Published in final edited form as:
Int J Comput Biol Drug Des. 2013 ; 6(0): 5–17. doi:10.1504/IJCBDD.2013.052198.

A New Iterative Method to Reduce Workload in the Systematic Review Process

Siddhartha Jonnalagadda
Department of Health Sciences Research Mayo Clinic Rochester, Minnesota 55902, USA
siddhartha@mayo.edu

Diana Petitti
Department of Biomedical Informatics Arizona State University Tempe, Phoenix 85281, USA
diana.petitti@asu.edu

Abstract

High cost for systematic review of biomedical literature has generated interest in decreasing overall workload. This can be done by applying natural language processing techniques to "automate" the classification of publications that are potentially relevant for a given question. Existing solutions need training using a specific supervised machine-learning algorithm and feature-extraction system separately for each systematic review. We propose a system that only uses the input and feedback of human reviewers during the course of review. As the reviewers classify articles, the query is modified using a simple relevance feedback algorithm, and the semantically closest document to the query is presented. An evaluation of our approach was performed using a set of 15 published drug systematic reviews. The number of articles that needed to be reviewed was substantially reduced (ranging from 6%–30% for a 95% recall).

Background

Systematic reviews of biomedical literature are the cornerstone of the development of evidence-based clinical practice guidelines. Systematic reviews are used not only to decide the comparative effectiveness of medical treatments, but also as additional input on decisions about payment for technologies internationally.

The steps to conduct a systematic review (Woolf, 1996; Higgins and Sally Green, 2011; Khan et al, 2001) are:

1. To define the review question and develop criteria for including studies
2. To search for studies addressing the review question
3. To select studies meeting the criteria for inclusion in the review
4. To collect data from the studies meeting the criteria for inclusion
5. To assess the risk of bias in the included studies by appraising them critically
6. Where appropriate, to analyze the data by undertaking meta-analyses
7. To address reporting biases

The results of systematic review are then presented in a report that interprets them and then draws conclusions.

It is nearly impossible to review the full text of all publications identified in step 2 of a well-conducted review. Therefore, step 3 of this process has historically involved human reviewers reading the abstracts of all publications identified in step 2 to determine whether

they meet the criteria for inclusion in the review. Reviewers only study the full text of a relevant publication if the abstract review suggests that the publication might contain data that would address the question posed.

A well-conducted, comprehensive systematic search for all publications related a topic often yields thousands, or even tens of thousands, of citations to publications. It is typical for only a few hundred of the identified publications to be judged as potentially relevant based on the abstract review. It is common for only a handful to ultimately be found to address the question posed (eg, Upadhyay et al, 2011). The abstract review to determine potential relevance is laborious and is known to be costly (ASHP Foundation, 2010; DFID, 2010)

Aphinyanaphongs and colleagues (Aphinyanaphongs et al, 2005) proposed the use of machine learning to reduce the workload in systematic review. Over subsequent years, several other approaches to replace manual (human) review of abstracts as a way to reduce the effort have been proposed. Table 1 describes these systems. All of them employ supervised machine learning, with differences in the machine-learning algorithm employed. Recently, Wallace et al (Wallace et al, 2010) described the application of active learning, a novel extension of supervised machine learning, as an approach to the same problem. Active learning starts with a small training set and interactively obtains a more responsive training set. The output of all these systems, however, is a model that encodes the knowledge learned from few training examples. The model classifies new documents according to whether they are relevant for the systematic review or not. This model might not be useful for a systematic review of other topics. Further, most machine-learning algorithms need parameter tuning; this has to be done manually by computer engineers. In addition to the machine-learning approaches, there are a few approaches that use semantic processing and rules that match question classes for similar tasks (Bray et al, 2008; Fiszman et al, 2010; Fiszman et al, 2008; Lu et al, 2008).

It is in this context that we explore whether semantic information can be automatically derived using distribution of the words in Medline abstracts as a generic strategy for automating the process of identifying potentially relevant publications from abstracts. We also explore the use of an iterative feedback system that eliminates the need for creating a separate training set in an online learning kind of set-up. Such a system could be readily used for any systematic review, even if the reduction in workload is not as high as a supervised machine-learning system.

