Siena's Clinical Decision Assistant with Machine Learning

Brendan Kish, Thomas Walsh, Katherine Small, Steven Gassert,
Kylie Small, Sharon Gower Small

Siena College Institute for Artificial Intelligence
515 Loudon Road
Loudonville, NY 12211
bt13kish@siena.edu, te21wals@siena.edu, smallk1@hawkmail.newpaltz.edu,
gasserts1@hawkmail.newpaltz.edu, ka12smal@siena.edu, ssmall@siena.edu

Abstract

This paper discusses Siena's Clinical Decision Assistant's (SCDA) system and its participation in the Text Retrieval Conference (TREC) Clinical Decision Support Track (CDST) of 2016. The overall goal of this track is to link medical cases to information that is pertinent to patient care. Participants were given a set of thirty topics in the form of medical case narratives and a snapshot of 1.25 million articles from PubMed Central (PMC). This snapshot was taken on March 28, 2016. New this year is that actual electronic health records (EHR) were used instead of synthetic cases. Admission notes from the MIMIC-III database were used to generate the topics. TREC describes the EHR notes as something that "describes a patient's chief complaint, relevant medical history, and any other information obtained during the first few hours of a patient's hospital stay, such as lab work." Each topic consists of three fields. There is a new field this year, the admission *notes*, which is the actual admission's data generated by the clinicians (mostly physicians, including residents, and nurses). The other two fields continue from last year: the *description* field, which is a layman-terms account of the patient visit, and a *summary field ,* which is typically a one or two sentence summary of the main points of the visit. The thirty topics were annotated in three major subsets: diagnosis, test and treatment, with ten of each type. SCDA used several methods to attempt to improve the accuracy of medical cases retrieved. SCDA used the metathesaurus Unified Medical Language System (UMLS) that was implemented using MetaMap (NIH, 2013), machine learning and query and document framing (Small and Stzalkowski, 2004). SCDA also used Lucene for initial document indexing and retrieval. The track received a total of 115 runs from 26 different groups. We submitted two *notes* runs where our highest P(10) run was 0.16 and three runs where we used just the summary field and our highest P(10) was 0.2767. The average P(10) from CDST TREC 2015 Task A was 0.33, with a low of .0867 and a high of 0.4733. Our best Task A run last year had a P(10) of 0.3767. The work described here was performed by a team of undergraduate researchers working together for just ten weeks during the summer of 2016. The team was funded under the Siena College Institute for Artificial Intelligence's National Science Foundation's Research Experience for Undergraduates Grant.

1. Introduction

The Clinical Decision Support Track (Simpson et al., 2014) is a program in the Text Retrieval Conference (TREC) (Voorhees, 2007). TREC is a program co-sponsored by the National Institute of Standards and Technology (NIST) and the U.S. Department of Defense and it focuses on supporting research in information retrieval and extraction, and to increase availability of appropriate evaluation techniques. The Clinical Decision Support Track was run for the third time in 2016. Teams were allowed to submit 5 runs, where each run was required to process only one of the 3 fields of the topic. A maximum of three of these were allowed to use a non-note field, that is the description or summary field.

The highest ranked articles for each topic submitted by the participants were pooled and judged by medical librarians and physicians trained in medical informatics. In particular, the judgment sets were created using two strata: all documents retrieved in ranks 1-20 by any run and a union with a 20% sample of documents not retrieved in the first set that were retrieved in ranks 21-100 by some run. Assessors were instructed to judge articles as "definitely relevant" for answering questions of the specified type about the given case report, "definitely not relevant," or "possibly relevant." The latter judgment may be used if an article is not immediately informative on its own, but the assessor believes it may be relevant in the context of a broader literature review. The teams searched for relevant documents using medical patient case notes. Thirty different case studies subdivided into three types: diagnosis, test and treatment, were used to search the corpus of related documents. It was up to the team's discretion to determine the relevant information contained within these case notes and create a search program that utilized this medical data.

