Issues in Information Systems
Volume 18, Issue 1, pp. 22-30, 2017

AUTOMATICALLY CLASSIFYING PUBMED ABSTRACTS AS BENCH OR BEDSIDE

Daniel McDonald, Utah Valley University, dmcdonald@uvu.edu
Michelle Ashton, Utah Education and Telehealth Network (UETN), mashton@uen.org

ABSTRACT

The PubMed database contains over 20 million research abstracts ranging from lab experiments and gene arrays to patient-facing research. We introduce a classification task that groups PubMed abstracts into categories of basic science and clinical research. We present a conditional probability and a decision tree algorithm and compare the two algorithms based on three different feature sets. The first feature set consists of semantic tags that appear as verbs in the abstract. The second feature set consists of tags that are nouns and appear as subjects or objects within a sentence. The third feature set consists of the first two feature sets combined. Algorithms are evaluated using precision, recall and f-measure measurements. The decision tree algorithm with features made up of both verb tags and tags from subjects and objects outperformed all other combinations achieving a precision of 97 percent and a recall of 96.8 percent. The lack of fallback rules when using the conditional probability algorithm hurt its performance. The decision tree algorithm was more robust to testing abstracts of different lengths and unseen feature values.

Keywords: Information retrieval, document classification, text mining

INTRODUCTION

In 2003 the National Institutes of Health (NIH) laid out a Roadmap for Medical Research that emphasized using basic research to improve health. In theory, the knowledge and advancements developed at the bench or in the lab should pass smoothly and quickly along to the next step of the research process. The more efficient the translation process, the faster patients receive advanced treatment. However, gaps in the process between published basic science research and further testing and validation slow the translation process. Automated tools that help users find and process relevant documents along the process may provide some assistance.

With now over 26 million published citations in PubMed and an estimated growth rate of four percent per year, it is increasingly difficult for researchers to filter through documents and build on the research of others (Lu, 2011). PubMed citations are a critical communication artifact between researchers. The difficulty of building on basic science research to perform clinical research has been cited as a barrier to Translational Research (TR) (Schnapp et al., 2009; Van der Laan & Boenink, 2015).

In this research, we introduce a document classification task that involves automatically grouping PubMed abstracts into basic research and clinical research. The relevance of documents to information-seeking users is not only influenced by the search task, but also by research that matches their stage in the translation process (McDonald & Chen, 2006). In this paper, we evaluate two algorithms with varying features based on their performance correctly classifying PubMed abstracts. We will first explore the text classification literature, especially as it relates to biomedical text. Next, we will introduce our algorithm and then present a research design to evaluate it. Finally, we will report on the results of the experiment, make conclusions regarding our classification approach, and then share our future research plans.

LITERATURE REVIEW

Research that affects the classification of PubMed abstracts has over a 25-year history. Over the years, there has been a combination of natural language processing (NLP) enhancements, machine learning enhancements and valuable resources created by the National Library of Medicine (NLM) that have contributed to improvements.

Research Overview

Led by DARPA and closely coordinated with the National Institute of Standards and Technology (NIST) in 1991, the U.S. government created the TIPSTER text project (which included TREC) to advance the state of the art in text processing. TREC currently has tracks related to biomedical informatics including a genomics track and a clinical decision support track.

A text processing event specific to the biomedical text domain was organized in 2004 titled the Critical Assessment of Information Extraction in Biology (BioCreative) (Hirschman, Yeh, Blaschke, & Valencia, 2005). An early BioCreative task was to identify genes and proteins from text and map them to standard identifiers in the fly, mouse, or yeast databases. An additional task included identifying concepts in full-text articles that supported Gene Ontology (GO) annotations. Current tasks in BioCreative include automatically annotating biomedical text, integrating ontologies and text, detecting drugs and compounds in text, and automatically extracting experimental methods from publications.