Methods

Our approach to reducing the workload of systematic review and eliminating the need for the systematic reviewers to interact with informatics professionals separately for each review topic is based on the use of distributional semantics (semantics empirically derived from text) (T. Cohen and Widdows, 2009; Jonnalagadda et al, 2012). Figure 1 depicts the architecture of the system. Abstracts are first uploaded to the system and then a semantic model of the terms is created during the preparation or preprocessing phase. It is also possible to use a previously created semantic model of terms using Medline abstracts to avoid preprocessing. A randomly selected document is presented to the reviewer, who annotates and classifies the document as potentially relevant or not relevant. The semantic model is then used as the basis for presenting the next document to the reviewer based on the similarity of the document to the terms in the document and to the document just classified. This document is annotated and classified as relevant or not relevant. The feedback from the relevance classification by experts and is used to present documents to the reviewer that are increasingly likely to be relevant, based on information from the documents that have been classified as relevant or not relevant to that point. The reviewer

can elect to end the process of classifying documents at any point, recognizing that stopping before reviewing all documents involves a trade-off of lower recall for reduced workload.

A. SEMANTIC SEARCH

In prior text-mining applications to systematic review, the documents are classified dichotomously as relevant (included) or not relevant (excluded). Our approach uses a semantic vector model of the terms present in the abstracts to rank the documents in order of their potential relevance. The ranking of documents is an important feature that distinguishes the approach we describe from prior approaches. The semantic vector model of terms, also referred to as "wordspace," is learned using the sliding windows of the words in the Medline abstracts. All Medline citations of the 2009 baseline (2009 MEDLINE®/PubMed® Baseline Statistics, 2010) that have abstracts (~9 million) are used for creating the term vectors so that the terms are more accurately represented in the wordspace. Using the cosine distances between vector representations of the modified query and the documents, the next most relevant document is calculated iteratively.

In a typical vector representation of terms and documents, each term is considered completely independent. Thus, a search on "diseased" and "sick" might result in a completely different document ranking as measured by the distance between the document vector and the term vector. Recent research (T. Cohen and Widdows, 2009) suggests that the semantic representations using distributional information of the terms, such as Hyperspace Analogue to Language (HAL) (Lund & Burgess, 1996) and Latent Semantic Analysis (LSA) (Landauer and Dumais, 1997), yield better results. We reduce the dimensionality of vectors using random indexing and construct a vector space of terms using the directional model (simplified version of HAL).

B. RANDOM INDEXING

First, we introduce random indexing. Geometric models of distributional semantics represent each term as a vector in high-dimensional space. Distributional semantic models, constructed based on millions of documents and/or millions of terms such as those represented in Medline abstracts, would be unmanageable by size. The models approaching corpora of this magnitude tend to reduce dimensionality first. Traditional, dimensionality reduction techniques, such as Singular Value Decomposition (SVD), are computationally expensive (the commonly utilized algorithm for SVD is cubic in complexity) (Trefethen and Bau, 1997). Recently, Random Indexing (Kanerva et al, 2000) emerged as a promising alternative to the use of SVD for the dimension reduction step in the generation of term-by-context vectors. Random Indexing and other similar methods are motivated by the Johnson–Lindenstrauss Lemma (Johnson and Lindenstrauss, 1984) that states that the distance between points in a vector space will be approximately preserved if they are projected into a reduced-dimensional subspace of sufficient dimensionality. Random Indexing scales at a rate that is linear to the size of the data, since the term-document or term-term matrix need not be stored in memory. This is accomplished by assigning to each document (in term-document models) or term (in sliding-window models) a sparse high-dimensional (on the order of 1000) elemental vector, a vector comprising of mostly zero elements with a small number (on the order of 10) set to either +1 or −1. These non-zero elements are determined at random, and because of the sparseness of the vectors, the resulting vectors are highly likely to be orthogonal or close-to-orthogonal to one another.

We use the Semantic Vectors package (Widdows and Cohen, 2010) to create elemental vectors for the terms in Medline abstracts using random indexing. Based on our previous experiments (Jonnalagadda, 2011), which revealed that using 2000-dimensional vectors and 5 seeds (number of +1s and −1s in the vector) are most optimal, we create the elemental

vectors for all the terms in each of the 9 million Medline abstracts. Among different types of distributional models implemented in the Semantic Vectors package (Widdows and Cohen, 2010) (Basic, Positional, Directional, and Positional + Basic), the Directional model was shown to optimally assign similar vectors (in direction) to terms appearing in similar context (Jonnalagadda, 2011).

