UTD at TREC 2014: Query Expansion for Clinical Decision Support

Travis Goodwin and Sanda M. Harabagiu
Human Language Technology Research Institute
University of Texas at Dallas
Richardson TX, 75080
{travis,sanda}@hlt.utdallas.edu

Abstract

This paper describes the medical information retrieval (MIR) systems designed by the University of Texas at Dallas (UTD) for clinical decision support (CDS) which were submitted to the TREC 2014. We investigated the impact of various knowledge bases for automatic query expansion in the four officially submitted runs. Each of these systems exploits both Wikipedia and PubMed corpus statistics in order to automatically extract keywords. Extracted keywords were then expanded by relying on structured medical knowledge bases, such as the Unified Medical Language System (UMLS), the Systemized Nomenclature of Medicine – Clinical Terms (SNOMED-CT), and Wikipedia as well as unsupervised distributional representations based Google's Word2Vec deep learning architecture. Our highest scoring submission achieved an inferred AP score of 0.056 and an inferred NDCG score of 0.205.

1 Introduction

In order to make biomedical information more accessible to physicians, and to anticipate their medical information needs, the Text REtrieval Conference (TREC) initiated the Clinical Decision Support (CDS) track which aimed to simulate the requirements of Medical Information Retrieval (MIR) systems and to encourage the creation of tools and resources necessary for their implementation. With this goal in mind, the focus of the 2014 track was the retrieval of biomedical articles relevant for answering generic clinical questions about medical "topics". The medical topics for the TREC-CDS 2014 track were medical case narratives authored by experts at the U.S. National Library of Medicine that served as idealized representations of actual medical records. These medical case reports consist of both a well-formed narrative summarizing the portions of a patient's medical record that are pertinent to the case, as well as one of three possible expected answer types. These answer types refer to the need to find out the diagnosis, the treatment or the test best suited for the patient described in the medical case. An example of a medical topic evaluated in TREC-CDS 2014 is provided in Table 1.

2 The Architecture

The MIR-CDS systems designed for the TREC-CDS track in 2014 used the open access subset of the PubMed Central[1] (PMC) collection of scientific articles as retrieved on January 21, 2014. This collection contained a total of 733,138 articles. MIR-CDS systems that participated in the TREC-MDS 2014 track were challenged with retrieving from this collection the scientific articles which were relevant to a given

[1] PMC is an online digital database of freely available full-text biomedical literature.

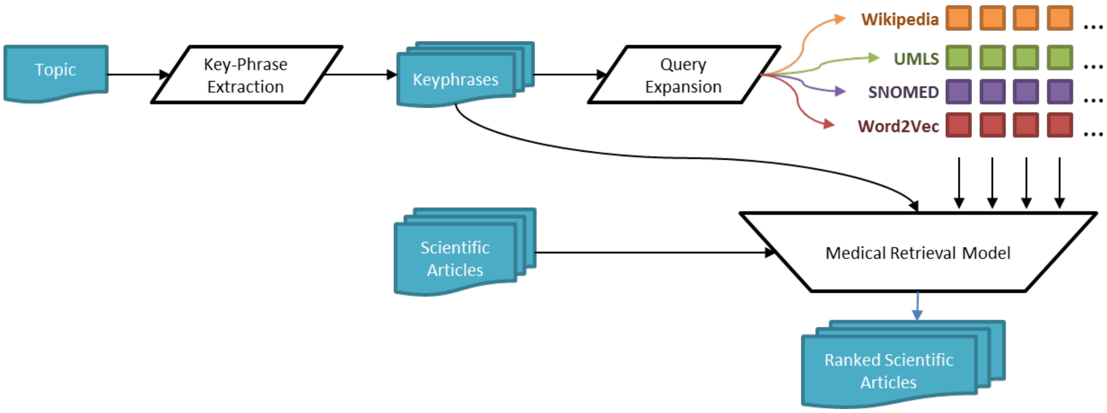


Figure 1: System Architecture for Medical Information Retrieval (MIR) Clinical Decision Support (CDS)

Topic 8 [**Diagnosis**]
Description: A 62-year-old man sees a neurologist for progressive memory loss and jerking movements of the lower extremities. Neurologic examination confirms severe cognitive deficits and memory dysfunction. An electroencephalogram shows generalized periodic sharp waves. Neuroimaging studies show moderately advanced cerebral atrophy. A cortical biopsy shows diffuse vacuolar changes of the gray matter with reactive astrocytosis but no inflammatory infiltration.
Summary: 62-year-old man with progressive memory loss and involuntary leg movements. Brain MRI reveals cortical atrophy, and cortical biopsy shows vacuolar gray matter changes with reactive astrocytosis.

