Semantic Matching Models for Medical Information Retrieval: A Case Study

Dawid Kasperowicz, *York University*, Toronto, Canada

Jimmy Huang, *York University*, Toronto, Canada

Abstract

This paper presents the continuing work in the Medical Records Track of TREC 2011. The goal is to develop tools that are useful for medical professionals to aid in the diagnosis of diseases and suggest potential treatments. The focus of the paper is on proposing new semantic matching models with extracted UMLS concepts. Two methods have been developed: 1) Term-Based Matching and, 2) CUI-Based Matching. These models are applied to rank 107,111 EMR, which contain concepts located in 34 predefined topics, utilizing their relevancy score returned from BioLabler as the basis of the calculation. The results are evaluated using the official measures Bpref, R-Prec and P10. The methods yield promising results as each advancing implementation of the methods yields higher performance, however; further research is required to formulate a more evolved conceptual matching model where the weights of candidate concepts would contribute more efficiently to the overall performance.

Keywords: *Medical Information Retrieval, EMR, UMLS, Model*

1. Introduction and Challenges

With an increasing interest of moving the healthcare paragon on a global scale to being electronically based, there emerges an opportunity to create information retrieval and knowledge discovery systems that are capable of aiding medical professionals in the treatment of patients. This can be accomplished by processing electronic medical records (EMR) in EMR systems for information and patterns. EMR systems are an evolving concept defined as a longitudinal collection of electronic health information about individual patients and populations, and it integrates healthcare information currently collected in both paper and electronic media for the purpose of improving the quality of healthcare [1]. These records consist of, but are not limited to: various dates of results, screenings surgeries, diagnoses of illnesses, in addition to lists of prescriptions, allergies, family history, laboratory results, discharges, etc.

The continued work presented in this paper was conducted for the Text REtrieval Conference (TREC) 2011 Medical Records track. The objective of the Medical Records Track in TREC 2011 is to advance research pertaining to content-based access to free-text fields of EMR [2]. Participants receive a dataset of 101,711 de-identified medical records that is made available for research through the University of Pittsburgh BLULab NLP Repository. In addition, 34 topics are given in which participants are required to answer by identifying EMR's that are most significant to the particular topic being examined, and rank the said EMR in descending order of importance. The 34 topics focus on identifying pre-defined diseases, conditions, and treatments and interventions. For example, topic 111 seeks to identify: "Patients with chronic back pain who receive an intraspinal pain-medicine pump". Hence, for a given topic, a result should be generated which returns a EMR list that is relevant to a given topic, ranked in decreasing likelihood that a particular record satisfies a topics requirements.

When attempting to retrieve contextual information from medical data, there are challenges which are faced; frequent use of acronyms and possible non-standardized acronyms, the presence of homonyms (worlds referring to two or more different entities) and synonyms (two or more words referring to the same entity) [2]. These limitations are present in traditional Information Retrieval (IR) due to the keyword-based matching model considering any given document a bag-of-words [3]. With processing EMR, using such a model is problematic as terminology within may be depending on adjacent terms, and failing to recognize such will result in

inadequate performance. In order to mitigate the above limitations and ensure accurate synonyms and name variants for medical terminology is present when processing the EMR, a biomedical text mining tool named BioLabeler[1], is employed. The tool extracts Unified Medical Language System (UMLS) concepts from biomedical texts to index documents and discover relationships between the documents.

The paper is organized as follows: Section 2 contains related work currently being conducted in the field; Section 3 encompasses the methods that are used for the work presented in this paper; Section 4 encloses the experimental results of the work implemented in this paper; Section 5 includes discussions related to the methods and results presented in this paper; and Section 6 is composed of the conclusion and future work on the continued work illustrated in this paper.

Figure 1: An Illustration of a Typical EMR in the Dataset

2. Related Work

There has been a great volume of research performed in regards to medical related IR. In 2003, the first year which the TREC Genomics Track was created, research was conducted on ad hock retrieval and information extraction, with a focus on Gene Reference into Function resource of the National Library of Medicine [4]. This track was particularly interested in researching how to access and manage the increasing quantity of biomedical information. By 2006, the importance was moved to returning relevant passages which discuss various aspects of a given topic that were derived based on biologists' information needs [5].

Outside the scope of conferences, one of the earliest work performed on medical IR occurred in 1995, by C. Friedman et al. on a system named MEDLEE. The system is a natural language text extraction system used to extract structures, and encode clinical information from textual patient records [6]. The system demonstrated to be comparable to experts in the field who examine reports to determine whether a specified disease is contained in the report.