C. DIRECTIONAL MODEL

The algorithm uses a sliding window that is moved through the text corpus to generate a term-term matrix, T, where T[i, j] is the number of times the word representing the jth column appears near the word representing the ith column. Two words are in the vicinity of each other if, and only if, the number of words separating them is less than an integer parameter known as the sliding-window radius. The directional model also takes into account the direction in which a word occurs with respect to another by having two columns for each word, with one column representing the number of occurrences to the left and the other column representing the number of occurrences to the right.

D. DOCUMENT RANKING

The documents and the query were mapped to the vector space as follows:

$$s(C) = \left\| \sum_{i=1}^{T} n[i] * s(t[i]) \right\|, \quad \text{Equation 1}$$

where C is a collection of terms such as a document or a term, s() is the unit semantic vector, t[i] is the i^{th} term, n[i] is the number of times t[i] occurs in C, and $\| \ \|$ is the norm operator.

The cosines of the document vectors, with respect to the query vector, are measured. The most relevant document d for the query q among a set of documents D is given by:

$$d = \{d \in D | (\forall x \in D, \cos(s(x), s(q)) \leq \cos(s(d), s(q))\} \quad \text{Equation 2}$$

E. RELEVANCE FEEDBACK

Our approach uses feedback from the reviewers (as shown in Figure 1) as they review the abstracts presented to them by the system to modify. This is “relevance feedback.” The incorporation of feedback is the second feature that distinguishes our approach from those described previously, although Wallace et al's (Wallace et al, 2010) prototype system also uses relevance feedback through active learning to create a training set.

The system first asks the reviewer to describe the study using salient words or a simple query. The initial vector was constructed by the initial query terms set (Q_0) given by the reviewer. If the reviewer decides not to give an initial query, Q_0 is an empty set.

The set of query terms after m documents (Q_m) were reviewed is given by:

$$Q_m = (Q_0 \cup P_m) - N_m \quad \text{Equation 3}$$

where N_m are the terms that appear in the documents reviewed as not relevant so far, P_m are the terms that appear in the documents reviewed as relevant so far, and Q_m is used to create the query vector after reviewing m documents by adding the vectors for each term in the set.

As evident from Equation 3, the query at any stage depends on the terms in the documents reviewed so far. The most relevant document to a query in the remaining documents is

selected using Equation 2. After sufficient experience with a variety of topics, a cut-off criteria could be decided by the users (or suggested by the system) based on the percentage of articles reviewed, the cosine similarity value of the document presented as most appropriate, and the number of successive irrelevant articles presented by the system.

F. DESCRIPTION OF THE SYSTEMATIC REVIEWS USED FOR THE EVALUATION

Our system needs no training set data for development. However, we need a set of annotated documents so that the annotation could be used to simulate the reviewer user of the system. We evaluated our system using information from 15 systematic reviews of drug classes that Cohen et al (A.M. Cohen et al, 2006) used to train their supervised machine-learning system. The different categories provided by Cohen et al are merged to create a binary classification of relevant or not relevant documents. Although the abstracts are not representative, given the high percentages of inclusion, this is the only publicly available collection. For these 15 systematic reviews, Cohen et al have made available (http://medir.ohsu.edu/~cohenaa/systematic-drug-class-review-data.html) the PubMed IDs of the abstracts that they manually reviewed, along with the corresponding classification of each abstract. Table 2 presents the 15 drug review topics along with the number of abstracts reviewed, the number of abstracts included, and the number of abstracts excluded by Cohen et al in their review.

In systematic reviews done with the aim of developing practice guidelines and clinical policy, all abstracts are first studied to identify and eliminate the clearly irrelevant ones from further review. The full texts of the possibly eligible documents are then reviewed to identify those that are actually relevant. In our evaluation, we use only abstracts, not fulltext, to make our final determination of relevance, because some full texts were not freely available. The work saved over sampling at X%, or WSS@X%, also defined by Cohen et al (A.M. Cohen et al, 2006), will be used for evaluation:

$$WSS@X\% = \frac{TN+FN}{N} - 1 + \frac{X}{100}, \quad \text{Equation 4}$$

where TN is the number of true negatives identified by the system, FN is the number of false negatives identified by the system, N is the total number of documents in the test set, and X is the recall rate.

