

The Siena College Medical Information Retrieval System (MIRS)

Larry Medsker, Sharon Small
Christopher Rivadeneira, Andrew Reynolds, and Mara Afzali
The Siena College Institute for Artificial Intelligence
Siena College, Loudonville, NY 12211
{lmedsker, ssmall, cs04riva, at31reyn, md10afza}@siena.edu

Abstract

The work done by our MIRS team of three students and two faculty mentors resulted in a baseline system for content-based medical record retrieval. We also made significant progress on an alternative system based on neural computing concepts. The task for the Medical Records TREC in 2012 was to process a list of thirty-four randomly selected queries against a large medical records database to simulate searches for patients who meet the criteria for participating in various clinical trials. The task was to analyze a data set of over 100,000 reports associated with hospital visits and identify patients whose situations were relevant to the queries. Our text retrieval process was done in two separate ways: one used an index created from standard Information Retrieval (IR) software called Lucene and an alternate method based on principles of neural computing. We submitted three runs to the TREC competition, two using our standard Lucene-based approach and one that used elements of neural network analysis.

The MIRS team was provided the code necessary to download the Medical Records corpus, consisting of an average of 15 reports from each of approximately 100K patient visits to a hospital. Teams were also provided a training set of 34 sample topic statements from TREC 2011. The records, which comprised one month of reports from multiple hospitals, came from the University of Pittsburgh NLP Repository and were de-identified in regard to specific patient names. The Medical Record track organizers from TREC also provided year 2011 judgment sets, produced by medical professionals at the Oregon Health Science University, that we then used in testing our MIRS software at different states of development. For each topic the system was required to search the medical records data corpus and return a ranked list of the top 10 relevant hospital visits, which were proxies for specific patients whose personal identification was made anonymous by the TREC organizers. It is not yet clear how traditional IR should perform on the identification of patients suitable for the clinical trials. Our first logical step was to run an experiment using traditional simple keyword informational retrieval. We used the open source IR system Lucene to index the NIST-supplied Medical Records corpus and to run our baseline experiments. This Lucene version became the first version of our MIRS system and these results were used as our baseline. Modules were proposed and implemented to improve the keyword identification. Then, the first round of experimentation was run with full error analysis. Modules were modified based on this error analysis, run again on the training collection and finally run on the test collection. Results were submitted to NIST before the deadline of August 11, 2012.

General Terms

TREC 2012, Lucene, Weka

Keywords

TREC 2012, medical records, information retrieval

1. Introduction

In May 2012, the Siena College Institute for Artificial Intelligence (SCIAI) formed a team comprising two faculty and three undergraduate students to work on the MIRS project. The team designed unique modules involving untried techniques that they hoped would increase the precision of results over the baseline run. They were

Report Documentation Page				Form Approved OMB No. 0704-0188	
Public reporting burden for the collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to a penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.					
1. REPORT DATE NOV 2012		2. REPORT TYPE		3. DATES COVERED 00-00-2012 to 00-00-2012	
4. TITLE AND SUBTITLE The Siena College Medical Information Retrieval System (MIRS)				5a. CONTRACT NUMBER	
				5b. GRANT NUMBER	
				5c. PROGRAM ELEMENT NUMBER	
6. AUTHOR(S)				5d. PROJECT NUMBER	
				5e. TASK NUMBER	
				5f. WORK UNIT NUMBER	
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) Siena College, The Siena College Institute for Artificial Intelligence, Loudonville, NY, 12211				8. PERFORMING ORGANIZATION REPORT NUMBER	
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)				10. SPONSOR/MONITOR'S ACRONYM(S)	
				11. SPONSOR/MONITOR'S REPORT NUMBER(S)	
12. DISTRIBUTION/AVAILABILITY STATEMENT Approved for public release; distribution unlimited					
13. SUPPLEMENTARY NOTES Presented at the Twenty-First Text REtrieval Conference (TREC 2012) held in Gaithersburg, Maryland, November 6-9, 2012. The conference was co-sponsored by the National Institute of Standards and Technology (NIST) the Defense Advanced Research Projects Agency (DARPA) and the Advanced Research and Development Activity (ARDA). U.S. Government or Federal Rights License					
14. ABSTRACT					
15. SUBJECT TERMS					
16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT Same as Report (SAR)	18. NUMBER OF PAGES 4	19a. NAME OF RESPONSIBLE PERSON
a. REPORT unclassified	b. ABSTRACT unclassified	c. THIS PAGE unclassified			

required to assess the state of the art prior to proposing their module to the entire team.