2. TREC 2016 Literature Review

While designing the experimental procedure for this year's clinical support track the team reviewed a significant amount of literature from the 2014 and 2015 tracks. The University of California, Los Angeles (UCLA) implemented the use of a manual run (Garcia-Gathright, et al., 2014). Their manual run utilized domain experts for query expansion. Our work utilized domain experts to annotate last year's queries to improve the performance of framing for our automatic runs. Similarly to UCLA, we also utilized MetaMap, UMLS and Lucene (McCandless et al., 2010). MetaMap is used to both relate biomedical text to the UMLS Metathesaurus and to flag Metathesaurus concepts that are present within biomedical texts. Lucene is a full text search engine library that is composed entirely in Java and is used to build the initial indices on the document corpus.

San Francisco State University (Bhandari et al., 2014) also used MetaMap but they translated their case reports into a list of structured medical concepts. Instead of using this method, we utilized our framing technique to add structure to the abstract of each case report to automatically score the retrieved documents relative to our query.

From the 2015 track we reviewed what Wayne State University did for their system. They were one of the top performing groups from the 2015 track. One concept that we used from their system was their idea of weighting certain query concepts. We implemented this by weighting our symptoms and passing those to Lucene as queries. They based their weighting on semantics or statistics (Balaneshin-kordon et al., 2015). We used machine learning in order to decide on the weights that we used for our symptom queries.

3. The SCDA System Main Components

The main focus of the 2016 SCDA system was to improve the framing component from 2015's system to make it more accurate and useful. This meant utilizing machine learning on the 2015 data to pinpoint areas in which we could make improvements to our Lucene queries and thus provide a better data set for framing. We also used machine learning on the to discover the optimal set of frame attribute weights. The remainder of this paper will discuss the modules of our SCDA system in detail as well as the results of our NIST evaluation.

3.1 Lucene Baselines

In order to run the initial retrieval on the corpus of documents, Apache Lucene 4.0.0 was utilized to create an index. Lucene is an open source search engine, written in Java, designed to function as a text search engine library.

Lucene was also used to generate the baseline run of our system by indexing the entire document and using the summary field of each topic to query our index. In another run we used the note field as the query, first automatically removing various characters that would make the Lucene query parser crash in the example note that was given.

3.1.2 Weighted Lucene run

Once we learned the best weights of symptoms found in the 2015 track using logistic regression, we incorporated these weights in another run by altering our Lucene search at the time of the query. Lucene supports the boosting of terms in a query and we used this feature in two of our runs. We did this by scanning the topic summary for the existence of symptoms using UMLS symptom finder. We then boosted these symptom terms by their corresponding weight. Something worth noting is that Lucene's boosting technique also supports multiple term symptoms. If we found a symptom in our query that was contained within a larger symptom, only the larger phrase was ranked. For example "stomach pain following ingestion" was ranked rather than just ranking "stomach pain" if both were found within the query.

3.2 The Framing Component

We expanded SCDA's Framing component using the 2015 model as a base, adding new attributes and altering the method for scoring. Last year's frame attributes included: age,

gender, time, and symptoms. We expanded these by adding: country, topic-category, and key symptom. In Figures 1-3 below we show a sample query frame for topic #20 where our P(10) = 1.0, as well as two data frames, one with a high score and one with a low score.

Query Frame:

<topic number="20" type="test">
<note>
This is a 87 year old female NH resident with a history of chronic atrial fibrillation, hypertension and hypothyroidism who presents to the [**Hospital Unit Name 10**]. She had been in her usual state of health until 5 days ago when she suddenly began to have abdominal pain. Her abdominal pain was initially intermittent lasting for a few hours at at time. No clear correlation with food. Yesterday, she noticed that her pain was much more severe, [**3301-9-5**] in severity and more localized to the right. This was accompanied by nausea and vomitting. She vomitted twice, with clear liquid emesis and was sent to [**Hospital3 **]. At [**Hospital1 **], she was noted to have elevated amylase/lipase to 538 and 516 with elevated bili to 4.1 and AST/ALT to 198/115 and was given ciprofloxacin, flagyl and 500cc NS and was transferred to the [**Hospital1 1**] emergency department. . At [**Hospital1 1**] EDVS 97.9 HR 83 157/92 RR 18 97% RA. Elderly F, oriented X 2, NAD, flat jvp, CTA decreased b/b, s1 s2 [**Last Name (un) **], decreased BS, + t at ruq, no edema
</note>
<description>
A 87 year old female NH resident with a history of chronic atrial fibrillation, hypertension and hypothyroidism who presents wit abdominal pain. She had been in her usual state of health until 5 days ago when she suddenly began to have abdominal pain. Her abdominal pain was initially intermittent lasting for a few hours at at time. No clear correlation with food. Yesterday, she noticed that her pain was much more severe and more localized to the right. This was accompanied by nausea and vomitting. She vomitted twice, with clear liquid emesis and was sent to a hospital. At the hospital, she was noted to have elevated amylase/lipase to 538 and 516 with elevated bili to 4.1 and AST/ALT to 198/115 and was given ciprofloxacin, flagyl and 500cc NS and was transferred to the emergency department. At the emergency department her vital signs were TM 97.9 HR 83 BP 157/92 RR 18 sat 97% RA.
</description>
<summary>
A 87 yo female reports several days abdominal pain, worse yesterday, severe and more localized to the right, accompanied by nausea and vomitting. Labs show elevated bilirubin, transaminitis, amylase and lipase.
</summary>
</topic>