Results

We assessed the performance of our relevance feedback-based system in terms of reduction in workload based on Cohen's 15 drug class reviews. Since the key questions these reviews try to address are unknown and the objective is to test the system, the initial queries are not set, although highly precise queries would result in better performance. The query vector becomes more and more relevant to the task as we use relevance feedback. Figure 2 shows how recall changes as more articles are reviewed. Considering both workload and recall, an ideal system for selection of abstracts for human review of the full text of articles would have 100% recall when the number of abstracts presented to the expert reviewer by the systems is exactly equal to the number classified as potentially relevant in the “gold standard” (manually reviewed) abstracts.

Figure 2 shows how recall changes as more articles are reviewed. Recall at a given proportion of articles reviewed varies substantially by topic. At an arbitrary threshold of a recall of 90%, the percentage of all abstracts reviewed is as low as 43% (attention deficit hyperactivity disorder) and as high as 95% (opiods). To assure attainment of 95% recall across all 15 topics, as might be required when using the system as a generic approach to

selection of abstracts for human review, it would be necessary to review 95% of all abstracts. If relevance feedback was not useful, the curves would have been straight lines, reflecting presentation of articles in random order. Figure 2 shows that relevance feedback leads to a large initial increase in recall. As more and more information is added, the increase in recall becomes more gradual.

Figure 3 shows the percentages of work saved in review over sampling at 95% recall comparing our system with results presented by Cohen et al for the supervised machine-learning system. Cohen et al (A.M. Cohen et al, 2006) used a 5 × 2 cross validation (half of each corpus for training and the other half for testing) on a supervised machine-learning system. Our results are based on the entire collection of documents for each review. Considering workload reduction, our results are broadly comparable to those of Cohen et al (as shown in Figure 3). Estimated workload reductions at 95% recall range from 6% to 30% for our system and from 0% to 68% for Cohen's workload; median estimated workload reduction at 95% recall is 13% for our system and 18% for Cohen's system. Using our system, we estimate a reduction in workload for all 15 reviews, whereas Cohen et al's system suggested a reduction in workload for 13 of the 15 reviews. Our system had a better performance than Cohen et al's did for 5 reviews.

Our findings provide strong support for the conduct of further research to create unsupervised systems to reduce workload in a systematic review process. Unlike supervised systems, they do not add additional workload of creating a training set or of building a trained model that might involve interacting with computer engineers.

Discussion

We described a system that reduces the workload of systematic review based on the use of semantic features of the document to identify potentially relevant documents. We have coupled this with feedback about relevance to the system based on classification by experts that results in documents more likely to be relevant when presented to the expert earlier.

Semantic features in the form of manually assigned MeSH terms have been previously used by Cohen in a similar attempt to reduce the workload of systematic review (A.M. Cohen, 2008). Our system is different from Cohen's in that the semantic features are created automatically. The system uses a directional model for creating the semantic vectors, which are created for terms that are paradigmatically related. If two terms can be substituted for each other in a sentence (ie, they occur in similar local contexts throughout the corpus), they are said to be in a paradigmatic relationship. Examples of terms in a paradigmatic relationship are p53 (gene) and gata1 (gene); AD and SDAT (synonyms); and poliomyelitis and polio (synonyms). The directional model approach enables semantic search, where the user need not enter all the synonyms for a particular concept to get all the relevant documents.

Using a traditional supervised machine-learning approach, it is possible that a document will not be classified as relevant because it uses different words or n-grams to convey a concept. Since documents are represented as a complex vector or logical combination of various features, in traditional approaches, it is not easy for users to modify the criteria for document selection. In the proposed system, the distributional semantic model assigns nearby vectors to contextually similar words. Therefore, even if an important (key) word is paraphrased or replaced by a similar word in an unannotated document, it is likely to be ranked high. In addition, the dynamically changing query set, which decides which document will be presented next, could be easily modified at any stage by removing or adding terms to the query set. In this way, the possibility of not finding documents that use different words

could be further minimized by allowing active participation of the users in defining the terms in the query set.