2. Research Outcomes for the Summer 2012

The open source IR system Lucene was downloaded and installed locally on our server. The NIST Medical Records corpus was formatted to the specifications required by Lucene [1] and an index was generated. The group then experimented with the baseline system using the 34 training topics. The software was run on sample queries from the 2011 TREC competition, and a full round of error analysis on the results for each topic statement was completed. This involved analyzing what worked when a relevant record was found and what was not working when an irrelevant record was returned or a relevant one was missed.

2.1 Part One: Andrew Reynolds and Chris Rivadeneira, in collaboration with Mara Afzali and Drs. Medsker and Small, designed and developed software that used the open source IR system Lucene to index the NIST-supplied Medical Records corpus and to run our baseline experiments. Modules were proposed and implemented, and rounds of experimentation were run with full error analysis. They generated a Lucene index for each topic statement and judged whether a record was relevant based on the Lucene score for the match with keywords and phrases found in the patients' hospital visit records.

2.2 Part Two: Mara Afzali and Larry Medsker, designed a process for translating indexed words and phrases into a format suitable for analysis by neural computing software. The process started with the Phase One output, which comprised a table of keywords extracted from each topic statement. That table was then expanded so that each keyword was mapped to an equivalent subset of medical keywords.

Keyword or Phrase	Additional words and phrases accepted by the counting software (it also accounts for plurals and negations).						
adult	man	woman					
received colonoscopy	colonoscopy						
revealed adecarcinoma	adecarcinoma	colorectal cancer	Colon Adenocarcinoma	Adenocarcinoma			
	(adenocarcinoma?)						
adult	received colonoscopy	revealed adecarcinoma	use in ----->	NN Training Set		Desired response	
				adult	received colonoscopy	revealed adecarcinoma	from the NN
				1	1	1	2
				1	1	0	0
				1	0	1	0
				1	0	0	0
				0	1	1	1
				0	1	0	0
				0	0	1	0
				0	0	0	0

Figure 1: Example of an expansion of initial keywords and translation to a format that is suitable for neural network analysis.

Each subset was used to produce a table of ones and zeros such that an entry from each subset produced a valid pattern of keywords representing text that was relevant to a particular topic statement. This table then became the data set that could be used for training a neural network. Furthermore, grouping the digital results for all the records associated with one patient visit for a given topic

statement gave a complete picture for a given patient and allowed the neural analysis to take advantage of statistical features in determining relevance of a patient to the topic statement for a given clinical trial. The trained neural network could then be used to identify patterns in a record if the latter were similarly mapped to the ones and zeroes format.