Figure 1: Topic #20

Topic Number	20
Age	Aged 80
Gender	Female
Time	Null
Country	United States (default)
Topic Category	Test
Key Symptom	Abdominal pain
Symptoms	abdominal pain, elevated bilirubin, nausea

Figure 1: Topic #20's corresponding frame using the summary field (Note we generalized ages to their decade)

Document #2845777 Frame:

Topic Number	20
Score	96.00
Document ID	2845777
Age	Null
Gender	Null
Time	Null
Country	United States (Default)
Topic Category	Test
Key Symptom Match	True
Symptom Matches	2
Symptoms	systemic lupus erythematosus, conditions, minor symptoms, pancreatitis, hepatitis, mesenteric vasculitis, gastrointestinal complications, gastrointestinal vasculitis, ischemia, vasculitis, abdominal pain, peritoneal irritability, oral ulcers, dysphasia, nausea, vomiting, enhanced complications, diffuse abdominal colonic pain, vomiting, illness, fever

Gastrointestinal manifestations are common in patients with systemic lupus erythematosus (SLE), and the incidence of these conditions varies according to the methods of evaluation and the type of manifestation; the incidence of gastrointestinal manifestations varied between 15% and 75% in the majority of previously published studies.1,2 The clinical picture may vary from minor symptoms, such as oral ulcers, dysphasia, nausea and vomiting, to more serious conditions, including pancreatitis, hepatitis and mesenteric vasculitis.

Mesenteric vasculitis is uncommon among gastrointestinal complications (2.2–9.7%) but may cause enhanced complications and mortality if it is not carefully diagnosed and promptly treated.1

Gastrointestinal vasculitis is characterized by ischemia of the digestive tract; this ischemia is caused by the deposition of circulating immune complexes. Clinically, this vasculitis presents with diffuse abdominal or colonic pain, vomiting and fever. The physical examination demonstrates abdominal pain upon touching, which may indicate peritoneal irritability. Gastrointestinal vasculitis does not usually appear as an isolated manifestation but is generally associated with other clinical manifestations of the illness and SLE activity. Consequently, the Systemic Lupus Erythematosus Disease Activity Index (SLEDAI) score is generally high in patients with gastrointestinal vasculitis.1,3–7

In the present study, we describe an unusual case of a lupus patient with mesenteric vasculitis who presented without any evidence of the illness, which is usually found in this clinical situation.

Figure 2: Topic #20 frame scored highly and moved to our top ten by our system with its corresponding text passage from a document judged to be relevant by NIST assessors

Document #4513752 Frame:

Topic Number	20
Score	40.04
Document ID	4513752
Age	Null
Gender	Null
Time	Chronic
Country	Germany
Topic Category	Null
Key Symptom Match	False
Symptom Matches	0
Symptoms	disease, known syndromes, haemorrhagic fever, renal syndrome, milder subtype nephropathia epidemica, cardiopulmonary syndrome hfrs, hantavirus infection, yosemite national park, causative agent, kidney injury, thrombocytopenia, nephropathia epidemica

The most common causative agent for hemorrhagic fever with renal syndrome in Germany is Puumala virus (PUUV) and a high percentage of patients with PUUV infection have gastrointestinal (GI) symptoms. The aim of the present study was to determine the prevalence of increased lipase levels and acute pancreatitis during nephropathia epidemica (NE) in 166 patients from Germany.