The TREC 2011 Medical Records Track builds off the research performed above and extends it into EMR. It is conventional to find various existing systems being leveraged in order to take the dataset and topics, and generate corresponding medical concept representations of them, as it was demonstrated to yield the top performing results in [7].

3. Methods

The methods consists of indexing the dataset using UMLS in order to identify key concepts in the EMR and use conceptual matching models that is based on the confidence weights for each concept found within a EMR. Furthermore, semantic indexing is utilized in order to map medical concepts in an ontology to various free text articulations, synonyms, acronyms and anonyms of medical entities that are found within the dataset to ensure a suitable ontological impression of a given medical entity is established.

In order to establish the above ontological impressions, an online biomedical text mining tool named BioLabeler is employed. BioLabeler is designed to create concept associations to any given text, perform stemming, stop word removal, and return candidate National Cancer Institute (NCI)[2] concepts with corresponding weights and normalized weights. These weights are established using the cosine similarity between the input data and those that are a part of the candidate concepts [8].

The EMR in the dataset are formatted as Extensible Markup Language (XML) files and contained the following fields: checksum, subtype, type, chief complaint, admission diagnosis, year, download time, update time, deid, and report text. Figure 1 illustrates a typical EMR in the dataset. When feeding the dataset to BioLabeler, only the report text field from the XML files are used to retrieve candidate concepts, while the entire topic string is given to

[1] http://www.biolabeler.com/bioLabeler/

[2] http://www.cancer.gov/

BioLabeler to retrieve candidate concepts for the topics.

BioLabeler permits selecting specific semantic types and specific UMLS sources to be used when analyzing text. There are 10 specific semantic types selected to retrieve candidate concepts; 6 of which are used to extract disease concepts and the remaining 4 are utilized to extract procedure concepts. The 6 disease semantic types consists of: Acquired Abnormality, Congenital Abnormality, Disease or Syndrome, Experimental Model of Disease, Mental or Behavioural Dysfunction, and Neoplastic Process. The 4 procedure semantic types consists of: Diagnostic Procedure, Health Care Activity, Laboratory Procedure, and Therapeutic or Preventive Procedure. To ensure the best range text articulations, synonyms, acronyms and anonyms of medical entities are used when analyzing text with BioLabeler, all 44 UMLS sources are used.

Once the processing of the dataset and topics is completed, two indexes for each EMR in the dataset and each topic are generated, specifically a candidate disease-based concept index and a candidate procedure-based index. These indexes include various information, including but not limited to: National Cancer Institute Concept Unique Identifier (CUI), medical concept abbreviations, normalized weight, and weight, which are used to determine the relevance of the generated candidate concepts to the patient record and topics. For instance, BioLabeler would consider Topic 101: "Patients with hearing loss" to contain a disease concept of: "Hearing Loss, Central#Central hearing loss#Central hearing loss#Central hearing loss#Central Hearing Loss" with a CUI of C0018776 to be highly ranked with a weight of 4.64 and normalized weight of 0.75.

3.1 Semantic Matching Models

In this section, two semantic matching models are proposed:

Term-Based Matching

For the term-based matching of the EMR to topics, the following formula is used to establish document relevancy: Given the disease-based concept vector of topics $Q_d = (C_1, C_2, ..., C_m)$ and the disease-based EMR concept vector $D_d = (C_1, C_2, ..., C_n)$, the conceptual score of the EMR is computed by matching the terms in the EMR with those that overlap with the topic, and rank them according to the overlapping concept weights in the EMR as follows:

$$Score_d(D) = \sum_{C_i \in Q_d} \frac{1}{n} \sum_{C_j \in D_d / \exists t \in Terms(C_i) \cap Terms(C_j)} w(C_j, D_d) \quad (1)$$

where $Score_d(D)$ is the conceptual score of the EMR obtained using the disease-based index, C_i is a concept in topic Q_d and C_j is a concept in EMR D_d, $Terms(C_i)$ and $Terms(C_j)$ are the lists of terms associated with concepts C_i and C_j respectively, $w(C_i, Q_d)$ and $w(C_j, D_d)$ represent the weight of concepts C_i in the topic Q_d and the weight of concept C_j in EMR D_d respectively. The final score of a EMR is based on adding the weights of both the procedure-based and disease-based EMR score, as indicated by formula 3.