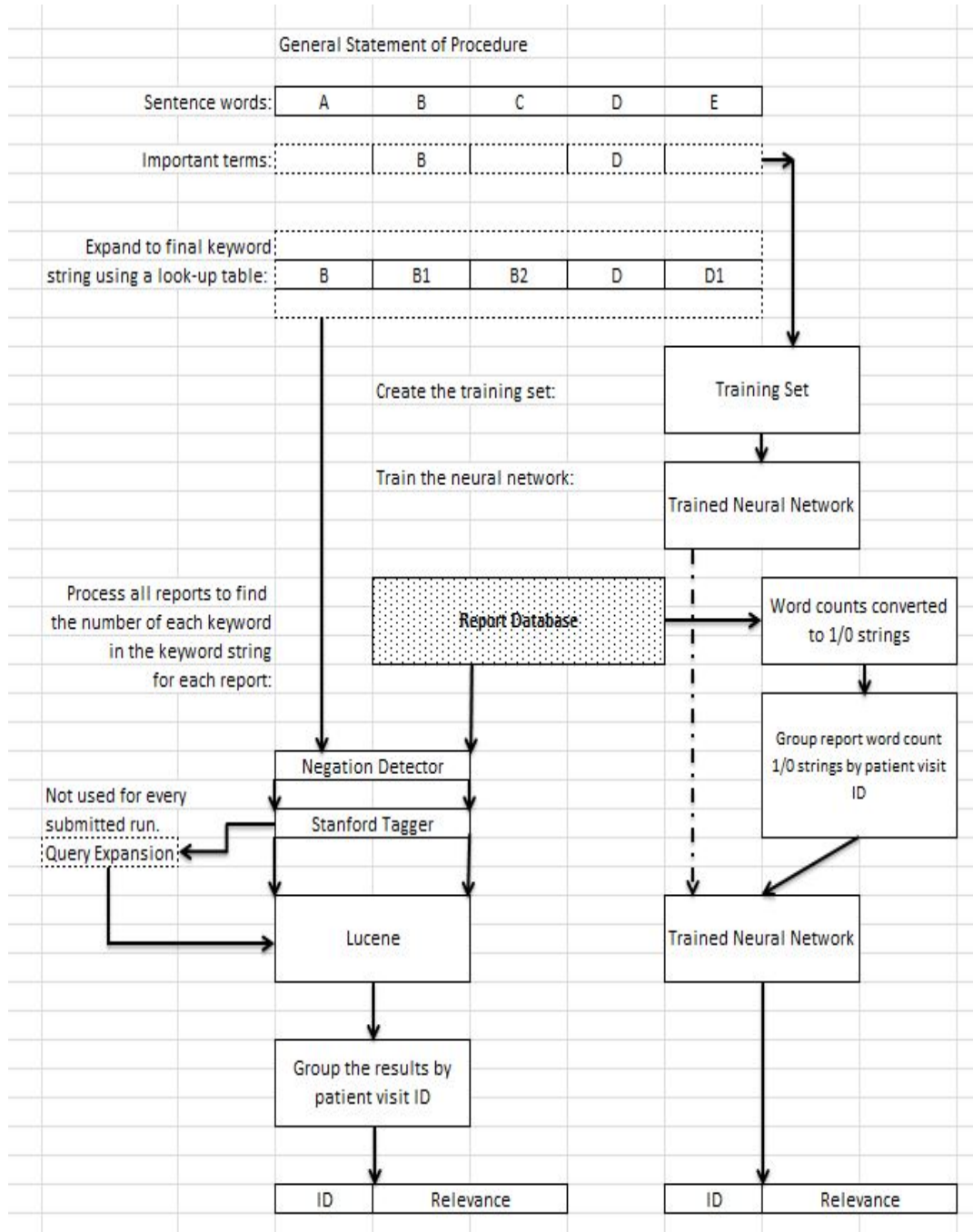


Figure 2. Structure of the overall MIRS process that uses a common initial text analysis and then branches to the standard Lucene or alternative neural computing methods.

Because this medical records application work was started from scratch and required detailed study and creativity, we were not able in the length of time available to complete the use of neural computing software; however, we were able to use statistical treatments on the neural network input data to get very encouraging results on the 2011 data used for testing our system. We therefore produced a run using a modified statistical method on the neural computing data for the 2012 records data and submitted the results to the TREC competition as a possible improvement on the Lucene method used in the other two submitted runs. The judging showed our best run to be at 49.79% precision. We have identified in this work clear next steps for building on the 2012 summer work and exploring further the neural computing approach.

For both methods, we used the judged set of 2011 Medical Records data from the TREC organizers to test the software we developed for each of the two approaches. One result of our work is a design for a comprehensive process that uses initial analysis of statements and medical records and then branches to a traditional IR approach and to a technique based on neural computing ideas (see Figure 2). Another result of our work is a better understanding of the challenges and location of the chokepoint for producing excellent results, namely the identification of the proper keywords from topic statements to accurately and efficiently reflect the significance of the various patient records. More specifically, the challenge is to map the important words in a statement into an expanded set of medically equivalent words that are relevant, and not misleading, for use in the automated searches.

3. Summary

We developed a model for future work: a hybrid system that uses traditional text retrieval components plus elements of neural computing techniques. Test runs using topic statements from the 2011 TREC competition, and run against the 2012 data, gave very encouraging results that is in excess of the relevance rates found by groups in the 2011 competition. In this research, we have identified a crucial need for building a comprehensive database of medical keywords and phrases that can link topic statements to medical reports. We have also created a preliminary hybrid model that may lead to effective integration of neural computing techniques with conventional text analysis capabilities. Further work on this research could lead to enhanced clinical care and improved medical trials and epidemiological studies.

Also, importantly, we now have a working prototype for our Medical Information Retrieval System, which will be the basis for future research at SCIAI and for participation in future TREC competitions. This broader impact of our 2012 Summer Scholars research will benefit additional Siena students beyond the ones who participated in the summer 2012 program. Our work, along with others in the summer 2012 IR work, also became the basis for a proposal to NSF for a Research Experiences for Undergraduates (REU) Site at Siena College that would bring top students from other colleges to work with us in the summers.