UCD IIRG at TREC 2012 Medical Track

James Cogley, Nicola Stokes, John Dunnion and Joe Carthy
School Of Computer Science and Informatics
University College Dublin, Belfield, Dublin 4, Ireland.
james.cogley@ucdconnect.ie,
{nicola.stokes, john.dunnion, joe.carthy}@ucd.ie

Abstract

This paper describes the participation of UCD IIRG in the TREC 2012 Medical Records track, which fosters research in the retrieval of electronic health records using free text fields. Our contributions to this track investigate several problem areas in the retrieval of medical documents. Multiple knowledge sources are investigated to alleviate the issue of vocabulary mismatch. Medical records are verbose documents that give a full picture of a patient's medical status including their family health information and their own medical history. A Condition Attribution and Temporal Grounding system is implemented to address such occurrences. A rule-based system is employed in order to extract the patient's demographic information from their medical record. All extracted information is then leveraged using Indri's structured query language. These methods are combined to identify patients who fit the exact criteria as described in natural language queries.

1 Introduction

The TREC 2012 Medical Records track fosters research in the retrieval of electronic health records using free text fields. These free text fields outline a set of *inclusion criteria* that specify a patient cohort, and must be satisfied in order for a record to be relevant. These criteria include demographic information (for example a patient's age, gender and ethnicity) as well as their medical conditions or treatments.

The comprehensive nature of medical records pose numerous problems in accurately identifying patients who fit a very specific set of criteria. Firstly, there is the issue of vocabulary mismatch. For example "hypertension" may be noted as "high blood pressure", "HNT" or "HBT". To address this we use several knowledge sources and tools to identify and expand key medical concepts provided in queries. Medical records often describe a patient's medical history including that of relatives. Such issues are resolved with the use of a Condition Attribution and Temporal Grounding system that determines if a patient experienced a condition and if it is relevant to their current state. In order to identify highly specific information in verbose documents we aim to add structure to documents using machine learning and rule-based methods that is then leveraged using Indri's structured query language.

This paper is structured as follows. Section 3 provides an overview of the task as well as the Indri retrieval model. Section 3 details the background technologies and motivation behind the systems developed at IIRG. Section 4 describes the submitted runs and the results of their evaluation. Discussion and conclusions are put forward in Section 5 and Section 6, respectively.

2 Search Task

The goal of the Medical Records track is to foster research on providing content-based access to the free-text fields of electronic medical records[1]. The 2012 task builds on the previous year with the aim of identifying cohorts of patients who fit a set of criteria that describe their medical status as well as key demographic information, known as *inclusion criteria*. The queries for this task take the form of natural language text specifying a set of inclusion criteria, e.g. *Children with dental caries*. The test corpus comprises 101,700 de-identified medical records obtained from the University of Pittsburgh NLP repository[2]. The unit of retrieval for this task is a *visit*, which is composed of multiple records that relate to the same medical episode. A mapping report is provided for the task in order to create these 17,265 visits.

2.1 Indri Structured Query Language

The systems described in this paper all use the Indri[3] IR engine, that combines inference model and language model approaches(Metzler and Croft, 2004). This model was chosen as the basis for the systems described in this paper because of its ability to handle phrases as well as its robust structured query language, one aspect of this query language is the ability for field restriction. Field restriction limits the matching of an expression to a particular field found within indexed documents. For example, the query "shakespeare.author" would ensure that documents matching shakespeare in the author field are returned. On the corpus side, field extents are identified using XMLlike markup, e.g. <author>shakespeare</author>.

3 System Background & Motivation

This section outlines the technologies and motivation behind UCD IIRG's submissions to the TREC 2012 Medical Track.

3.1 Demographic Information

A key aspect in identifying patient cohorts is the resolution of demographic information. Demographic information includes attributes such as age and gender as well as ethnicity. Age group information is identified in the Pittsburgh dataset through their own de-identification process. Gender and ethnicity is extracted using a set of regular expression rules.

3.2 Condition Identification, Condition Attribution & Temporal Grounding

In order to identify medical conditions in text, an implementation of the IndexFinder algorithm was used (Zou et al., 2003). However as medical records must give a complete description of a patient's healthcare profile, they are necessarily word-heavy. This includes a description of familial medical conditions (e.g. *father's diabetes*) as well as past medical history (e.g. *past admission with wrist fracture*). A Condition Attribution and Temporal Grounding system (Cogley et al., 2012) is used to resolve such occurrences.