Table 1: Example of a medical topic used in the TREC-CDS 2014 dataset

medical topic. Retrieved articles were judged relevant if they provided information of the specified semantic answer type useful for the given medical case description.

In this paper, we describe the four systems we have designed for the MIR-CDS evaluation offered by NIST in the 2014 TREC-CDS track. The remainder of this paper is outlined as follows. Section 2 documents the system architecture of the MIR-CDS systems we have developed. Section 3 explains how we automatically extract keywords from a given medical topic, Section 4 details the query expansion methods we have evaluated, and Section 5 illustrates the scientific article retrieval models we have used. In Section 6, we present the experimental evaluation which is then analysed in Section 7. Finally, Section 8 summarizes the conclusions.

3 System Architecture

The design of our MIR-CDS system benefits from our past participation in the 2011 and 2012 TRECMed evaluations, where we designed a medical information retrieval system for patient cohort identification [5, 6]. As illustrated in Figure 1, the MIR-CDS system is provided with (a) a medical topic, and (b) a collection of scientific articles and it is expected to return a ranked list of scientific articles which are relevant to the topic. Given the medical topic, we have designed a key-phrase extraction module which uses basic syntactic analysis as well as Wikipedia information to discover multi-word medical key-phrases. Examples of medical key-phrases extracted by this module are: “normocytic anemia” and “coronary artery aneurysm”. We then considered the expansion of medical key-phrases according to several medical knowledge bases as well as distributional information. This allowed us to perform scientific article retrieval using two methods: (1) Lucene-based Boolean retrieval, and (2) distributional vector retrieval.

4 Keyword Extraction

We represent a given medical topic by transforming the medical case description into an unordered set of medical key-phrases (MKPs). Our process for extracting MKPs consists of three steps:

1. Find in the medical case description the longest non-overlapping sequences of words which correspond to titles of articles from Wikipedia, denoted as MKP_1.

2. Syntactically process the medical case description with the shallow parser available from OpenNLP [1] and derive all the noun phrases, generating a list of medical key-phrases MKP_2

3. Combine the entries in MKP_1 and MKP_2 creating MKP_{1+2} and filter it by removing any entry which occurs either in less than 5 or more than 60% of the scientific articles from the TREC-CDS dataset.

This allows us to remove rarely occurring or highly common medical key-phrases while retaining useful multiple-word expressions such as "coronary artery aneurysm."

5 Query Expansion

In written text, and especially in medical records, the morphology of words varies significantly both between and within documents. In order to account for these variations in text, we store each extracted keyword in several forms: its original surface form, a WordNet [4] lemmatized form, an unabbreviated form (based on a list of common medical abbreviations), a form in which hyphens are padded with spaces, a form with all hyphens replaced with spaces, and a form with all punctuation removed. These forms are used as exact synonyms for the purposes of keyword expansion and retrieval.

Unfortunately, these slight variations in morphology are not enough to capture the diverse ways in which keywords may be expressed in medical records. Indeed, the topics presented in this task often require extensive domain knowledge within the field of medicine to be properly understood. For example, a keyword such as *fatigue* may be referred to as *weariness*, *lack of energy*, *malaise*, or *feeling tired.* These synonyms all denote hearing loss, but use alternate phrasings. In our approach, we consider all of these semantically related words as 'expansions', and we refer to the process of generating them as query expansion. The following subsections describe the four methods of query expansion that we have considered.

UMLS Expansions of *fatigue*)
weariness
lack of energy
fatigue extreme
fatigues
time tired
tired all the time
lacking in energy

UMLS Expansion

The Unified Medical Language System (UMLS) is a resource for coordinating health and medical vocabularies [3]. UMLS contains three major components: the "Metathesaurus" (which includes data from SNOMED, RxNorm, MeSH, and other collections), the "Semantic Network" which provides general categories and relationships, and the "SPECIALIST Lexicon and Lexical Tools"[2]. In the UMLS Metathesaurus, each unique medical concept is associated with a "concept unique identifier" or CUI such that all entries in the Metathesaurus which refer to the same concept share the same CUI. We exploit this information by expanding each key-phrase so that it

[2] UMLS is described at *http://www.nlm.nih.gov/research/umls/quickstart.html.*

includes all concepts in the MetMetathesaurus which share the same CUI.