In traditional supervised machine learning, externally supplied instances are used to create a model that classifies future instances (Kotsiantis, 2007). In our framework, no instances are supplied initially to the system, and predictions are made using the classification of test set instances. Our approach has similarities with active learning, but it is designed to be easier to adapt. The application of active learning to assist systematic review is also novel (Wallace et al, 2010). Future work might involve comparing the performance of these two methods and perhaps integrating them. A second application of our system would be to obtain a balanced training set for a traditional supervised system. The first 1000 (or so) documents reviewed using our system would have higher number of relevant documents than the same number of documents selected at random.

Across the 15 topics we examined, our system was not able to assure a high rate of recall (90%–95%) with a substantial reduction (40%) in workload reliably. The acceptability of both our system and other systems that attempt to substitute modern informatics approaches for human labor is not yet known. If those who rely on systematic review to develop guidelines and policy demand 100% recall and informatics approaches such as ours are not able to guarantee 100% recall, the approaches may be doomed. Applications of the approaches to assure more frequent updating of systematic reviews might be more acceptable than use for de novo review. However, our system provides a framework that only complements the manual review process by passively using feedback from the reviewer to determine the order of review. Further empiric work with policymakers should accompany the development of approaches like ours.

Conclusion

We proposed the use of distributional semantics and user feedback as an approach to reduce workload in systematic review. The system might be immediately useful as an enhancement to existing traditional supervised learning systems by creating balanced training sets. Even though the system currently does not use sophisticated features such as n-grams, MeSH terms, and UMLS identifiers and does not have a separate training phase, its performance is comparable to a well-known existing system that also attempts to reduce workload of systematic review (albeit by applying supervised machine learning). Future work would involve integrating the system with an active learning system and incorporating the above features.

References

2009 MEDLINE®/PubMed® Baseline Statistics. Apr 19. 2010 Retrieved from http://www.nlm.nih.gov/archive//20100419/bsd/licensee/2009_stats/2009_LO.html

Aphinyanaphongs Y, Statnikov A, Aliferis CF. A comparison of citation metrics to machine learning filters for the identification of high quality MEDLINE documents. Journal of the American Medical Informatics Association: JAMIA. 2006; 13(4):446–455. doi:10.1197/jamia.M2031. [PubMed: 16622165]

Aphinyanaphongs Y, Tsamardinos I, Statnikov A, Hardin D, Aliferis CF. Text categorization models for high-quality article retrieval in internal medicine. Journal of the American Medical Informatics Association: JAMIA. 2005; 12(2):207–216. doi:10.1197/jamia.M1641. [PubMed: 15561789]

Aronson AR. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. AMIA. 2001

ASHP Foundation. Demonstrating Pharmacists' Value: A Systematic Evidence Review Request for Proposals. Demonstrating Pharmacists' Value: A Systematic Evidence Review Request for Proposals. Feb. 2010 Retrieved May 17, 2011, from http://www.ashpfoundation.org/

MainMenuCategories/ResearchResourceCenter/FundingOpportunities/DemonstratingPharmacistsValueASystematicEvidenceReviewRequestforPro posals.aspx

Bekhuis T, Demner-Fushman D. Towards automating the initial screening phase of a systematic review. Studies in Health Technology and Informatics. 2010; 160(Pt 1):146–150. [PubMed: 20841667]

Bray, BE.; Fiszman, M.; Shin, D.; Rindflesch, TC. Using semantic predications to characterize the clinical cardiovascular literature. AMIA … Annual Symposium Proceedings / AMIA Symposium. AMIA Symposium; 2008. p. 887

Cohen AM, Hersh WR, Peterson K, Yen P-Y. Reducing workload in systematic review preparation using automated citation classification. Journal of the American Medical Informatics Association: JAMIA. 2006; 13(2):206–219. doi:10.1197/jamia.M1929. [PubMed: 16357352]

Cohen, Aaron M. Optimizing feature representation for automated systematic review work prioritization. AMIA … Annual Symposium Proceedings / AMIA Symposium. AMIA Symposium; 2008. p. 121-125.