3.3 Indexing

The systems described in Sections 3.2 and 3.1 were used to mark field extents describing age, gender, ethnic origin as well as information regarding past, present and familial medical conditions. The markup tags used in the system are outlined in Table 1.

Following the identification of required information, visit documents were created. The visit documents are created using a simple shell script to concatenate a visits constituent reports. The script reads the mapping file provided by TREC, which mapped reports to visits using unique identifiers. Indexes were then built from these visit documents. The index was left unstemmed in order to avoid problem instances such as 'AIDS' stemmed to 'AID'.

Field Tag	Usage
CONDITION	Identifies a condition
PASTCOND	Identifies a former condition
FAMCOND	Identifies a familial condition
AGE	Identifies an age
AGEGRP	Identifies an age group
GEN	Identifies patient gender
ETH	Identifies patient ethnicity

Table 1: Fields marked in Text

3.4 Structured Retrieval

Following the identification of fields in documents using the systems described in previous sections,

[1] http://trec.nist.gov/act_part/tracks12.html

[2] http://www.dbmi.pitt.edu/nlp/report-repository

[3] http://sourceforge.net/projects/lemur/

this information may be manipulated using Indri's structured retrieval and field restrictions. As a result highly specific queries (1) may be translated to structured queries (2).

1. Elderly adults with a past admission for fractures

2. elderly.AGEGRP adults with a past admission for fractures.PASTCOND

3.5 Query Expansion

Medical literature is a rich source of synonymy and in an IR context vocabulary mismatch is an often encountered issue (Limsopatham et al., 2011). To combat these problems, query expansion is employed. Both manual and automatic methods are used, as detailed below. Concept Re-Ranking (Stokes et al., 2007), a method developed to address the issues arising when a document contains multiple references to the same concept term, is also enforced.

3.5.1 Manual

The query is submitted to PubMed[4]. Pubmed creates chunks from this query which represent entries in MeSH (Johnston et al., 2002). A manual systematic lookup is then performed on these entries at `http://www.ncbi.nlm.nih.gov/mesh` i.e. All synonyms are then taken from these results and added to the original query. No manual filtering of appropriate terms was conducted.

3.5.2 Automatic

Concepts are identified using the MetaMap tool. Expansions are then generated using one or more of the following: the MetaMap tool; queries to PubMed; the knowledge graph FreeBase[5]. The queries are then automatically generated using these expansions and a rule based system translates the queries to Indri's structured query language.

3.6 Submitted Systems

Four systems were submitted by UCD IIRG to this year's Medical Track. They are as follows.

- `UCDCSI1` A manual run using MeSH based expansions, Indri's structured query language to specify demographic information and Concept Re-Ranking.

- `UCDCSI2` A manual run using MeSH based expansions, Indri's structured query language to specify demographic information such as ages and Concept Re-Ranking. Furthermore it uses field-based retrieval in order to utilise more specific information regarding medical conditions namely determining the experiencer and whether or not it occurred in the past.

- `UCDCSI3` An automatic run using MeSH based expansions, Indri's structured query language to specify demographic information such as ages and Concept Re-Ranking.

- `UCDCSI4` An automatic run using MeSH based expansions and Indri's structured query language to specify demographic information such as ages without Concept Re-Ranking.

4 Experimental Results

This section describes the performance of the four runs submitted to the TREC 2012 Medical Track by UCD IIRG. The submissions consisted of two manual runs UCDCSI1, UCDCSI2 and two automatic runs UCDCSI3 and UCDCSI4. Four metrics, infAP, infNCDG, R-prec and Precision @ 10 (P @ 10) are used to evaluate all submissions. 88 runs were submitted in total to the track. Of the 88 submitted,6 were manual with the remaining 82 automatic. Tables 2 and 3 display the results of UCD IIRG's submissions and the hypothetical Max and Median runs, respectively. Table 4 shows the hypothetical max, median automated runs and `UCDCSI4` .