Clinical and laboratory data during the acute phase of the disease were obtained from medical reports and files from 456 patients during acute hantavirus infection. Patients in whom serum lipase levels were determined during acute course of the disease were included in the study.

Lipase levels at the time of diagnosis were determined in 166 of the 456 NE patients (36 %). Of the 166 patients, 25 (15 %) had elevated lipase levels at the time of admission to hospital or first contact with general practitioner/nephrologist. In total 7 patients had a threefold increased serum lipase above the normal range. Abdominal pain was not more often present in the group of patients with elevated serum lipase compared to the lipase-negative group (9/25 vs 58/141). Abdominal ultrasound and CT scans revealed no signs of pancreatitis in any of the patients. Patients with elevated serum lipase had higher serum creatinine peak levels ($p = 0.03$) during the course of the disease.

Elevated lipase levels were common in our patient cohort and might reflect a more severe form of NE. NE does not lead to acute pancreatitis.

Figure 3: Topic #20 frame scored low and removed from our top ten by our system with its corresponding text passage from a document judged to be not relevant by NIST assessors

In order to create frames from queries and passages of text, the text was taken through a number of different steps. First, MetaMap was used on the text to generate a list of negated concepts. For example, upon processing the phrase *"cardiac arrest was ruled out"*, the function would add to the negated list any concept triggered in MetaMap for the frame *"cardiac arrest"*. Later, any concept in the candidate target concept list that matched a concept negated in the same phrase was removed. The text was further automatically modified to replace potentially problematic phrases, especially those that would cause problems for the parser (for example, the Latinate medical terminology "status post" was replaced with "after") based on a dictionary we generated from 2014 analysis.

The text was then run through the Stanford Parser, in order to detect semantic roles and relationships. The parser's output was stored as a set of hierarchical clauses. This clausal hierarchy was searched for words that triggered concepts using MetaMap. Using the typology of "semantic types" employed by MetaMap to categorize triggered concepts. If trigger concepts were found with one of eight designated types, the relevant concept was added to the symptom list variable for the frame of the larger given area of text. For example, the sentence *"64-year-old woman with uncontrolled diabetes, now with an oozing, painful skin lesion on her left lower leg"* would have, among its many triggered concept referents from Metamap's database, a concept referent for skin lesions, likely classed under the semantic class [anab] (Anatomical Abnormalities). Since [anab] is one of the designated semantic types for denoting symptoms, the noun clause containing it, *"oozing, painful skin lesion"* is added to the symptoms list.

Referring to the temporality typology suggested by the medical professionals employed by the UCLA team in 2014, our frame's time attribute functions to classify conditions into classes of "*acute*", "*progressive*" and "*chronic*". The text of each triggered symptom clause was searched for temporal wording describing the symptom, and if it was found, the appropriate time class was saved to the frame's time attribute.

The country attribute was determined by searching the document or topic for any match with the Java system locales. If no match was found, the attribute is set to "United States" as a default. The topic category attribute was filled using a process similar to finding the time attribute. The entire document was searched for wording relating to one of the three topic categories: diagnosis, test, or treatment. When this wording was found, the frame would be assigned the related category. The highest ranked symptom (according to the symptoms list derived from machine learning) from among the query frame's symptoms becomes the key symptom attribute. If the document frame's symptom list contains a match to the key symptom, the key symptom match is set to true.

3.2.1 Frame Scorer

After the Framing process was complete, SCDA had to rank each frame created by a document passage in order of its relevance to the query frame. The 2015 scoring algorithm simply looked for equality of the contents of each frame attribute. The total score of the frame was then calculated as the average of the scores from each individual frame attribute. In the 2016 scoring algorithm, the total score of the frame depends on

which individual frame attributes are matches. For example, a match in the Key Symptom attribute is weighted more heavily than a match in the Gender attribute. The individual weight for each attribute was determined using Machine Learning as described next.