Wikipedia Expansions of *fatigue*
fatique
fatigue failure
tatt
tired all the time syndrome
chronic fatigue

Wikipedia Redirect Expansion

Wikipedia is a common resource for natural language processing tasks. The English version is comprised of 3,731,340 user-generated articles covering almost any notable topic[3]. In addition to articles, Wikipedia also contains pages called *redirects* which do not contain content themselves, but rather redirect – or send – the user to another article (or section of an article). Redirects typically embody alternate names, spellings, forms, closely related words, alternately punctuated or encoded forms, less specific forms in which the redirected name is the primary topic, or more specific forms of some other page. Our system uses Wikipedia redirects[4] to generate synonyms, and alternate (or mis-) spellings for keywords. We do this by expanding each keyword so that it includes all article titles that redirect (send the user) to the same article as the key-phrase.

SNOMED CT Expansion

The National Library of Medicine provides a resource called the Systematized Nomenclature of Medicine–Clinical Terms (SNOMED CT) [9]. SNOMED CT is a clinical terminology being maintained by the International Health Terminology Standards Development Organisation and is one of several designated standards used in United States Federal Government systems for the electronic exchange of medical records[5]. SNOMED CT catalogs both medical concepts and various relationships between them. We leverage these relationships by expanding a medical key-phrase so as to include all SNOMED CT entries which partake in the child side of an IS_A or PART_OF relationship[6] with the key-phrase.

SNOMED Expansions of *fatigue*
complaining of debility and malaise
frailty
heavy feeling
senile asthenia
tired
sensation of heaviness in limbs
psychogenic fatigue
feeling tired
attacks of weakness
asthenia

Word2Vec Expansion

In addition to medical ontologies, we have also considered the role of unsupervised distributional semantics for query expansion. We used the publicly available Gensim library provided in [7] to train a Word2Vec distributional word representation using the TREC-CDS scientific article corpus. Word2Vec is a highly sophisticated deep-learning neural network architecture which learns how to represent words as multi-dimensional vectors. The model operates without human supervision by considering the textual context surrounding words. These contexts may be represented in a variety of ways, although in this work we utilize the Skip-gram model which is able to capture discontinuous multi-word sequences. Word2Vec attempts to project words onto a multi-dimension vector space such that the proximity between two vectors indicates the semantic "similarity" between their associated words. We used this property by

[3] The number of articles is based on September 6, 2011.

[4] Redirect and article data is based on the May 26, 2011 English Wikipedia data dump.

[5] More information on SNOMED CT is available at *http://www.nlm.nih.gov/research/umls/Snomed/snomed_main.*

[6] We consider children up to 2 levels (grandchildren) from the parent concept.

Word2Vec Expansions of *fatigue*
tiredness
malaise
sickness
instability
neuropathy

associated each key-phrase with its vector representation and using the twenty most "similar" (as measured by cosine similarity) words in the vocabulary as expansions.

6 Scientific Article Retrieval

Lucene-based Scientific Article Retrieval

We constructed an index of the scientific articles using Apache Lucene[7]. When representing a medical case topic, we represented each key-phrase as a disjunctive Boolean query containing all of its possible expansions. This allowed us to represent the entire topic by combining each medical key-phrase query into one large Boolean query. As such, we determined the relevance of a scientific article to a medical topic by determining the BM25 score achieved for each query using Apache Lucene's implementation of the BM25 ranking function [8].

Distributional Scientific Article Retrieval

Latent Dirichlet Allocation (LDA) presents a way to efficiently learn latent topic distributions of words automatically from a dataset [2]. We trained an LDA representation on the scientific article collection used for TREC-CDS. In Figure 2, we illustrate the first fifteen words and their weights in some of the inferred latent topics. We then used our trained LDA model to represent each query and document as their latent topic distribution vectors. This allowed us to determine the relevance of any document and query pair by calculating the cosine of the angle between their latent topic distribution vectors.

