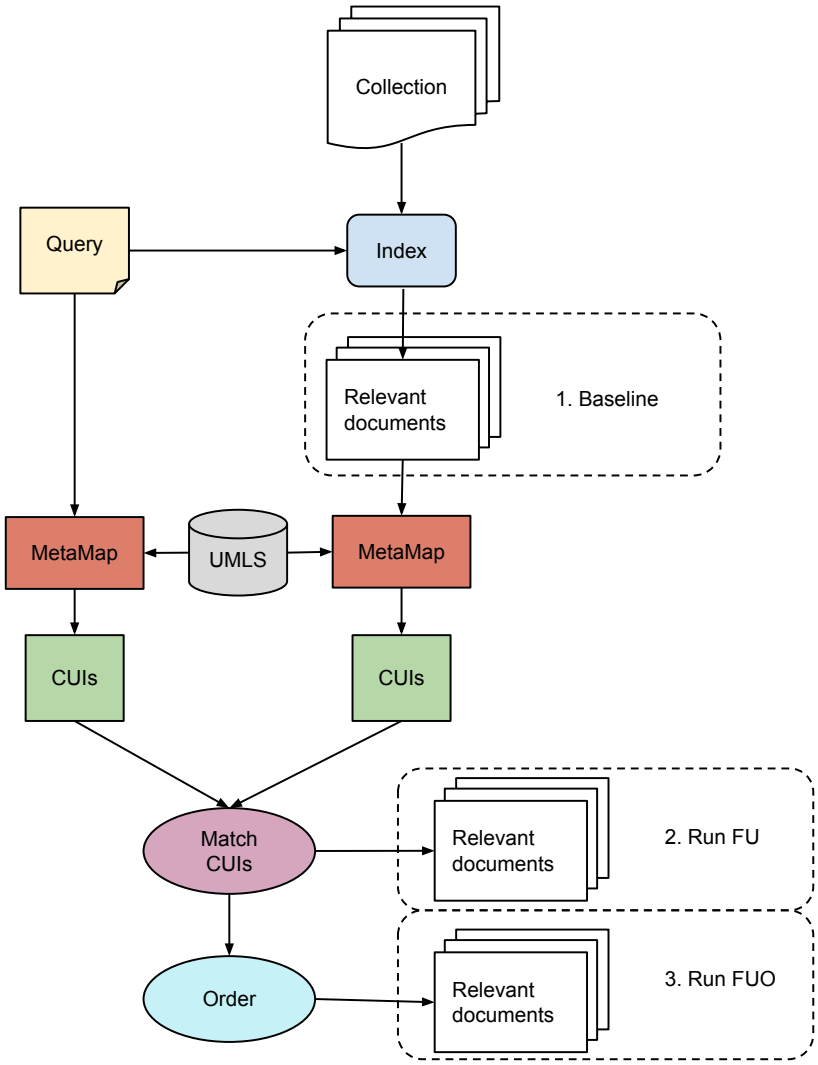

Figure 1: Architecture of the approach

index is shown in Figure 3 and the result of relevant documents was sent as *Baseline*.

3. **Collect**: the next step was to obtain each relevant document returned from the *Baseline* and get the title and abstract of the file. To do this, we need the IDs of the documents returned by Indri and pick up each document from the index.

4. **MetaMap**: afterwards, the title with the abstract of each relevant document was processed using MetaMap to obtain the medical concepts of UMLS with its CUIs. The MetaMap output example, for query 1, is presented in Figure 4.

5. **Match CUIs**: having the CUIs of the query and of each relevant document, we can verify if the document contains any CUI from the query.

   If any identifier does not matches, then the file will not be taken into account to generate the *Run 2 FU* (Filter UMLS concepts) output.

6. **Order by percentage**: the last run sent by our group, we have taken into account all the previous steps and also we counts how many times the CUI of the original query are repeated in the documents.

   Formula 1 shows how to get the percentage of CUI matched:

$$RV_d = NC_d / NC_q \quad (1)$$

   where:

   - RV_d is the rank value of document d

```
<topic number="1">
    <disease>melanoma</disease>
    <gene>BRAF (V600E)</gene>
    <demographic>64-year-old male</demographic>
</topic>
```

```
<query>
    <type>indri</type>
    <number>1</number>
    <text>melanoma braf v600e 64-year-old male</text>
</query>
```

Figure 2: Example of original (upper) and normalized (bottom) query 1.

```
26451873
ASCO_186257-199
23211290
24614711
...
```

Figure 3: Example of output of relevant Indri documents for query 1.

- NC_q is the number of CUIs of document d
- NC_d is the number of CUIs of query q

This percentage is obtained for each document and is sorted in descending order to rank the documents and obtain the *Run 3 FUO* (Filter UMLS concepts and Order results).

4 Results

The results are not satisfactory because *Baseline* is superior to the other *Runs*, so we can conclude that sorting the relevant files according to the UMLS CUIs detected with MetaMap is not a good idea (see Table 1).

Run ID	P@10	R-prec	infNDCG
SINAI_Baseline	0.4980	0.3082	0.4573
SINAI_FU	0.4820	0.2978	0.4510
SINAI_FUO	0.1100	0.0565	0.1080

Table 1: Evaluation scores for 3 automatic runs for scientific abstracts.