CUI-Based Matching

For CUI-based matching of the EMR to topics, the following formula is utilized:

$$Score_d(D) = \sum_{C_i \in Q_d} w(C_i, Q_d) * w(C_i, D_d) \quad (2)$$

$Score_d(D)$ is the conceptual score of a given EMR which is obtained using the disease-based index. $w(C_i, Q_d)$ and $w(C_i, D_d)$ represent the weight of concept C_i in query Q_d and EMR D_d respectively. The same formula is used to calculate procedures. The final score of a EMR is based on adding the weights of both the procedure-based and disease-based EMR score, as indicated by formula 3.

$$Score_f(D) = Score_d(D) + Score_p(D) \quad (3)$$

4. Experimental Results

Six runs are generated on the conceptual indexes, where a term-based matching model is employed for the first run to compute a relevance score for each EMR in the dataset with respect to a given topic, and a CUI-based matching model is employed for the remaining 5 runs to compute a relevance score for each EMR in the dataset with respect to a given topic. The 6 runs conducted consists of: 1) Term-based matching of the EMR to topics with dropped records, 2) CUI-based matching of EMR to topics without dropped files, 3) CUI-based matching of EMR to topics without dropped files using the top concept, 4) CUI-based matching of EMR to topics without dropped files using up to 5 top concepts, 5) CUI-based matching of EMR to topics without dropped files using the top concept with only the Medical Subject Headings (MSH) UMLS source, and 6) CUI-based matching of EMR to topics without dropped files using up to 5 top concepts with only the MSH source.

Our main goal for the 6 runs is to evaluate the impact of semantic indexing and conceptual matching models presented above. Each of the 6 runs returned up to 1000 EMR ranked in descending order. Table 1 summarizes the results of the 6 runs in the order

presented in Section 2, according to the following official measures: Bpref, R-Prec and P10. The official measures are calculated using the TREC_eval package provided by TREC.

The Bpref measure is an IR metric function that is based on binary relevance [9]. It is designed to judge the number of times a relevant document is retrieved before a non-relevant document [10]. The R-Prec measure is a function where the precision of a document is measured after R documents are retrieved where R is the number of relevant documents for a given topic [10]. The P10 measure counts the number of relevant documents in the top 10 documents in the ranked list returned for a given topic [10].

Table 1: Results of 6 Runs

Run	Bpref	R-Prec	P10
1	0.0834	0.0159	0.0382
2	0.1742	0.0600	0.1147
3	0.1405	0.0852	0.1970
4	0.2009	0.0969	0.2000
5	0.2528	0.1725	0.2559
6	0.2817	0.0639	0.2143

5. Discussions

Examining Table 1, it is clear that the performance of the 6 runs is improving. There are many potential factors which may contribute to the improvement of the runs. For run 1, the main reason for its low performance is related to the accuracy of concept extraction performed by BioLabeler, along with the limitations of term-based matching between concepts of topics and EMR. In addition, certain topics request to return records containing a specific gender and age. BioLabeler does not take these conditions into account when returning candidate concepts, and no pre-processing is performed on the EMR in the dataset prior to submitting them to BioLabeler for analysis in order to extract such information. Therefore, EMR which do not meet the criteria for the 34 topics are included in the ranking process. The same holds true for the remaining 5 runs. In addition, TREC had announced 845 EMR had been dropped from the dataset and are not to be used in the runs. Run 1 includes these 845 records when ranking documents, therefore causing further irrelevant EMR to be included in the results. However, due to these records representing less than a percentage point of the dataset, they do not significantly obstruct the performance of run 1.

An improvement is seen when moving from term-based matching to CUI-based matching, seeing an average improvement of 0.1266 in the Bpref, 0.0798 in the R-prec, and 0.1582 for P10, as is demonstrated in Figure 2. Using CUI-based matching eliminates the limitations of term-based matching between concepts of topics and EMR by removing syntactic, semantic, homonyms, synonyms, etc., limitations. In addition, ensuring the 845 dropped records are not incorporated when ranking the records further ensures no irrelevant document are incorporated in ranking, although their significance is nominal which would not drastically effect performance.