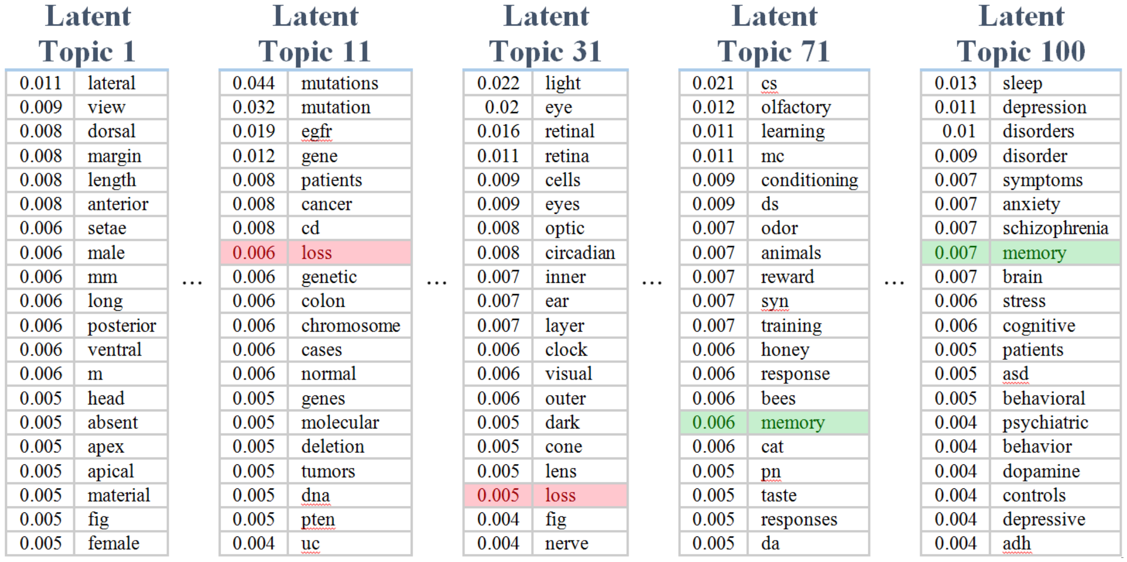
7 Performance Evaluation

For our experiments, we use the thirty topics authored for the TREC-CDS task in 2014. Twenty-six groups participated in the TREC-CDS task, yielding a total of 71 fully-automated systems and 11 "manual" submissions which incorporated user-intervention. The scientific articles collected across all 85 submissions were pooled and manually judged to determine their relevance the associated topic. Supervised by the Oregon Health & Science University (OHSU), physicians judged the twenty top-ranked documents returned by each system as well as a 20% random sample of the retrieved documents between ranks 21 and 100, yielding a total of 37,949 topic-article pairs. In our experiments, we evaluate the performance achieved on this task according to these relevance judgments.

We submitted four official "runs" to the 2014 TREC-CDS task. Our first system, NQE, captures the performance achieved when no query expansion is performed. In this way, standard Lucene-based BM25 retrieval is performed based on the key-phrases extracted. Our second system, LDA, captures the performance when Latent Dirichlet Allocation (LDA) is used to determine the relevance between a medical topic and a scientific article. Our third system, W2V, captures the performance when Word2Vec (W2V) a state-of-the-art distributional semantic approach is used to expand queries. Our fourth system, MIR, captures the performance when all query expansion techniques except Word2Vec (UMLS, SNOMED, Wikipedia) are used to expand each medical key-phrase.

In our evaluations, we consider four different performance measures of the quality of retrieved scientific articles. The inferred Average Precision (infAP) estimates the expected average precision for a topic when only a subset of the retrieved articles have relevance judgments [10]. Likewise, the inferred Normalized Discounted Cumulative Gain (infNDCG) extends the commonly used NDCG measure to account

[7] Apache Lucene is an open source high-performance search engine library written in Java. Apache Lucene is available at `http://lucene.apache.org/core/`



Latent Topic 1		Latent Topic 11		Latent Topic 31		Latent Topic 71		Latent Topic 100	
0.011	lateral	0.044	mutations	0.022	light	0.021	cs	0.013	sleep
0.009	view	0.032	mutation	0.02	eye	0.012	olfactory	0.011	depression
0.008	dorsal	0.019	egfr	0.016	retinal	0.011	learning	0.01	disorders
0.008	margin	0.012	gene	0.011	retina	0.011	mc	0.009	disorder
0.008	length	0.008	patients	0.009	cells	0.009	conditioning	0.007	symptoms
0.008	anterior	0.008	cancer	0.009	eyes	0.009	ds	0.007	anxiety
0.006	setae	0.008	cd	0.008	optic	0.007	odor	0.007	schizophrenia
0.006	male	0.006	loss	0.008	circadian	0.007	animals	0.007	memory
0.006	mm	0.006	genetic	0.007	inner	0.007	reward	0.007	brain
0.006	long	0.006	colon	0.007	ear	0.007	syn	0.006	stress
0.006	posterior	0.006	chromosome	0.007	layer	0.007	training	0.006	cognitive
0.006	ventral	0.006	cases	0.006	clock	0.006	honey	0.005	patients
0.006	m	0.006	normal	0.006	visual	0.006	response	0.005	asd
0.005	head	0.005	genes	0.006	outer	0.006	bees	0.005	behavioral
0.005	absent	0.005	molecular	0.005	dark	0.006	memory	0.004	psychiatric
0.005	apex	0.005	deletion	0.005	cone	0.006	cat	0.004	behavior
0.005	apical	0.005	tumors	0.005	lens	0.005	pn	0.004	dopamine
0.005	material	0.005	dna	0.005	loss	0.005	taste	0.004	controls
0.005	fig	0.005	pten	0.004	fig	0.005	responses	0.004	depressive
0.005	female	0.004	uc	0.004	nerve	0.005	da	0.004	adh