There has also been substantial text processing research in the biomedical domain outside of BioCreative and TREC. There has been systems to extract gene, protein, enzyme and molecular interactions and pathways from text (Carol Friedman, Kra, Yu, Krauthammer, & Rzhetsky, 2001; Humphreys, Demetriou, & Gaizauskas, 2000; McDonald, Chen, Su, & Marshall, 2004; Novichkova, Egorov, & Daraselia, 2003; Ono, Hishigaki, Tanigami, & Takagi, 2001; Zhou & He, 2008). There has been algorithms to automatically assign MeSH terms or other annotations to PubMed abstracts (Chen, Müller, & Sternberg, 2006; Ibushi, Collier, & Tsujii, 1999). There has also been systems for clinical radiology data management and systems to index findings and diagnoses (C. Friedman, Alderson, Austin, Cimino, & Johnson, 1994; Hersh, Mailhot, Arnott-Smith, & Lowe, 2001). There continues to be a great potential for text analysis applications in systems biology (Ananiadou, Kell, & Tsujii, 2006).

In order to make sense of the text processing research, we organize the literature around three main knowledge acquisition tasks, namely the find, process and analyze and visualize tasks. The three main tasks are shown in Figure 1. First, users must find documents relevant to their current research. These tasks fall in the area of Information Retrieval. Algorithms that enhance document finding in the biomedical domain includes algorithms such as query expansion, indexing of findings and diagnoses for search and document classification (Hersh et al., 2001; Imambi & Sudha, 2011; Pentoney, Harwell, & Leroy, 2014; Poulter, Rubin, Altman, & Seoighe, 2008). Second, once relevant documents have been identified, they can be processed using information extraction techniques. In this stage, entities such as genes, proteins, and enzymes are identified along with other items of interest such as clinical evidence. Entity relationships, such as protein-protein interactions or molecular pathways would also be extracted here. Third is the task of analyzing and visualizing relationships between information. This third area is often referred to as text mining. Different from finding and processing, the analyze step typically involves combining textual components from many documents and may integrate knowledge from ontologies and other data sources to produce the analysis. Key in this step is that relationships between data is aggregated and analyzed. Notice from the diagram that the tasks do not necessarily have an order. The three tasks do not own particular algorithms either. Our research is in find stage of the diagram, using the task of document classification.