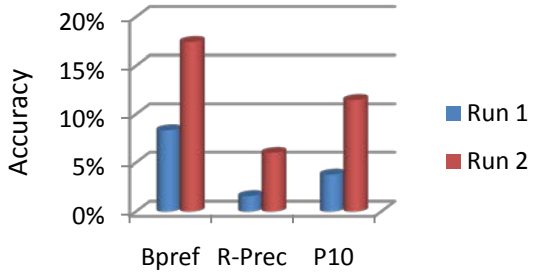

Figure 2: Performance Measures Comparison between Run 1 and Run 2

When comparing run 2 containing the improved results, which uses all candidate concepts from the EMR, to runs 3 and 4, it is evident that the latter runs contain an increased overall performance, as is evident in Figure 3. Run 3 secures an improvement of 0.0571 for Bpref, 0.0252 for R-Prec and 0.0823 for P10 when applying a threshold of only the most significant concept in a given EMR, while run 4 experiences an improvement of 0.1175 for Bpref, 0.0369 for R-Prec and 0.0853 for P10 when applying a threshold of up to the 5 most significant concepts in a given EMR. From these statistics, it is clear that applying a threshold for the amount of candidate concepts to be used for conceptual matching between EMR and topics yield improved results. The question now remains, which threshold yields the most optimal results.

Comparing run 3, which limits the use of candidate concepts to the most significant concept, with run 5, which likewise limits the use of candidate concepts to the most significant concept in addition to leveraging only the MSH UMLS source opposed to all available sources; demonstrates utilizing a limited number of UMLS sources yields improved performance, as is demonstrated in Figure 4. It is observed that run 5 contains an improvement of 0.1123 for Bpref, 0.0873 for R-Prec and 0.0589 for P10 over run 3. Likewise, a similar performance improvement of 0.0808 for Bpref, 0.0143 for P10 is witnessed in Figure 5 comparing run 4, which utilizes all UMLS sources

with a threshold of up to 5 most significant concepts being utilized for concept matching, and run 6, which utilizes only the MSH source with a threshold of up to the 5 most significant concepts being utilized for concept matching.

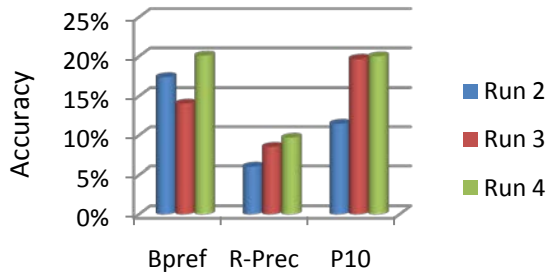

Figure 3: Performance Measures Comparison between Run 2 – Run 4

These statistics establish there is a need to utilize a precise amount of UMLS sources, opposed to utilizing all available sources in order to retrieve optimal performance. What now needs to be determined is which combination of UMLS sources yield the most optimal results.

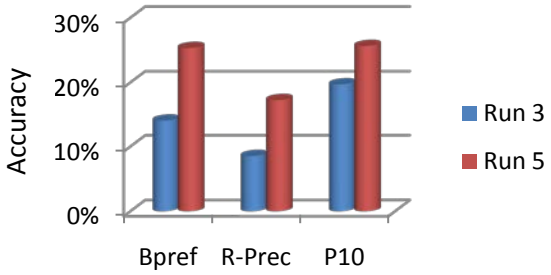

Figure 4: Performance Measures Comparison between Run 3 and Run 5

6. Conclusion and Future Work

The TREC Medical Records Track exhibits a challenging task. This paper presents the participation effort done in the Medical Records Track of TREC 2011. The participation effort constitutes of using BioLabeler for indexing 101,711 and 34 topics EMR for procedure-based and disease-based concepts. Six runs are conducted where a term-based matching model is employed for the first run to compute a relevance score for each EMR in the dataset with respect to a given topic, and a CUI-based matching model is employed for the remaining 4 runs to compute the relevance score for each EMR in the dataset with respect to a given topic. Runs 1-4 use all 44 available UMLS concepts provided by BioLabeler, while the remaining use only MSH. Runs 1 and 2 utilize all candidate concepts provided by BioLabeler, while runs 3 and 5 utilize the most relevant candidate concept, and runs 4 and 6 employ up to the 5 top candidate concepts.