Figure 2: The fifteen most likely words associated with selected latent topics produced by the Latent Dirichlet Allocation (LDA) on the TREC-CDS corpus of scientific articles. The position of words "loss" and "memory" has been highlighted in red and green, respectively.

System	infAP	infNDCG	R-Prec	P@10
NQE	0.056	0.205	0.170	0.307
LDA	0.002	0.028	0.015	0.027
W2V	0.044	0.173	0.137	0.263
MIR	0.056	0.201	0.167	0.327

Table 2: Performance Evaluation for official submissions to TREC-CDS 2014

for incomplete relevance judgments [11]. We also list the R-Precision (R-Prec) denotes the precision at the R-th position in the ranking results where R is the number of relevant documents, and the precision of the first 10 documents (P@10). Table 7 summarizes the scores provided by NIST for our four submissions. Additionally, due to the complexity of each topic, we present the performance of our system on a per-topic basis in Figure 3.

8 Analysis

As shown in Table 7, the NQE and MIR systems achieve comparable performance. This suggests that incorporating medical knowledge bases does not yield a significant increase in the quality of retrieved scientific articles without further processing. The poor performance of the Word2Vec module likewise suggests that more analysis is needed to ascertain how many expansions should be considered, and how those expansions should be weighted. Finally, the LDA baseline performs exceptionally poorly, indicating that more sophisticated latent topic modelling approaches should be considered.

As shown in Figure 3, the performance of our three Lucene-based systems are vary similarly between topics. Clearly, our best performance is achieved on Topic 4. Our weakest performance performance is experienced on Topics 8 and 12, which will be analysed in the following sub-sections.

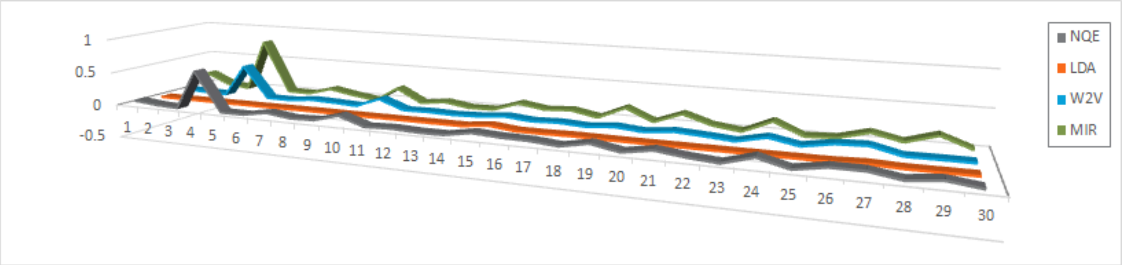


Figure 3: Per-topic performance comparison of all four submitted systems as measured by the inferred Average Precision (infAP).

Key-phrase Analysis

The key-phrase detection algorithm we have described in Section 2.1 leverages two sources of information to determine likely key-phrases: Wikipedia and syntactic noun phrases. This allows us to anticipate sparsity in Wikipedia and mitigate noise from syntactic analysis. Additionally, we preempt further noise by filtering according to corpus statistics to remove very rare or highly frequent candidate key-phrases in the scientific article collection. Despite these measures, errors in automatically determining key-phrases still arise. Consider, for example, Topic 8, as given in Table 1.

