LitLinker: Capturing Connections across the Biomedical Literature

Wanda Pratt[1,2]
Meliha Yetisgen-Yildiz[2]
[1] Biomedical and Health Informatics, School of Medicine, University of Washington, Seattle, USA.
[2] The Information School, University of Washington, Seattle, USA

ABSTRACT
The explosive growth in the biomedical literature has made it difficult for researchers to keep up with advancements, even in their own narrow specializations. In addition, this current volume of information has created barriers that prevent researchers from exploring connections to their own work from other parts of the literature. Although potentially useful connections might permeate the literature, they will remain buried without new kinds of tools to help researchers capture new knowledge that bridges gaps across distinct sections of the literature. In this paper, we present **LitLinker**, a system that incorporates knowledge-based methodologies, natural-language processing techniques, and a data-mining algorithm to mine the biomedical literature for new, potential causal links between biomedical terms. Our results from a well-known text-mining example show that LitLinker can capture these novel, interesting connections in an open-ended fashion, with less manual intervention than in previous systems.

General Terms: Algorithms

Categories and Subject Descriptors
I.2.7 Natural Language Processing
J.3 Life and Medical Sciences – Medical Information Systems
H.3.3 Information Search and Retrieval
H.2.8 Database Applications – Data Mining

Keywords
Text mining, knowledge-based systems

1. INTRODUCTION
With the explosive growth in the scientific literature, researchers have much information at their fingertips, but sifting through that information to identify nuggets of useful knowledge has become even more difficult. Because huge quantities of information are published each day, most researchers struggle to keep abreast of work within their own narrow specialization and spend little or no time examining the literature from other related disciplines. However, such isolation can stagnant research progress; many innovations occur only when traditional field boundaries are bridged. This problem is particularly acute in the domain of biomedicine. For example, MEDLINE, the primary bibliographic database for biomedicine, contains over 12 million references to journal articles, and over 2000 new references are added each day.[1] Obviously, no one is able to read about advancements across this entire body of published literature, even when they are motivated by the potential to save people's lives. Tools are needed to help them capture and explore the knowledge in the literature.

We have developed a system, called **LitLinker**, which provides a first step toward meeting this need. LitLinker bridges traditional field boundaries to identify and link together previously obscured connections in the biomedical literature. Our approach incorporates knowledge-based methodologies, natural-language processing techniques, and data-mining algorithms to mine the biomedical literature for new, potentially causal links between biomedical terms. The goal in this research is to develop a new type of retrieval approach that both will help researchers bridge gaps across specializations and will improve their ability to identify new research directions. To test LitLinker's performance, we explore how well it works on a known text-mining example.

2. RELATED WORK
Other researchers have shared this research goal and have been working in the general area of biomedical text mining for nearly fifteen years. Swanson was responsible for much of the earliest work in biomedical text mining.[2, 3] He used a combination of citation analysis and manual review. The former was used to determine novelty by detecting disjoint literatures. The latter was used to identify plausible new connections across disjoint biomedical literatures by examining the titles from search results. In an early example of this process, he identified a hidden connection between the disjoint literatures on migraines and on magnesium.[4] He noticed this hidden

K-CAP'03, October 23-25, 2003, Sanibel Island, FL, USA.

connection by identifying several linking, intermediate medical terms, such as *epilepsy* and *calcium channel blockers*, that occurred frequently in the titles of both the magnesium literature and the migraine literature. Thus, those linking terms were able to act as a bridge between a medical problem, *migraines*, and the mineral, *magnesium*, which did not occur in the titles from the migraine literature. The key to his approach was to assume one-level of transitivity held between correlated terms. In other words, the assumption is that if *migraine* is correlated with *epilepsy*, and *epilepsy* is correlated with *magnesium*, then *migraine* is correlated with *magnesium*. His text-mining success even spawned clinical studies to investigate several of his newly found connections. Clearly, Swanson's work introduced seminal ideas for text mining; however, a limiting factor for his approach is the large amount of manual intervention required. Although his more recent research with Smallheiser incorporates an interactive tool called Arrowsmith, much work still is required to setup customized lists of stop words and to sort through the many spurious connections that Arrowsmith generates.[5] For the migraine example, we identified over 2700 unique words in his set of migraine titles, but through the manual creation of a customized list of stop words (words that are assumed to provide little meaning and are eliminated from consideration), they reduced the number of words considered to about 200, thus, eliminating over 90% of the words manually. They recommend that this set of stopwords be revised for each new discovery task.