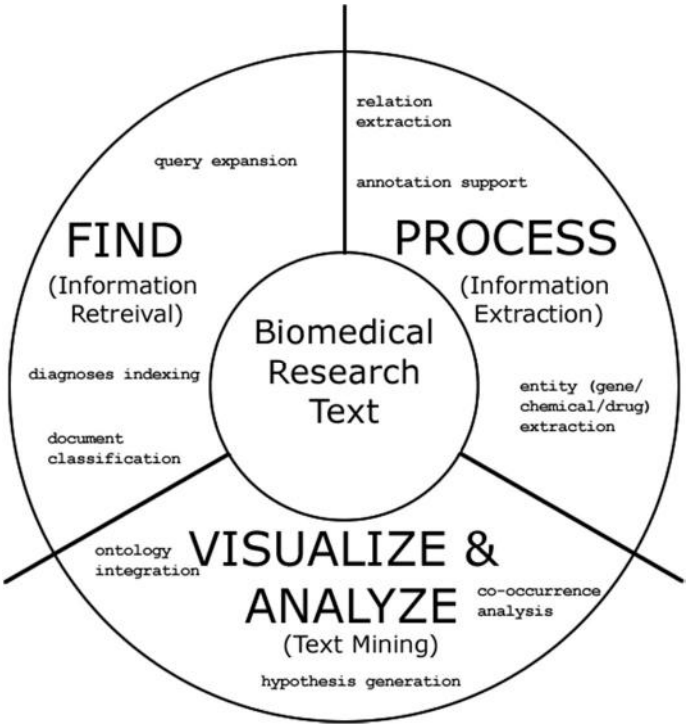

Figure 1. Major tasks addressed by automated text processing

Document classification

Using PubMed as a corpus, document classification has been used to enhance document relevance measures, group together research on related topics, as well as to identify publication types (Chen et al., 2006; Houston et al., 2000; Imambi & Sudha, 2011; Kilicoglu, Demner-Fushman, Rindflesch, Wilczynski, & Haynes, 2009; Poulter et al., 2008; Sarker & Molla-Aliod, 2010). The task of identifying publication types has the most overlap with our research. Sarker and Molla-Aliod used regular expression rules to separate medical literature into three different article types: Systematic Reviews, Meta-analyses and Randomized Controlled Trials (Sarker & Molla-Aliod, 2010). The task was considered valid because the medical community considered these publication types to be high quality. Other research focused solely on identifying systematic review articles. Hunt and McKibbon identified some key phrases as good indicators of systematic reviews (Hunt & McKibbon, 1997). Montori et al. used a set of terms from abstracts, titles, and meta-data to identify systematic reviews (Montori, Wilczynski, Morgan, & Haynes, 2005). Other approaches relied more on meta-data and search filters. Shonjania and Bero evaluated different search strategies using less visible search interfaces to search PubMed (Shojania & Bero, 2001). Related to labeling types of publications, other research identified publications with particular methodological attributes. Haynes et al. used combinations of query terms to detect studies meeting basic methodological standards (Haynes, Wilczynski, McKibbon, Walker, & Sinclair, 1994). Kilicoglu et al. presented an algorithm that identified scientifically rigorous clinical research studies (Kilicoglu et al., 2009). In their approach, they use an ensemble of supervised machine learning algorithms trained on high-level semantic features. Their advanced approach, tested on 2,000 citations, achieved nearly 74 percent precision and 62 percent recall. In the research classifying PubMed abstracts into publication types, performance numbers were typically high while classification rules were typically manually created. Kilicoglu et al. stands out for its advanced algorithmic approach and unique classification groupings.

RESEARCH GAP

Document classification has commonly been used on PubMed abstracts. However, classification has been used to identify types of research papers, such as systematic review articles or articles with a certain level of rigor. Instead of focusing on labeling high quality research, our task is to identify publications at different stages of the translational process, namely basic science and clinical research. There are many PubMed abstracts that can be classified into our two categories, especially compared to previous tasks, such as finding review articles. In addition to having a unique task, we are also using novel inputs into the classification algorithm. We do not use existing abstract text or meta data as input to our algorithm. Rather, we tag the abstract text with lexical-semantic tags that are used as input to the classification algorithm. We expect this approach will generalize better than using text directly from the abstracts.

RESEARCH QUESTION

Different than identifying systematic reviews or rigorous clinical studies, our algorithm classifies research into basic science or clinical research. We explore just what type of performance can be expected for such a task. Secondly, we compare two supervised machine learning algorithms each with inputs based on lexical-semantic tags. We explore how well the different algorithms generalize to the testing data set without over training or "memorizing" inputs. We also explore the types of lexical-semantic tags that are most informative in the classification task, i.e. NOUN groups or VERB groups.

Our primary research questions include the following:

1. How high of precision and recall can be achieved given the two classification groups of basic research and clinical research?
2. Would a conditional probability approach perform better than a decision tree approach?
3. Does using verbs as features in the algorithms improve classification performance?
4. How does including both verbs and nouns together as inputs impact the classification performance?

EXPERIMENTAL DESIGN

The PubMed abstracts for our training and testing sets were taken from PubMed searches. In order to obtain abstracts that fit into the bench science and clinical research categories, we used different keywords to obtain different result sets. We first searched using the keywords "p53" which is a tumor suppressor. We extracted the first 100 abstracts that were returned. We later searched for "patients" and extracted the first 100 abstracts that were returned. We had an upper-class biology student go through the 200 abstracts and identify them as most related to basic science or clinical research. The result was that eight-five abstracts were identified as basic science, sixty-nine abstracts were tagged as clinical research, twenty-nine abstracts were identified as systematic review articles, twelve abstracts were tagged as both basic science and clinical, and five abstracts were tagged as research on tools and techniques. Because of our focus on differentiating between basic science and clinical research, we kept only the 154 abstracts from those two categories. We randomly selected 20 percent of the abstracts (31 abstracts) to be the test set, which left 123 abstract in the training set. The training set was left with 67 basic science abstracts and 56 clinical abstracts (54 percent basic science). The testing set had 18 basic science abstracts and 13 clinical abstracts (58 percent basic science).