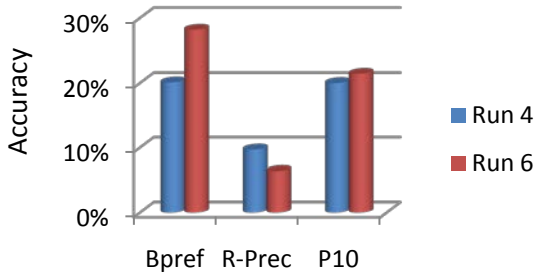

Figure 5: Performance Measures Comparison between Run 4 and Run 6

Employing the use of BioLabeler to perform indexing proves to yield results which have room for improvement. By pre-processing the dataset to extract information such as gender, age, etc., to pre-screen the EMR, along with being selective with UMLS sources, and setting a relevancy threshold for candidate concepts returned from BioLabeler, it is expected to observe an increase in performance, as exemplified in run 5.

Future work will focus on pre-processing the dataset prior to submitting EMR to BioLabeler for analysis, perform a more in-depth examination of available UMLS sources to BioLabeler to ensure only value added sources are to be used to create accurate medical concepts. In addition, other tools will be sought out to be either used in conjunction with BioLabeler, or exclusively, in an attempt to disambiguate and exclude irrelevant concepts from both the dataset and topics. Furthermore, there is a need for a more evolved conceptual matching model where the weights of candidate concepts would contribute more efficiently to the overall performance of future runs.

Acknowledgements

This research is jointly supported by NSERC of Canada and the Early Researcher/Premier's Research Excellence Award. The authors would like to thank Robert Saggiorato of the Ontario Hospital Association for his valuable insight and resources, in addition to Jose Luis Marina and Csar Silgo from the BioLabeler team for their invaluable help while conducting the experiment. Furthermore, the authors

would like to thank Dr. Mariam Daoud for her suggestions and support, in addition to Miao Jun for his aid.

References

[1] Tracy D Gunter and Nicolas P Terry, "The Emergence of National Electronic Health Record Architectures in the United States and Australia: Models, Costs, and Questions," *Journal of Medical Internet Research*, vol. 7, no. 1, March 2005.

[2] Mariam Daoud, Dawid Kasperowicz, Jun Miao, and Jimmy Huang, "York University at TREC 2011: Medical Reords Track," in *Proceedings of the 20th TREC 2011: Medical Records Track*, 2011.

[3] Ming Zhong and Xiangji Huang, "Concept-Based Biomedical Text Retrieval," in *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2006, pp. 723-724.

[4] William Hersh and Ravi Teja Bhupatiraju, "TREC Genomics Track Overview," in *Proceedings of the Twelfth Text Retrieval Conference*, 2003, pp. 14-23.

[5] William Hersh, Aaron M Cphen, Phoebe Roberts, and Hari Krishna Rekapalli, "TREC 2006 Genomics Track Overview," in *Proceedings of the 15th Text REtrieval Conference*, 2006.

[6] C Friedman, G Hripcsak, W DunMouchel, S B Johnson, and P D Clayton, "Natural Language Processing in an Operational Clinical Information Syste," *Natural Language Engineering*, vol. 1, no. 1, pp. 83-108, 1995.

[7] Wei Zhou, Clement Yu, Neil Smalheiser, Vetle Torvik, and Jie Hong, "Knowledge-Intensive Conceptual Retrieval and Passage Extraction of Biomedical Literature," in *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information retreival*, New York, 2007, pp. 655-662.

[8] Mariana Luís, José Luís Marina, and Alberto Pascual-Montano, "BioLabeler and Moara in the First Round of the CALBC challenge," in *Proceedings of the First CALBC Workshop*, Hinxton, Cambridgeshire, 2010.

[9] Tetsuya Sakai, "Alternatives to Bpref," in *Proceedings of the 30th Annual International AMC SIGIR Conference on Research and Development in Information Retrieval*, New York, 2007, pp. 71-78.

[10] Chris Buckley and Ellen M Voorhees, "Retrieval Evaluation with Incomplete Information," in *Proceedings of the 27th Annual International AMC SIGIR Conference on Research and Development in Information Retrieval*, 2004, pp. 25-32.

[11] Bing Liu, *Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data*, 1st ed. New York: Springer, 2007.

[12] Xiangi Huang, Ming Zhong, and Luo Si, "York University at TREC 2005: Genomics Track," in *Proceedings of the Fourtheenth Text REtrieval Conference*, 2005.

[13] M M Beaulieu et al., "Okapi at TREC-5," in *Proceedings of the Fifth Text REtrieval Conference (TREC-5)*, 1996, pp. 143-165.

Correspondence Address

Dawid Kasperowicz and Jimmy Huang
Information Retrieval and Knowledge Management Research Lab
School of Information Technology
York Univeristy, Toronto, Canada
Email: {dawidk, jhuang}@yorku.ca