Cohen, Aaron M.; Ambert, K.; McDonagh, M. Cross-topic learning for work prioritization in systematic review creation and update. Journal of the American Medical Informatics Association: JAMIA. 2009; 16(5):690–704. doi:10.1197/jamia.M3162. [PubMed: 19567792]

Cohen T, Widdows D. Empirical distributional semantics: Methods and biomedical applications. Journal of Biomedical Informatics. 2009; 42(2):390–405. [PubMed: 19232399]

DFID. DFID Systematic Review Pilot. DFID Systematic Review Pilot. Feb. 2010 Retrieved May 17, 2011, from http://www.dfid.gov.uk/r4d/PDF/Publications/DFID_Systematic_Review_FAQ.pdf

Fiszman M, Bray BE, Shin D, Kilicoglu H, Bennett GC, Bodenreider O, Rindflesch TC. Combining relevance assignment with quality of the evidence to support guideline development. Studies in Health Technology and Informatics. 2010; 160(Pt 1):709–713. [PubMed: 20841778]

Fiszman, M.; Ortiz, E.; Bray, BE.; Rindflesch, TC. Semantic processing to support clinical guideline development. AMIA … Annual Symposium Proceedings / AMIA Symposium. AMIA Symposium; 2008. p. 187-191.

Frunza O, Inkpen D, Matwin S, Klement W, O'Blenis P. Exploiting the systematic review protocol for classification of medical abstracts. Artificial Intelligence in Medicine. 2011; 51(1):17–25. doi: 10.1016/j.artmed.2010.10.005. [PubMed: 21084178]

Higgins, JPT.; Sally Green, P. Cochrane handbook for systematic reviews of interventions. Vol. Vol. 4. John Wiley & Sons; 2011.

Johnson WB, Lindenstrauss J. Extensions of Lipschitz mappings into a Hilbert space. Contemporary mathematics. 1984; 26(189–206):1–1.1.

Jonnalagadda, S. An effective approach to biomedical information extraction with limited training data (PhD Dissertation, Arizona State University) (PhD). Arizona State University; Phoenix: 2011.

Jonnalagadda S, Cohen T, Wu S, Gonzalez G. Enhancing clinical concept extraction with distributional semantics. Journal of Biomedical Informatics. 2012; 45(1):129–140. doi:10.1016/j.jbi.2011.10.007. [PubMed: 22085698]

Kanerva, P.; Kristofersson, J.; Holst, A. Random indexing of text samples for latent semantic analysis. Proceedings of the 22nd Annual Conference of the Cognitive Science Society; Citeseer. 2000.

Khan, KS.; ter Riet, G.; Glanville, J.; Sowden, AJ.; Kleijnen, J., et al. Undertaking systematic reviews of research on effectiveness: CRD's guidance for carrying out or commissioning reviews. NHS Centre for Reviews and Dissemination; 2001.

Kotsiantis, SB. Emerging artificial intelligence applications in computer engineering : real word AI systems with applications in eHealth, HCI, information retrieval and pervasive technologies. IOS Press; Amsterdam: 2007. Supervised Machine Learning: A Review of Classification Techniques.

Landauer TK, Dumais ST. A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. Psychological review. 1997; 104(2):211–240.

Lu, Z.; Kim, W.; Wilbur, WJ. Evaluating relevance ranking strategies for MEDLINE retrieval. AMIA … Annual Symposium Proceedings / AMIA Symposium. AMIA Symposium; 2008. p. 439

Lund K, Burgess C. Hyperspace analog to language (HAL): A general model of semantic representation. Language and Cognitive Processes. 1996

Matwin S, Kouznetsov A, Inkpen D, Frunza O, O'Blenis P. A new algorithm for reducing the workload of experts in performing systematic reviews. Journal of the American Medical Informatics Association: JAMIA. 2010; 17(4):446–453. doi:10.1136/jamia.2010.004325. [PubMed: 20595313]

Trefethen, LN.; Bau, D. Numerical linear algebra. Society for Industrial Mathematics; 1997.