In our error analysis of the 2015 results, we discovered that we could improve the way we were previously determining whether two given symptoms from the query frame and a document frame were a match. MetaMap generates a list of content phrases, which are then checked against the topics symptoms to determine if the symptoms are a match. In 2015, symptoms were scored as a match if the content phrase was equal to a symptom from the query frame. In our error analysis we determined that symptoms such as "vomit" and "vomiting" were erroneously considered to not match. In 2016, we compared the two strings using Apache codec implementation of the Double Metaphone algorithm. The Double Metaphone algorithm improves upon the original metaphone algorithm which uses information variations and inconsistencies in English language and does a better job of matching words and phrases that sound familiar. After the double metaphone comparison has taken place, we then determine the longest common substring found between the two strings. If the substring length is larger than half of the smaller of the two input strings and the result of the Double Metaphone comparison is zero then the Frame Scorer will score the two symptoms as a match.

Example of 2015 SCDA Scoring algorithm:

Query Frame:

Gender	Female
Age	Child
Symptoms	Cough, Chest Pain, Left Lung Mass

Document Frame:

Gender	Undetected	Gender Score:	null
Age	Child	Age Score:	100
Symptoms	Cough, Chest Pain, Fever	Symptoms Score:	66
		Total Score:	83

After several rounds of analyzing results using the 2015 data we made another modification to our scoring algorithm. This change to our scoring algorithm lies in the way we treated frames when certain data types were not populated. For example, in the 2015 version of our scoring algorithm, when the query frame detected the gender of the patient, and the document we were scoring it against did not mention a gender (or our frame builder failed to locate it), we would not include that attribute in the calculation of the final overall score. In the 2016 version we instead assigned a score of 0 for that type.

Example of 2016 SCDA Revised Scoring algorithm:

Query Frame:

Gender	Female
Age	Child
Symptoms	Cough, Chest Pain, Left Lung Mass

Document Frame:

Gender	Undetected	Gender Score:	0
Age	Child	Age Score:	100
Symptoms	Cough, Chest Pain, Fever	Symptoms Score:	66
		Total Score:	55.33

3.3 Machine Learning

SCDA 2016 utilized WEKA for our machine learning. We tested out many different variables and algorithms to try and see what we could utilize in order to improve our accuracy. We completed three different experiments using the output from framing. Experiment one contained the attributes; docID, genderMatch, ageMatch, timeMatch, diagnosisMatch, categoryMatch, countryMatch, symptomMatches, and topic number. Experiment two contained the same attributes as one, with a substitution of symptomScore in place of symptomMatch. Experiment three was much more simplified and contained the attributes; docID, genderMatch, ageMatch, timeMatch, symptomMatches, and topic number. We wanted to see if we could find a pattern between any of these attributes and relevant and non-relevant documents. We did this by running framing on the 2015 topics and documents and using those attributes from the frames in the machine learning to see how WEKA used the attributes to predict relevant and non relevant documents. There were some patterns that we saw. For example, genderMatch, ageMatch, and timeMatch were consistently weighted the highest. Though utilizing the results of these experiments did not realize a significant improvement in our scores when we applied these weights to our frame scoring algorithm, so we had to look for a different approach.

We thought to try and just use the symptoms to help us make our frame scorer more precise. We decided to take every symptom that was identified during framing and use them in our machine learning. Each symptom was given a weight by WEKA based on how often it appeared in relevant versus non-relevant documents. After getting this weighted list we ordered it based on the weight it was given for relevant documents. The higher the weight, the more chance there is that symptom appeared in a document classified as relevant. We ended up using this list to help make our frame scorer more precise because during our testing incorporating this data we did see a significant boost in our 2015 precision. We also added these symptom weights to one of our Lucene runs. We used the weighted symptoms list to help add different symptoms to the queries that we passed to Lucene.

4. The SCDA Architecture

In a standard run, we used Lucene as described above to generate a list of the top 20 documents for a topic. This list, containing document ids and scores, is passed to the Framer. The topic is framed to create a Query Frame, and each returned document's abstract is framed and then scored against the Query Frame. The Framer returns a new re-ranked list of the highest scoring documents based on their frame's score. Finally, the weights learned from the 2015 machine learning process are applied to further improve the ranked results.