To generate the input for the machine learning algorithms, we processed all the abstracts using our content tagging algorithm (McDonald et al., 2004). The algorithm tokenizes the document and recognizes sentence boundaries. The tokenizing process separates hyphenated words, adds spaces to punctuation, recognizes abbreviations, and matches parenthesis and quotations. Once tokenized, an abstract's words and phrases are tagged using hybrid semantic/syntax tags. The tagging process is aided by the use of an extensive dictionary of approximately one million word-tag entries. After tagging, the text is processed several times more to combine tags into topic categories. The topic categories are part of a large category hierarchy. Nouns and verbs are placed into different categories. There are just over 3,400 different categories in the hierarchy into which terms (words or phrases) can be assigned. This tagging is considered a placement of each phrase into a taxonomy. In addition, a grammatical parse took place that assigned tags into subject, verb or object roles where applicable.

Three different feature sets were created from the tagging of the title and abstract text. First, tags were used as features if the phrases appeared in a grammatical triple as a verb phrase. Second, tags were extracted from the document and used as features if the tags appeared in a grammatical triple as a subject or object. Finally, tags were included as features if they appeared in a grammatical triple as the subject verb or object.

These three feature sets were used by two different algorithms and performance was measured using precision, recall, and f-measure in their ability to classify documents. The first algorithm utilized the conditional probability (CP) of tags occurring with the different categories of abstracts. The equation for the algorithm is shown in Figure 2, where *O* is the classification outcome, either basic science or clinical research, T is a tag from the testing document that matches one from the training set, and *n* is the total number of matching tags between the testing abstract and the training set of abstracts. To train the algorithm, the training abstracts were processed three times to obtain three different sets of tags. Each tag from the training set was given a tag probability $P(T_n)$ indicating how common that tag was in the training set. Next, for each tag, a probability was calculated for the abstract being a basic science $P(O_{basic\ science}|T_n)$ abstract or a clinical abstract $P(O_{clinical}|T_n)$. For example, the tag AUTOPHAGY occurred 16 times in the testing set of abstracts. Such a tag had a tag probability of .006. Furthermore, every time the tag occurred, it occurred in a basic science abstract. So if this tag appeared in a testing document, it would contribute .006 to the basic science outcome and nothing to the clinical outcome. To process the testing set, all the tags from the testing documents were compared to the conditional probability scores from the training set. The outcome with the highest probability determined the abstract classification.

$$\sum_n P(O_{basic\ science}|T_n) \times P(T_n) \quad compared\ to \quad \sum_n P(O_{clinical}|T_n) \times P(T_n)$$

Figure 2. calculation for determining the probability of being basic science or clinical research

Historically, language modeling approaches have performed well when testing sets match well to training sets, but performance falters when values are encountered that do not appear in the training set. Currently, we just ignore unrecognized tags with no fall back algorithm in place.

In order to address the expected lack of generalization of our probabilistic algorithm, we decided to compare it to the performance of a decision tree. We used the open-source machine learning software Weka to run the experiment.(Hall et al., 2009) With 123 training abstracts, we had a relatively small training set. To avoid a large feature space, we decided not to use individual tags as inputs, but rather use numeric inputs. We decided to separate the tags into six groups, which would be used as inputs. We grouped all the tags that had 85-100 percent probability of occurring in a clinical abstract together. We labeled this group, CLINICALTOP. Next, we grouped tags with 70-85 percent probability of occurring in a clinical abstract together. That group was labeled CLINICALMED. Finally, tags with 55-70 percent probability of being in a clinical abstract were grouped into the label CLINICALLOW. We did the same thing for the basic science group. In total, we had six inputs as shown in Table 1. Once the groups had been established and the tags we assigned to different groups, we were able to create the training data. For each abstract, a count was tallied for each of the six groups. Each time a tag occurred that was part of one of the groups, that group's total would be incremented. Table 1 shows an example of one of the training records.