ID	infAP	infNDCG	R-prec	P @ 10
UCDCSI1	0.168	0.406	0.280	0.4915
UCDCSI2	0.121	0.346	0.237	0.3851
UCDCSI3	0.089	0.286	0.195	0.283
UCDCSI4	0.105	0.319	0.223	0.340

Table 2: UCD IIRG Submissions

`UCDCSI1` had the best performance among UCD IIRG submissions across all metrics. The manual runs significantly outperformed their automatic

[4] `http://www.ncbi.nlm.nih.gov/pubmed`

[5] `http://www.freebase.com/`

counterparts, owing mainly to the much more sophisticated manner of query structuring that uses weighting among concepts.

ID	infAP	infNDCG	R-prec	P @ 10
MAX	0.395	0.722	0.515	0.802
MEDIAN	0.200	0.464	0.326	0.551
UCDCSI1	0.168	0.406	0.280	0.492

Table 3: Manual Max, Automatic Median & UCDCSI1

The top-performing IIRG submission achieves moderate performance, with no significant difference between it and the hypothetical median run.

ID	infAP	infNDCG	R-prec	P @ 10
MAX	0.423	0.746	0.543	0.815
MEDIAN	0.170	0.424	0.294	0.470
UCDCSI4	0.105	0.319	0.223	0.340

Table 4: Automatic Max, Automatic Median & UCDCSI4

Although the automated submission UCDCSI4 performs moderately well, it does so using very basic rules in order to translate natural language queries into Indri's complex query language that requires the knowledge of domain experts. The ramifications of this are discussed in the next section.

5 Discussion

In this section, we present a discussion of the results of UCD IIRG's submissions to TREC Medical Track 2012, both in respect to one another and the hypothetical max and median runs.

Figure 1 displays the per topic *infAP* score of the maximum and median participant results along with the authors' top performing submission, UCDCSI1.

UCDCSI1 matches the best performing runs on six occasions (136, 138, 143, 158, 178, 184). Of these six queries, the median also matched on performace (143, 178). Both of these queries are relatively simple e.g. *Patients who have had a carotid endarterectomy*, which explains their high performance. UCDCSI1 outperforms the median runs on queries with age group information, *Children with dental caries* and secondary information, *Patients with esophageal cancer who develop pericardial effusion*.

In total, UCDCSI1 equals the score of the median run on 16 topics (137, 143, 145, 147, 148, 149, 152, 155, 156, 165, 168, 174, 177, 179, 182, 185). These queries ranged from quite simple criteria such as *Patients with Ischemic Vascular Disease* that achieved high scores to very complex queries such as topic 141, *Adult inpatients with Alzheimer's disease admitted from nursing homes with pressure ulcers* or queries that had limited expansions such as *Patients with inflammatory disorders receiving TNF-inhibitor treatments* that produced very low scores.

There is a significant difference in the performances of UCDCSI1 and UCDCSI2 with UCDCSI1 achieving consistently higher scores as shown in Table 4. These results indicate that the introduction of further information relating to conditions in the query may be too strict.

The automatic submissions UCDCSI3 and UCDCSI4 are significantly outperformed by the max and median runs. An important point of note is the increase in performance of UCDCSI4 over UCDCSI3 on the removal of concept Re-Ranking. There are two causes for this. First, as they are automatic runs they are susceptible to the inaccuracies of the concept identification tool. Secondly, only medical conditions were identified in the queries, giving them a much higher weighting than equally important concepts such as treatments, medications and demographic information.

6 Conclusion

As part of TREC Medical track 2012, we investigated several problem areas in the retrieval of medical documents. These submissions investigated the areas of query expansion, automated query translation from natural language to Indri's structured query language and the use of specific demographic information about patients including their age, gender and information relating to their medical status. Four submissions were made in investigation of these problem areas. Two runs were manually created with the others using automatically generated queries. The manual query UCDCSI1 matched, on average the median run. However, this approach encountered difficulty with queries that had few expansions as well as verbose queries. The use of more specific information such as the attribution and