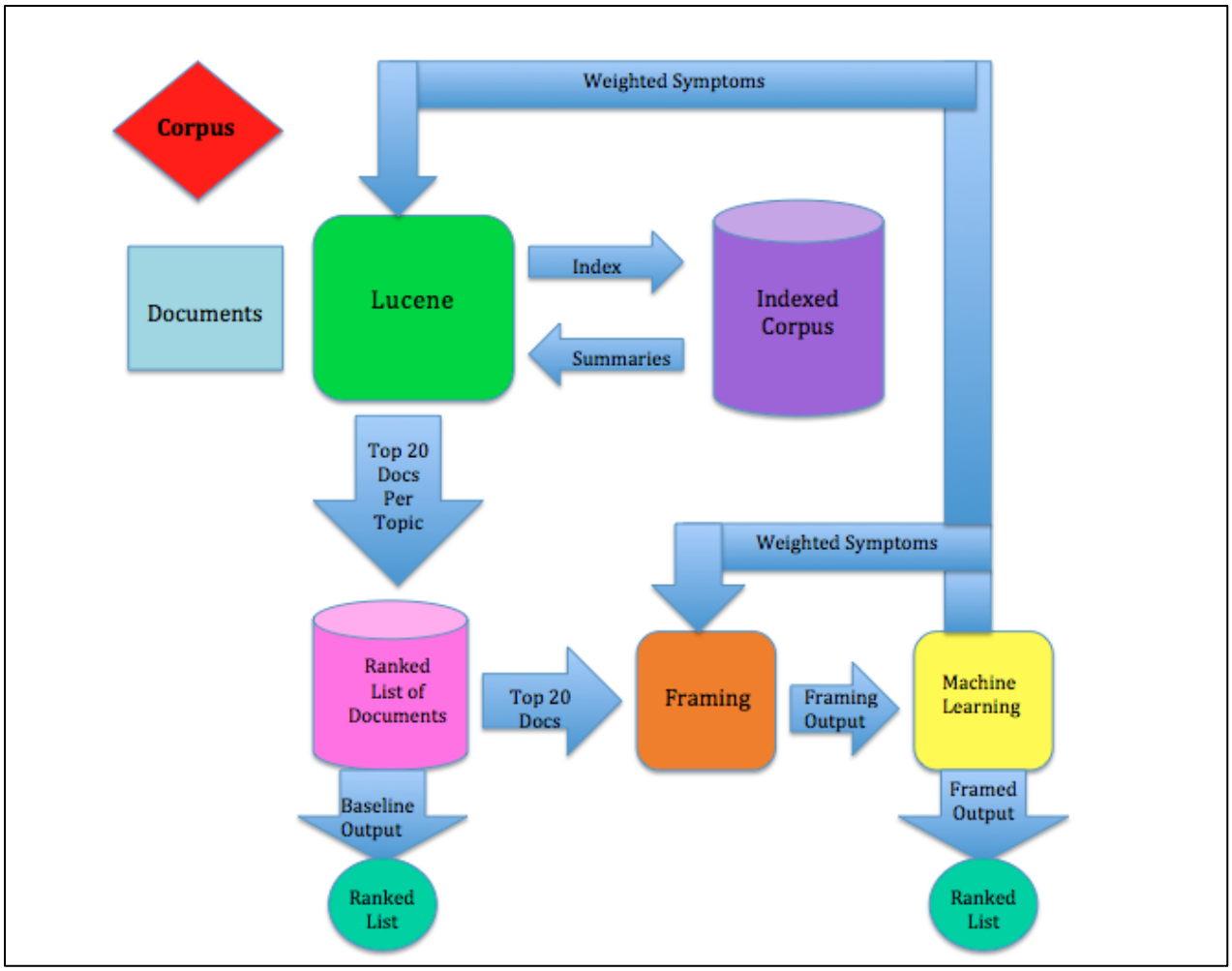

Figure 4: SCDA Architecture

6. Early Conclusions

Lucene Baselines
Our initial Lucene queries performed over 10% worse than during the 2015 track. This is most likely due to the different and more realistic nature of the topics.

Eleven topics in the baseline did not return any relevant documents. Due to the nature of the system, in the situation where Lucene generates no relevant documents, framing of course cannot improve the output, (specifically, topics 2, 4, 5, 7, 10, 12, 15, 19, 21, 23, and 27).