Table 1. An input for one of the basic science abstracts

CLINICALTOP	CLINICALMED	CLINICALLOW	BASICSCITOP	BASICSCIMED	BASICSCILOW
5	3	2	19	20	11

In Weka, there is an implementation of the decision tree algorithm J48, which we used as our classifier.
Our experiment is a two by three design, with the two algorithms being compared given three different sets of features each. We have formed for the following four hypotheses for our experiment:
H_1: Precision and recall of over 90 percent can be achieved in the given classification task.
H_2: A conditional probability approach will outperform the decision tree only with all the inputs NOUN + VERB tags included as input.

H_3: Using VERB tags alone will NOT outperform noun tags alone.
H_4: Using NOUN + VERB tags as inputs will outperform other feature combinations.

RESULTS

The results of our experiments are shown in Table 2. First off, the decision tree (DT) approach with noun and verb tags as features produced the highest performance with a precision of 97 percent, a recall of 96.80 percent and an f-measure of 96.80 percent. This outcome provides support for H1 that precision and recall over 90 percent can be achieved given our task of classifying basic science and clinical research.

The conditional probability (CP) approach was not able to achieve above 90 percent precision and recall. In all feature sets, the CP approach had false positives of basic science abstracts. The average numbers in Table 2 hide the fact that the algorithm was over identifying in the basic science category. With verbs alone as input, the CP algorithm had 100 percent recall on the basic science abstracts and 100 percent precision on the clinical research abstracts. However, using CP, basic science abstracts had a precision of 69 percent and clinical abstracts had a recall of 38 percent. Thus as shown in Table 2, you see the precision of 82.1 percent and recall of 74.2.

Using only verb tags as features in the algorithms, the conditional probability approach outperformed the decision tree approach by all measures. This outcome was a surprise to us. We had hypothesized just the opposite, that the CP algorithm would start out performing worse, but with more training data (NOUN + VERB tags), the CP approach would improve faster and outperform the decision tree. Thus, there is no support for H2. The decision tree showed larger improvements as more tags were added to its feature set and had the highest overall performance. Both algorithms did indeed improve by adding more tags (NOUNS + VERBS) to the feature set, the DT algorithm improved by a greater amount.

Table 2. Experimental Results

		Verb Tags	Noun Tags	Noun + Verb Tags
Conditional Probability (CP)	Precision	**82.10%**	84.80%	89.40%
	Recall	**74.20%**	83.90%	87.10%
	F-measure	**78.00%**	84.30%	88.25%
Decision Tree (DT)	Precision	68.70%	**92.10%**	**97.00%**
	Recall	67.70%	**90.30%**	**96.80%**
	F-measure	67.90%	**90.40%**	**96.80%**

Using just verb tags as features performed worse than using just noun tags as features. This result was true with both the CP algorithm and the DT algorithm. This outcome supports H3 that the algorithms using verb tags alone will not outperform algorithms using noun tags alone. One possible reason is that there are fewer tags from verbs than there are nouns from subjects and objects. Each abstract averaged nine verb tags and 19 noun tags. Having more tags provided more information for classification.
As expected, the noun + verb tag feature set outperformed all other feature sets. The performance of the CP and DT algorithms using features based on noun and verb tags provided support for H4.

DISCUSSION

A primary goal of this research was to create and test an algorithm that could correctly classify abstracts into basic science and clinical research categories with a precision and recall value over 90 percent. This goal was met with the decision tree algorithm, J48, using a shallow taxonomy of semantic noun and verb tags. The performance was achieved using a training set of 123 abstracts.

A surprise, however, was that the algorithm based on conditional probability did not improve as fast as the decision tree algorithm as training data was added. A success factor of the CP algorithm is how many testing document tags appeared in the training documents or tag overlap. If there is not a lot of overlap of salient tags, then the algorithm's performance suffers. Table 3 shows the percent of missing tags by feature set between abstracts that were correctly classified and abstracts incorrectly classified. With the verb tag inputs, incorrectly classified abstracts suffered from 45 percent missing tags, while the correctly classified abstracts had fewer missing tags at 38 percent. The lower performance of the CP algorithm with verb tag inputs resulted from having less tag overlap with the training set.