Upadhyay A, Earley A, Haynes SM, Uhlig K. Systematic review: blood pressure target in chronic kidney disease and proteinuria as an effect modifier. Annals of Internal Medicine. 2011; 154(8): 541–548. doi:10.1059/0003-4819-154-8-201104190-00335. [PubMed: 21403055]

Wallace BC, Trikalinos TA, Lau J, Brodley C, Schmid CH. Semi-automated screening of biomedical citations for systematic reviews. BMC Bioinformatics. 2010; 11:55. doi: 10.1186/1471-2105-11-55. [PubMed: 20102628]

Widdows, D.; Cohen, T. The Semantic Vectors Package: New Algorithms and Public Tools for Distributional Semantics. Fourth IEEE International Conference on Semantic Computing; 2010. p. 43

Woolf S. Manual for conducting systematic reviews. Agency for Health Care Policy and Research. 1996

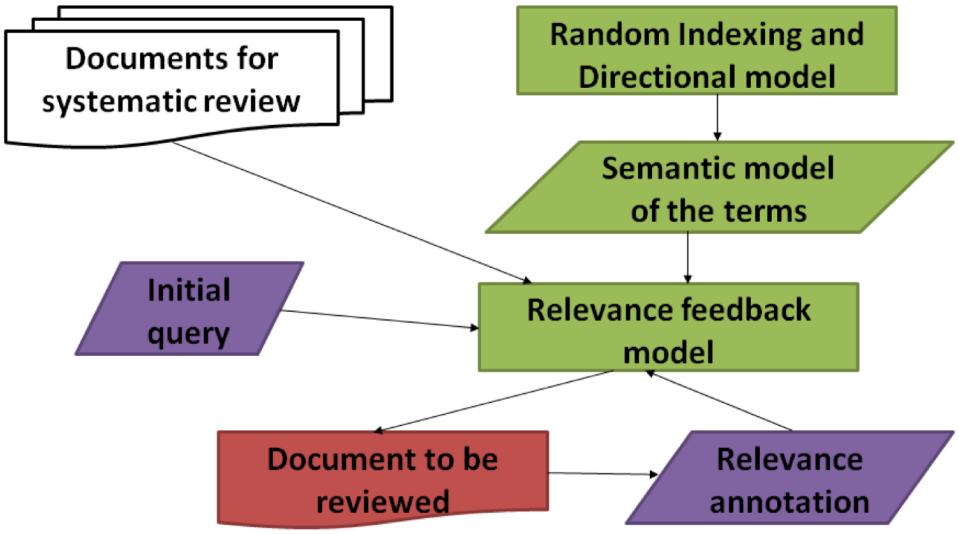

Figure 1. Architecture of the system

A semantic model is built for the terms present in the documents that need to be systematically reviewed. Using the initial query and our relevance feedback algorithm that uses the expert review for the documents annotated for relevance so far, the next document that is most likely to be eligible is presented.

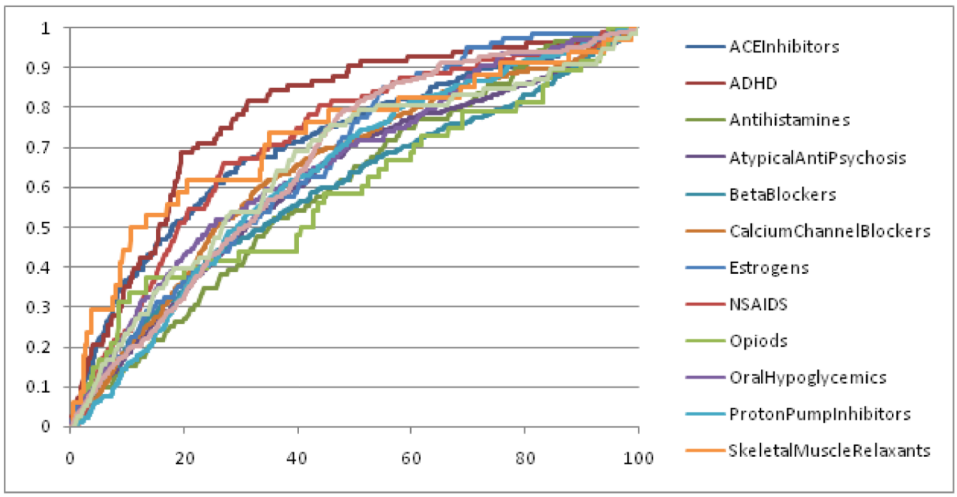

Figure 2. Performance of the system on 15 drug systematic reviews

The X-axis represents the percentage of abstracts reviewed and the Y-axis represents the recall. Plots are shown for all the 15 drug reviews.