A total of 103 Runs have been sent for this task and we have calculated the average obtained for each measurement used. The measures have been 0.5460, 0.2672 and 0.4290 for P@10, R-prec and infNDCG respectively.

As we can see, our score obtained in the measure P@10 does not reach the average. On the other hand, we have exceeded the average of the R-prec measurement in the *Baseline* and *SINAI_FU* sent.

Finally, the precision after 10 documents received (P@10) is not a strong point of our systems and we will have to improve for future participations.

5 Conclusion

The results are not as expected as the *Baseline* obtains better results than the other two Runs that are using MetaMap in order to obtain the medical concepts. Run 1 (*SINAI_FU*) is very close to the *Baseline* results, but they are still lower.

We have analyzed the documents eliminated in Run 2 (*SINAI_FU*) and verified that MetaMap does not detect some medical entities that exist in the query. This makes the accuracy somewhat lower when deleting these documents.

In future work we will use other medical dictionaries to detect the largest number of entities. Also, we will analyze the field of genes (Eden et al., 2009; Köhler et al., 2013) in more detail to give more exhaustive results. It is possible to find a relationship between genes and diseases (Hristovski et al., 2005) in the query, so we will try to generate a graph to find the most appropriate relationships (Martinez et al., 2014; Montejo-Ráez et al., 2014).

C0025202: melanoma [Neoplastic Process]
C0086582: Males [Organism Attribute]
C1706180: Male Gender, Self Report [Qualitative Concept]
C3273990: BRAF NP“¨004324.2:p.V600E [Cell or Molecular Dysfunction]
C2984289: Melanoma Pathway [Functional Concept]
C3539018: NCI CTEP SDC Melanoma Sub-Category Terminology [Intellectual Product]
...

Figure 4: Example MetaMap output for query 1.

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

AACR. 2019. American Association for Cancer Research. https://www.aacr.org/.

Alan R Aronson and François-Michel Lang. 2010. An overview of metamap: historical perspective and recent advances. *Journal of the American Medical Informatics Association*, 17(3):229–236.

Alan R Aronson and Thomas C Rindflesch. 1997. Query expansion using the umls metathesaurus. In *Proceedings of the AMIA Annual Fall Symposium*, page 485. American Medical Informatics Association.

Alan R Aronson. 2001. Effective mapping of biomedical text to the umls metathesaurus: the metamap program. In *Proceedings of the AMIA Symposium*, page 17. American Medical Informatics Association.

ASCO. 2019. American Society of Clinical Oncology. https://www.asco.org/.

Olivier Bodenreider. 2004. The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 32(suppl_1):D267–D270.

Bruce Croft and Jamie Callan. 2016. The Lemur Project. https://www.lemurproject.org/.

Eran Eden, Roy Navon, Israel Steinfeld, Doron Lipson, and Zohar Yakhini. 2009. Gorilla: a tool for discovery and visualization of enriched go terms in ranked gene lists. *BMC bioinformatics*, 10(1):48.

Carol Friedman, George Hripcsak, et al. 1999. Natural language processing and its future in medicine. *Acad Med*, 74(8):890–5.

Dimitar Hristovski, Borut Peterlin, Joyce A Mitchell, and Susanne M Humphrey. 2005. Using literature-based discovery to identify disease candidate genes. *International journal of medical informatics*, 74(2-4):289–298.

Sebastian Köhler, Sandra C Doelken, Christopher J Mungall, Sebastian Bauer, Helen V Firth, Isabelle Bailleul-Forestier, Graeme CM Black, Danielle L Brown, Michael Brudno, Jennifer Campbell, et al. 2013. The human phenotype ontology project: linking molecular biology and disease through phenotype data. *Nucleic acids research*, 42(D1):D966–D974.

David Martinez, Arantxa Otegi, Aitor Soroa, and Eneko Agirre. 2014. Improving search over electronic health records using umls-based query expansion through random walks. *Journal of biomedical informatics*, 51:100–106.

Arturo Montejo-Ráez, Eugenio Martínez-Cámara, M Teresa Martín-Valdivia, and L Alfonso Ureña-López. 2014. Ranked wordnet graph for sentiment polarity classification in twitter. *Computer Speech & Language*, 28(1):93–107.

National Library of Medicine. 2019a. MetaMap - A Tool For Recognizing UMLS Concepts in Text. https://metamap.nlm.nih.gov/.

National Library of Medicine. 2019b. UMLS - Unified Medical Language System. http://uts.nlm.nih.gov.

Thomas C Rindflesch, Lawrence Hunter, and Alan R Aronson. 1999. Mining molecular binding terminology from biomedical text. In *Proceedings of the AMIA Symposium*, page 127. American Medical Informatics Association.

Kirk Roberts, Dina Demner-Fushman, Ellen M Voorhees, William R Hersh, Steven Bedrick, Alexander J Lazar, and Shubham Pant. 2017. Overview of the trec 2017 precision medicine track. *TREC, Gaithersburg, MD*.

U.S. National Library of Medicine. 2019. MedlinePlus - Health Information from the National Library of Medicine. https://medlineplus.gov/.