In our error analysis the eleven topics where P(10)=0.0 were examined first. A common thread between these topics was the use of semi-complex and more rare medical terminology: *interparenchymal hemorrhages (topic 12), tracheobronchomalacia (topic 15), biphenotypic ALL (topic 21)*. Many of these symptoms were not present or common in the 2015 and 2014 track and our system was not designed with this degree of medical terminology in mind. Another new addition to the 2016 topics was the common use of

medical acronyms: "*78 year old female with PMHx HTN, dCHF, Diabetes, CKD*" (topic 19), "*94 M with CAD s/p 4v-CABG, CHF, CRI presented with vfib arrest*" (topic 22). These acronyms may have held valuable information which would have directed our system to more relevant documents that were unfortunately passed over.

It is also possible that the version of Lucene used in our system (version 4.0.0) is not suited to handle some of these terms as it is not the latest release and was built to run with Java 6. Using a more recent release of Lucene could yield better results. Furthermore, the use of other indexing/querying software such as Indri might be very helpful. If other querying software yielded different ranked lists, this could be extremely helpful for our framing. Finally, returning a larger list of ranked documents in Lucene could potentially improve the framing but would greatly increase the time needed to frame and score documents. If time allowed for the framing and scoring of the top 100 documents we may have found relevance after framing for our lower scoring topics.

Machine Learning

Machine learning was a new method that we tried out with our system that we had not tried in the previous year. In our initial testing we saw an increase in precision when we implemented the weights that we found in machine learning. It is hard to say how much machine learning helped in our 2016 results without further experiments on the 2015 results. But given that our Lucene baseline went down from last year (2015 our Lucene P(10) was 0.3767) to this year (2016 our P(10) was 0.23) but our framing score went slightly up, from 0.2667 to 0.2770, implies our framing changes realized a strong improvement. Without further analysis it is hard to say what made this jump in precision occur. It could have been the machine learning that we implemented or it could have been some other minor tweaks that were done in our framing process.

We also saw an improvement between the Lucene run with and without the weights. It was a positive change. In the Lucene run without weights our P(10) was .2300 and in the Lucene run with weights incorporated from machine learning our P(10) was .2467.

7. References

Balaneshin-kordan, Saeid, Alexander Kotov, and Railan Xisto. 2015. WSU-IR at TREC 2015 Clinical Decision Support Track: Joint Weighting of Explicit and Latent Medical Query Concepts from Diverse Sources.

Bhandari, Aayush, James Klinkhaer and Anagha Kulkarni. 2014. San Francisco State University at TREC 2014: Clinical Decision Support Track and Microblog Track. In Proceedings of The Twenty-Third Text Retrieval Conference (TREC 2014).

Cormack, G. V., Clarke, C. L. A., and Buttcher, S. 2009. Reciprocal Rank Fusion outperforms Condorcet and Individual Rank Learning Methods.

Garcia-Gathright, Jean I., Frank Meng and William Hsu. 2014. UCLA at TREC 2014 Clinical Decision Support Track: Exploring Language Models, Query Expansion, and Boosting. In Proceedings of The Twenty-Third Text Retrieval Conference (TREC 2014).

Mourão, André, Flávio Martins and João Magalhães. 2014. NovaSearch at TREC 2014 Clinical Decision Support Track. In Proceedings of The Twenty-Third Text Retrieval Conference (TREC 2014).

Mourão, André, Flávio Martins, and João Magalhães. 2013. NovaSearch at TREC 2013 Federated Web Search Track: Experiments with rank fusion. In Proceedings of The Twenty-Second Text Retrieval Conference (TREC 2013).

McCandless Michael, Erik Hatcher and Otis Gospodnetic. 2010. Lucene in Action. Second Edition. Manning Publications.

National Library of Medicine (NLM). 2013. MetaMap- A Tool For Recognizing UMLS Concepts in Text. Software.

Simpson, Matthew S., Ellen M. Voorhees and William Hersh. 2014. Overview of the TREC 2014 Clinical Decision Support Track. In Proceedings of The Twenty-Third Text Retrieval Conference (TREC 2014).

Small, Sharon and Tomek Strzalkowski. *2004. HITIQA: A Data Driven Approach to Interactive Analytical Question Answering.* Proceedings of Human Language Technology Conference. Boston, Massachusetts.

Voorhees, Ellen M. 2007. Overview of TREC 2007. In Proceedings of The Sixteenth Text Retrieval Conference (TREC 2007).