Also, Arrowsmith functions in a closed-discovery fashion; the user must enter both a starting a point and a candidate or broad-category for a candidate target concept. Arrowsmith's task is to generate the linking concepts between the starting concept and the target concept. Thus, Arrowsmith helps users explore hypotheses in the literature, but does not generate completely new connections.

Others have built on this work and added components to reduce the required manual intervention for that first step of identifying intermediate linking concepts. Lindsay and Gordon developed a process that followed the same basic architecture as Arrowsmith, but they added a variety of techniques to weight terms using information-retrieval methods, such as term frequency and inverse document frequency.[6] They evaluated the performance, in terms of precision and recall, for generating the linking terms, where Swanson's identified linking terms served as the gold standard. Gordon and Dumais also explored alternative techniques for identifying the linking literature. They used latent semantic indexing to extract close terms that occur in overlapping sets of documents. They followed Swanson's discovery approach and used his Raynaud's Disease and Fish Oil example to compare the performance of latent semantic indexing with the performance of term frequency and inverse document frequency methods used by Lindsay and Gordon.[7] In previous work, we used a knowledge-based approach to identify and prune potential linking terms. We found this approach produced more precise results than other approaches.[8] However, none of these researchers pursued or evaluated how easy the target term (e.g., *magnesium*) would be to identify.

In recent biomedical text mining work, Weeber et. al. also based their work on Swanson's approach.[9] They added both a natural-language processing component to identify biomedical terms and a knowledge-based approach to help prune spurious connections based on the semantic type of the connection term. These approaches still required a significant manual component for pruning possible connections, and none examined replicating Swanson's results in a purely automated, open-ended fashion.

In contrast, data-mining researchers have begun adapting their purely automated approaches for use with text. The most common approaches have been using association rules or term clustering. Much of that work and other work in information access has focused on identifying interesting phrases or correlations among terms within each document in a set of documents[10-12] or generating document categories and summaries,[13-16] rather than focusing on discovering new connections across documents. Although they make valuable contributions toward helping users see existing knowledge, none of these techniques incorporated Swanson's notion of an intermediate, linking literature, and, thus, do not emphasize finding previously unknown connections.

3. OUR APPROACH

In developing LitLinker, we have taken an approach that builds upon Swanson's fundamental idea of utilizing a linking literature, and incorporates techniques from both data mining and knowledge-based systems. Our approach offers several unique contributions. In contrast to Arrowsmith and the approaches that followed it, LitLinker uses data-mining techniques to identify correlations among concepts, and uses those correlations for open-ended discovery. In contrast to text-mining approaches from the data-mining community, LitLinker uses an intermediate literature to capture discoveries that are likely to be novel, and uses knowledge-based heuristics to limit the search space for such open-ended discovery.

A high-level view of the overall process within LitLinker is illustrated in Figure 1. LitLinker starts with a provided **starting concept**, which specifies the concept that the researcher wants to investigate. From Swanson's famous example of finding a connection between *migraine* and *magnesium*, this starting concept would be *migraine* (which Swanson refers to as the C-term).

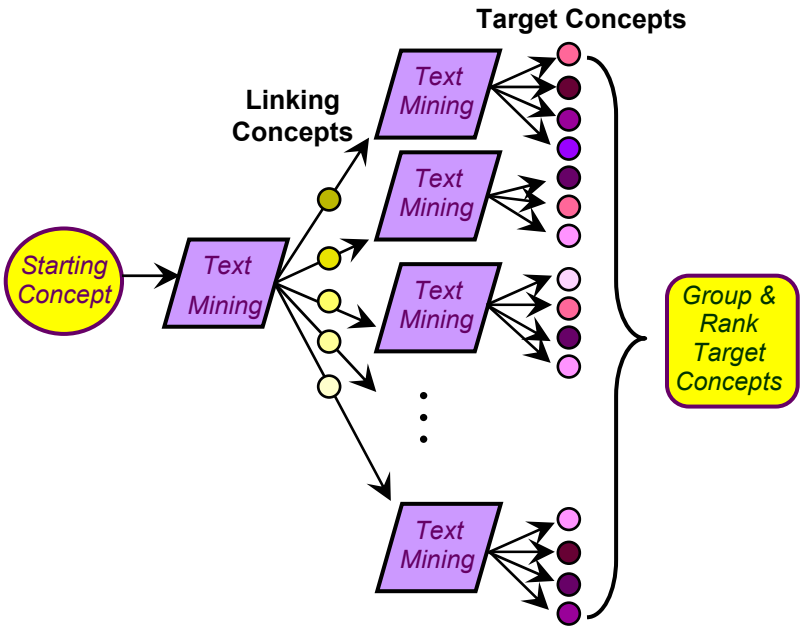

Figure 1 – The Discovery Process in LitLinker.