Table 3. Percent of unmatched tags from the testing set

	Verb Tags	**Noun Tags**	**Noun + Verb Tags**
Correctly Classified	**38%**	42%	42%
Incorrectly Classified	**45%**	42%	42%

Training set gaps, however, does not explain the lower performance of the CP algorithm given non verb-only feature sets. However, abstract length may provide some explanation of the lower performance. Table 4 shows the average number of tags per abstract per feature set. In the noun tag feature set, correctly classified abstracts had 30 tags while incorrectly classified abstracts had 25 tags. In the noun + verb tag feature set, correctly classified abstracts averaged 45 tags, while incorrectly classified abstracts averaged 38 tags. It appears that longer abstracts resulted in more tags. More tags given the same percent of unmatched tags means more total tags were matched between training and testing sets, which resulted in better classification performance.

Table 4. Average number of tags per abstract

	Verb Tags	**Noun Tags**	**Noun + Verb Tags**
Correctly Classified	15	**30**	**45**
Incorrectly Classified	15	**25**	**38**

Another possible factor differentiating the performance of the CP algorithm and the DT algorithm is the fact that we included only tags as features in the DT algorithm that had conditional probabilities above fifty-five percent. We grouped tags into TOP (85-100), MEDIUM (70-85), and LOW (55-85) categories for the decision tree. In other words, tags that were less predictive were ignored. This grouping removed some noise of those tags that appeared in both basic science and clinical research abstracts. In summary, the decision tree algorithm was more robust to testing abstracts of different lengths and unseen feature values.

CONCLUSIONS

Contributions from this paper include first, the introduction of the classification task with groupings of basic science and clinical abstracts. These two categories represent many abstracts in PubMed. Second was the creation and testing of an algorithm that could correctly classify abstracts into basic science and clinical research categories with a precision and recall value over 90 percent. In the process, we were able to conclude that tags serving as verbs in the abstracts are indeed useful to the classification and that including verbs does not diminish the impact of noun tags. Finally, using lexical-semantic tags as inputs proved to be a good representation of PubMed abstracts or in other words, our document representation generalized well to the testing set of abstracts. We used a relatively small training set, but the performance for the classification algorithm was still very high.

FUTURE WORK

Future direction for this research is to integrate the classification algorithm into an actual PubMed information retrieval system. We want to test the algorithm to see if the classification groupings actually have an impact on different search-related tasks. If the groupings do in fact help medical information seekers, we intend to add additional research categories to our classification to provide a finer-grained alignment of the medical abstracts to researchers needs.

REFERENCES

Ananiadou, S., Kell, D. B., & Tsujii, J.-i. (2006). Text mining and its potential applications in systems biology. *Trends in Biotechnology, 24*(12), 571-579. doi:http://dx.doi.org/10.1016/j.tibtech.2006.10.002

Chen, D., Müller, H.-M., & Sternberg, P. W. (2006). Automatic document classification of biological literature. *BMC Bioinformatics, 7*(1), 1-11. doi:10.1186/1471-2105-7-370

Friedman, C., Alderson, P. O., Austin, J. H., Cimino, J. J., & Johnson, S. B. (1994). A general natural-language text processor for clinical radiology. *Journal of the American Medical Informatics Association, 1*(2), 161-174.

Friedman, C., Kra, P., Yu, H., Krauthammer, M., & Rzhetsky, A. (2001). Genies: A natural-language processing system for the extraction of molecular pathways from journal articles. *Journal of Bioinformatics, 17*(1), S74-S82.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The weka data mining software: An update. *SIGKDD Explor. Newsl., 11*(1), 10-18. doi:10.1145/1656274.1656278

Haynes, R. B., Wilczynski, N., McKibbon, K. A., Walker, C. J., & Sinclair, J. C. (1994). Developing optimal search strategies for detecting clinically sound studies in medline. *Journal of the American Medical Informatics Association, 1*(6), 447-458.