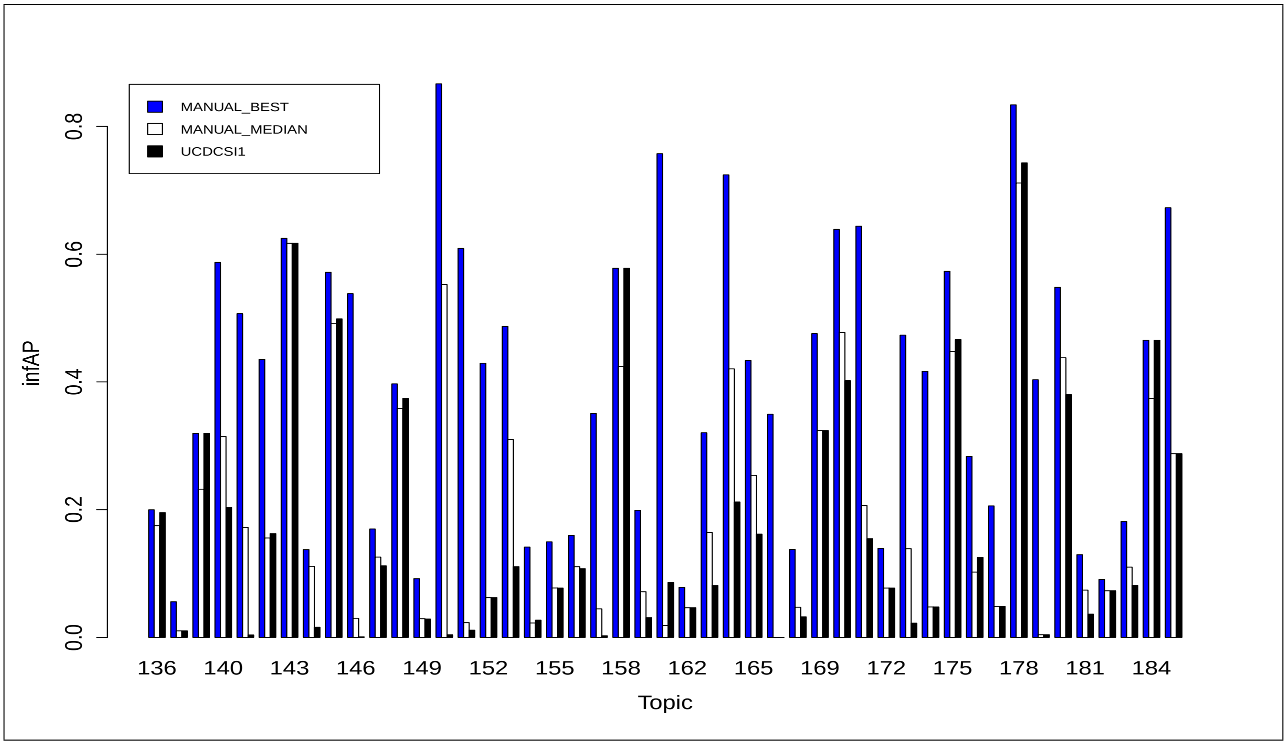

Figure 1: infAP Max, Median and UCDCSIrun1 scores per topic

temporal grounding of medical conditions saw a decrease in performance. The automatic runs were outperformed by the max and median runs, highlighting difficulties in concept identification as well as the automated translation of natural language queries to a structured format.

This leads to future work in performing more accurate concept identification as well as further refining the translation of natural language queries. Furthermore the expansion resources used for this task proved to be limited, thus failing to resolve fully the effects of vocabulary mismatch.

References

J. Cogley, N. Stokes, J. Carthy, and J. Dunnion. 2012. Analyzing patient records to establish if and when a patient suffered from a medical condition. In *2012 Workshop on Biomedical Natural Language Processing*.

W. D. Johnston, S. J. Nelson, and B. L. Humphreys. 2002. Relationships in medical subject headings (mesh).

N. Limsopatham, C. Macdonald, and I. Ounis. 2011. University of glasgow at medical records track 2011: Experiments with terrier. In *TREC 2011*.

D. Metzler and W.B. Croft. 2004. Combining the language model and inference network approaches to retrieval. In *Information Processing and Management Special Issue on Bayesian Networks and Information Retrieval*, volume 40, pages 735–750.

N. Stokes, Y. Li, L. Cavedon, E. Huang, J. Rong, and J. Zobel. 2007. Entity-based relevance feedback for genomic list answer retrieval entity-based relevance feedback for genomic list answer retrieval entity-based relevance feedback for genomic list answer retrieval. In *The Proceedings of TREC Genomics Track 2007*.

Q. Zou, W. W. Chu, C. Morioka, G.H. Leazer, and H. Kangarloo. 2003. Indexfinder: A knowledge-based method for indexing clinical texts. In *AMIA Annual Symposium*.