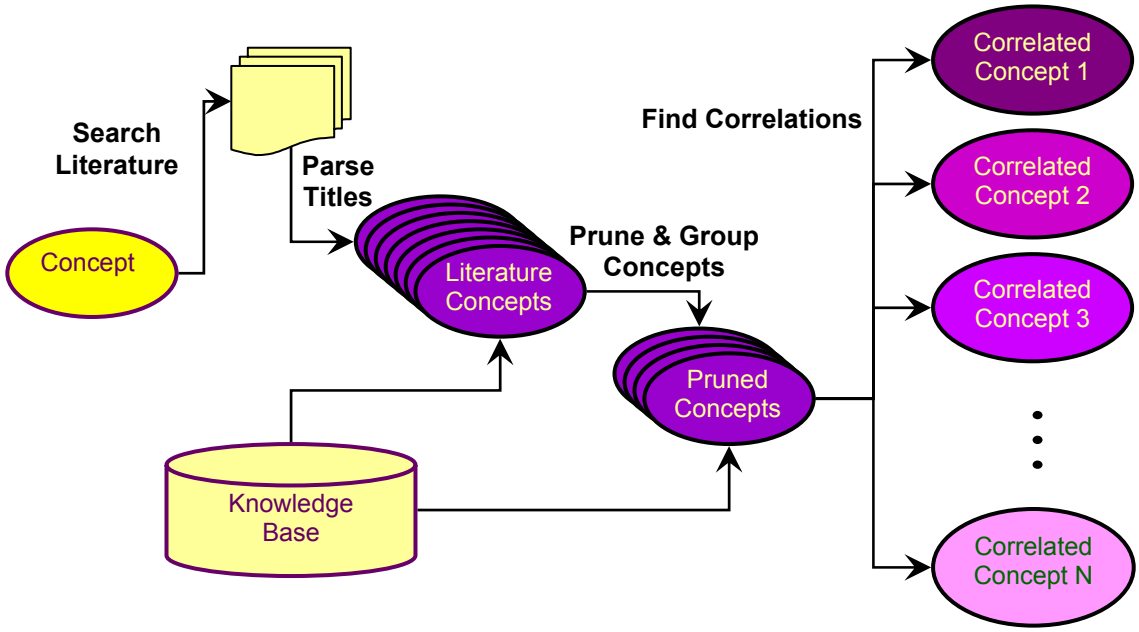

Figure 2 – The Text Mining Process.

Next, LitLinker goes through a text-mining process to find a set of terms that are correlated with the starting concept and could form the first step in connecting the starting concept to a new concept. We refer to this first set of correlated terms as the **linking concepts** (aka B-terms in Swanson's papers).

For each of those linking concepts, LitLinker uses the same text-mining process to identify a set of terms that are correlated with the linking concept. We call these final terms our **target concepts** (aka A-terms in Swanson's papers). In Swanson's migraine example, the identified target concept was *magnesium*.

Finally, LitLinker groups and ranks the target concepts by the number of linking concepts that connect the target to the starting concept. Thus, it provides an organized list of possibilities for this open-discovery process.

LitLinker uses a knowledge base as an integral component throughout the process. The knowledge base helps LitLinker both to identify all the medical terms, and to limit the search space by pruning away unhelpful terms

and forming groups of similar terms. The data-mining component plays a key role in determining which concepts are correlated with each other.

In the following sections, we describe in detail each of the major steps in the text-mining process, as well as the final step of assembling the target concepts.

3.1 Searching the Literature

To collect the literature for the starting concept as well as the linking concepts, LitLinker connects to the National Library of Medicine's (**NLM**) web-based client/server interface to MEDLINE and other databases, called Entrez.[17] It searches for all citations that contain the desired concept in the title. For the evaluation, we used *migraine* as the start concept. Because Swanson's work resulted in publications that directly discussed both *migraine* and *magnesium*, we limited the searches to all articles published prior to 1988, the dates over which Swanson performed his original quest. The originating search for *migraine* resulted in 2571 citations.