Hersh, W., Mailhot, M., Arnott-Smith, C., & Lowe, H. (2001). Selective automated indexing of findings and diagnoses in radiology reports. *Journal of Biomedical Informatics, 34*(4), 262-273. doi:http://dx.doi.org/10.1006/jbin.2001.1025

Hirschman, L., Yeh, A., Blaschke, C., & Valencia, A. (2005). Overview of biocreative: Critical assessment of information extraction for biology. *BMC Bioinformatics, 6*(1), 1-10. doi:10.1186/1471-2105-6-s1-s1

Houston, A. L., Chen, H., Schatz, B. R., Hubbard, S. M., Sewell, R. R., & Ng, T. D. (2000). Exploring the use of concept spaces to improve medical information retrieval. *Decision Support Systems, 30*, 171-186.

Humphreys, K., Demetriou, G., & Gaizauskas, R. (2000). *Two applications of information extraction to biological science journal articles: Enzyme interactions and protein structures.* Paper presented at the Pac Symp Biocomput.

Hunt, D. L., & McKibbon, K. A. (1997). Locating and appraising systematic reviews. *Annals of Internal Medicine, 126*(7), 532-538. doi:10.7326/0003-4819-126-7-199704010-00006

Ibushi, K., Collier, N., & Tsujii, J. (1999). Classification of medline abstracts. *Genome Informatics, 10*, 290-291.

Imambi, S. S., & Sudha, T. (2011). Classification of medline documents using global relevant weighing schema. *International Journal of Computer Applications, 16*(3), 45-48.

Kilicoglu, H., Demner-Fushman, D., Rindflesch, T. C., Wilczynski, N. L., & Haynes, R. B. (2009). Towards automatic recognition of scientifically rigorous clinical research evidence. *Journal of the American Medical Informatics Association, 16*(1).

Lu, Z. (2011). Pubmed and beyond: A survey of web tools for searching biomedical literature. *Database, 2011*. doi:10.1093/database/baq036

McDonald, D., & Chen, H. (2006). Summary in context: Searching versus browsing. *ACM Transactions on Information Systems (TOIS)*.

McDonald, D., Chen, H., Su, H., & Marshall, B. (2004). Extracting gene pathway relations using a hybrid grammar: The arizona relation parser. *Bioinformatics*.

Montori, V. M., Wilczynski, N. L., Morgan, D., & Haynes, R. B. (2005). Optimal search strategies for retrieving systematic reviews from medline: Analytical survey. *BMJ, 330*(7482), 68.

Novichkova, S., Egorov, S., & Daraselia, N. (2003). Medscan, a natural language processing engine for medline abstracts. *Journal of Bioinformatics, 19*(13), 1699-1706.

Ono, T., Hishigaki, H., Tanigami, A., & Takagi, T. (2001). Automated extraction of information on protein–protein interactions from the biological literature. *Bioinformatics, 17*(2), 155-161.

Pentoney, C., Harwell, J., & Leroy, G. (2014). Does query expansion limit our learning? A comparison of social-based expansion to content-based expansion for medical queries on the internet. *AMIA Annual Symposium Proceedings, 2014*, 976-983.

Poulter, G. L., Rubin, D. L., Altman, R. B., & Seoighe, C. (2008). Mscanner: A classifier for retrieving medline citations. *BMC Bioinformatics, 9*(108).

Sarker, A., & Molla-Aliod, D. (2010, Dec 10, 2010). *A rule-based approach for automatic identification of publication types of medical papers.* Paper presented at the 15th Australasian Document Computing Symposium, Melbourne, Australia.

Schnapp, L. M., Vaught, M., Park, D. R., Rubenfeld, G., Goodman, R. B., & Hudson, L. D. (2009). Implementation and impact of a translational research training program in pulmonary and critical care medicine. *Chest, 135*(3), 688-694. doi:10.1378/chest.08-1449

Shojania, K., & Bero, L. (2001). Taking advantage of the explosion of systematic reviews: An efficient medline search strategy. *Effective Clinical Practice, 4*(4), 157-162.

Van der Laan, A. L., & Boenink, M. (2015). Beyond bench and bedside: Disentangling the concept of translational research. *Health Care Analysis, 23*(1), 32-49. doi:10.1007/s10728-012-0236-x.

Zhou, D., & He, Y. (2008). Extracting interactions between proteins from the literature. *Journal of Biomedical Informatics, 41*(2), 393-407. doi:http://dx.doi.org/10.1016/j.jbi.2007.11.008