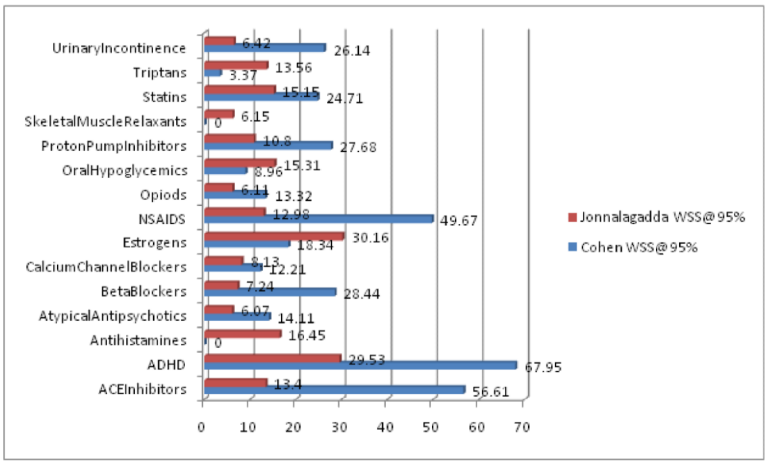

Figure 3. Comparison of our system with Cohen et al, 2006's

Jonnalagadda WSS@95% is the label for the respective percentage of work saved over a sampling of 95% recall for the current system. Cohen WSS@95% is the label for the respective percentage of work saved over a sampling of 95% recall for Cohen et al.'s system.

Table 1

Summary of Methods Proposed for Aiding in Systematic Review

Ref #	Title	Year	Machine-Learning Algorithm(s) Used	Comments
1	Text categorization models for high-quality article retrieval in internal medicine	2005	Naïve Bayes, Adaboost, SVM	First known method
2	A comparison of citation metrics to machine-learning filters for the identification of high-quality MEDLINE documents	2006	Support Vector Machines (SVM)	
3	Reducing workload in systematic review preparation using automated citation classification	2006	Perceptron based voting	
4	Optimizing feature representation for automated systematic review work prioritization	2008	SVM	Extensive research on Machine-learning features
5	Cross-topic learning for work prioritization in systematic review creation and update	2009	SVM	
6	A new algorithm for reducing the workload of experts in performing systematic reviews	2010	Factorized version of Complement Naïve Bayes (FCNB)	
7	Semi-automated screening of biomedical citations for systematic reviews	2010	ensemble of SVMs	Uses active learning
8	Toward automating the initial screening phase of a systematic review	2010	Evolutionary SVM	
9	Exploiting the systematic review protocol for classification of medical abstracts	2011	FCNB	

Ref #: the citation in the References section (1-,9 respectively, correspond to Aphinyanaphongs et al, 2006; Aphinyanaphongs et al, 2005; Bekhuis and Demner-Fushman, 2010; A.M. Cohen et al, 2006; A.M. Cohen, 2008; A.M. Cohen et al, 2009; Frunza et al, 2011; Matwin et al, 2010; Wallace et al, 2010); *Title: title of the paper; Year: the year in which the article is published.*

Table 2

Drug Class Reviews Used for Validation of the Method

Drug Class	UMLS	Total Abstracts	Included Abstracts	Excluded Abstracts
ACE inhibitors	C0003015	2544	183	2361
Attention deficit hyperactivity disorder	**C1263846**	851	84	767
Antihistamines	C0019590	310	92	218
Atypical antipsychotics	C0040615	1120	363	757
Beta blockers	C0001645	2072	302	1770
Calcium channel blockers	C0006684	1218	279	939
Estrogens	C0202006	368	80	288
Non-steroidal antiflammatory drugs (NSAIDS)	C0003211	393	88	305
Opioids	**C0029104**	1915	48	1867
Oral hypoglycemics	**C0571635**	503	139	364
Proton pump inhibitors	C0358591	1333	238	1095
Skeletal muscle relaxants	C0037250	1643	34	1609
Statins	C0360704	3465	173	3292
Triptans	C1567966	671	218	453
Urinary incontinence	**C0042024**	327	78	249

Each row in the above table corresponds to a class of drugs. Certain rows that are named after a medical condition, such as “Attention deficit hyperactivity disorder,” correspond to the class of drugs that treat the condition. For more detail about the drug class, UMLS (Aronson, 2001) codes are assigned (**bold for medical conditions** and normal for actual drug classes).