Our key-phrase extraction system extracts the key-phrases “memory loss”, “brain MRI”, “atrophy”, “gray matter”, “astrocytosis”, “neurologist”, “jerking”, “lower extremities”, “neurologic examination”, “cognitive deficits”, “electroencephalogram”, “sharp”, “wave”, “neuroimaging”, “cerebral atrophy”, and “gray matter”. Although all these phrases are important to diagnosing the patient described in the topic, a significant amount of semantic meaning is lost when the key-phrases are removed from their contexts. For example, the presence of the term “neurologist” is unlikely to convey the same impact to a document’s relevance as the presence of “astrocytosis.” Likewise, there is a subjective difficulty in determining the ideal boundary for each extracted keyphrase: should a system consider “astrocytosis” or “reactive astrocytosis”? Is “reactive astrocytosis” a specific class of astrocytosis, or simply a description? In order to answer these types of questions, more medical information must be considered when determing key-phrase boundaries. Unfortunately, due to the extraordinary variation possible in query expansion and relevance model implementations, the ideal representation is unlikely to be objectively true for all systems. As such, it is difficult to measure the cost-vs-reward of incorporating more sophisticated forms of query representation. For this reason, we focus primarily on the impact of individual query expansion methods.

Query Expansion Analysis

Using the medical key-phrase “fracture”, from topic 12, it is clear that UMLS and SNOMED provide the largest number of potential expansions. However, many the expansions provided by UMLS consist of phrasal expressions (e.g. “tired all the time”) which are unlikely to appear in scientific discourse. Likewise, the expansions from SNOMED are often highly verbose (e.g. “complaining of debility and malaise”) and need further language processing in order to use – in fact, it might be reasonable to consider these expansions as new “queries” with their own individual key-phrases (“debility” and “malaise”) which need further expansion. The Wikipedia expansions, on the other hand, are much more concise and easier to locate within the scientific article collection. Finally, the Word2Vec expansions are surprisingly potent – capturing many of the terms yielded by structured knowledge bases such as “tiredness” or “malaise”. Unfortunately, the expansions also quickly degenerate into co-occurring words which are not directly related to the medical key-phrase.

9 Conclusions

The Clinical Decision Support (CDS) track introduced in the 2014 Text REtrieval Conference (TREC) evaluated the problem of retrieving scientific articles relevant to a medical case description. This entailed a wide variety of difficulties, such as working with highly complex medical texts for which each term requires a high degree of domain knowledge to comprehend, and the lack of training data imposed by the fact that is the first iteration of this track. Given these difficulties, we considered four approaches to this task: (1) a straightforward BM25 approach using Wikipedia-inspired key-phrase detection, (2) an extended approach incorporating medical knowledge bases for query expansion, (3) a further extended approach which additionally incorporates unsupervised corpus statistics based on distributional word embeddings, and (4) a separate approach which considers the cosine similarity between latent topic model (LDA) representations of queries and documents. Our first and second submissions achieved promising performance, while the Word2Vec and LDA-based systems were the weakest. For future improvement, we plan to investigate how to properly incorporate medical knowledge bases, as well as latent knowledge for query expansion.

Bibliography

[1] Jason Baldridge. The opennlp project. *URL: http://opennlp. apache. org/index. html,(accessed 2 February 2012)*, 2005.

[2] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022, 2003.

[3] Olivier Bodenreider. The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 32(suppl 1):D267–D270, 2004.

[4] Christiane Fellbaum. *WordNet*. Wiley Online Library, 1998.

[5] Travis Goodwin, Bryan Rink, Kirk Roberts, Sanda M Harabagiu, and R Tx. Cohort shepherd: Discovering cohort traits from hospital visits. In *Proceedings of The 20th Text REtrieval Conference*. Citeseer, 2011.

[6] Travis Goodwin, Kirk Roberts, and Sanda M Harabagiu. Cohort shepherd ii: Verifying cohort constraints from hospital visits. Technical report, DTIC Document, 2012.

[7] Radim Řehůřek and Petr Sojka. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta, May 2010. ELRA. `http://is.muni.cz/publication/884893/en`.

[8] Stephen E Robertson, Steve Walker, Susan Jones, Micheline M Hancock-Beaulieu, Mike Gatford, et al. Okapi at trec-3. *NIST SPECIAL PUBLICATION SP*, pages 109–109, 1995.

[9] Michael Q Stearns, Colin Price, Kent A Spackman, and Amy Y Wang. Snomed clinical terms: overview of the development process and project status. In *Proceedings of the AMIA Symposium*, page 662. American Medical Informatics Association, 2001.

[10] Emine Yilmaz and Javed A Aslam. Estimating average precision with incomplete and imperfect judgments. In *Proceedings of the 15th ACM international conference on Information and knowledge management*, pages 102–111. ACM, 2006.

[11] Emine Yilmaz, Evangelos Kanoulas, and Javed A Aslam. A simple and efficient sampling method for estimating ap and ndcg. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 603–610. ACM, 2008.