This literature collection step saves only the titles of each article to a database. In the biomedical literature, titles tend to summarize the main point of the article. For example, a common form of a title is to state the disease and the studied treatment or correlated factor (e.g., *Relief of common migraine by exercise* or *Asymmetric cerebral blood flow patterns in migraine*). Thus, by limiting the collection to titles, we both reduce the number of terms that must be processed and include only the highly descriptive terms of an article. In addition, for the evaluation, we wanted to compare our results to those of other researchers who used this same example and who limited their analyses to only titles as well.

3.2 Identifying Literature Concepts

As part of the text mining process, the first challenging step is to identify the relevant concepts from the literature. This step is a critical component for insuring the quality and the accuracy of the connections generated by LitLinker. To identify biomedical concepts, LitLinker uses a knowledge-based, natural-language-processing approach to process the document titles. A key part of our approach is to use a knowledge base to help identify domain-specific terms. The biomedical domain already has a large, publicly available knowledge base called the **Unified Medical Language System** (**UMLS**). This knowledge base was created by the NLM and contains over 875,000 biomedical concepts as well as 2.1 million concept names.[18] The system was created by unifying hundreds of other medical knowledge bases and vocabularies to create an extensive resource that provides synonymy links as well as parent-child relationships among the single or multi-word concepts.

LitLinker uses MetaMap, a tool created by NLM that maps from free text to biomedical concepts from the UMLS.[19] MetaMap uses the Xerox tagger to assign syntactic parts of speech and then uses the tags to identify phrases. It uses the UMLS to find the closest matching known concept to each identified phrase.

LitLinker sends each collected title to MetaMap and saves every concept that MetaMap identifies in a data field linked to the title. LitLinker also groups together all the synonymous terms by merging any that have the same concept id, and then assigns a preferred name to that group. Weeber et. al. also used MetaMap to identify concepts, although it is unclear whether they merged synonymous terms.[9]

3.3 Pruning Concepts

The previous step of identifying biomedical concepts generates many concepts, but only a subset of those concepts would make medically plausible linking terms. We found three classes of problems: (1) many terms were too general (e.g., *problem, test*, *therapeutic*); (2) some terms were too closely related to the start term (e.g., *headache*, *retinal migraine*); and (3) some terms just did not make sense as plausible connections.

For the first class of problems, we chose a pragmatic approach similar to that used in information retrieval to prune away uninformative terms that are so general or prevalent that they are correlated with most biomedical terms. Our original plan was to use the UMLS knowledge base for this determination. We observed that many general concepts, such as *disease* and *drug*, appear on the second level (children of roots – which were the names of the biomedical vocabularies) and third level (children of the concepts on the second level) of the UMLS hierarchy. Thus, LitLinker marks the concepts that are in those levels as too general, and prunes away all such terms. However, such a process eliminated only a small fraction of the general terms. The problem is that the UMLS is a conglomeration of many different biomedical vocabularies, each of which makes different assumptions about the meaning of the levels in the hierarchy, and many very general terms appeared much lower in some vocabulary's hierarchy. Thus, we added another component that would prune terms that seemed too pervasive in the literature; LitLinker eliminated any concepts that appeared in the titles of more than 10,000 MEDLINE documents.

For the second class of problems, where the identified concepts were closely related to the starting concept, LitLinker used the UMLS to determine all the parents and children of the start concept, and then eliminated all those related terms. Thus, for the migraine example, LitLinker eliminated terms such as *headache* and *retinal migraine* as potential linking concepts because they were too closely related to the starting concept of *migraine*.

For the third class of problems, the challenge was to create an automated, generalizable approach to pruning away those implausible, uninteresting terms. We again

turned to our knowledge base for help with such an approach. In the UMLS, each concept is connected through an *isa* link to one or more semantic types from a small set of general medical terms in a hierarchy that the NLM calls the Semantic Network. For example, the term *migraine* has a semantic type of *disease or syndrome*, and *magnesium* has a semantic type of *biologically active substance*. We selected a subset of semantic types that were plausible for terms that could be correlated with a medical condition or disease and a potential treatment (see Table 1). Then, LitLinker pruned away any concepts that did not match the semantic-type criteria. A similar approach has been used in other work on biomedical text mining,[8, 9] and in organizing search results.[16] In contrast, Swanson addressed this problem by manually creating a query-customized list of stop words to filter out the uninteresting concepts, but such a level of word-based customization could be difficult to scale to new kinds of query concepts and connections.

Table 1 The allowable UMLS semantic types for pruning implausible medical concepts.

Semantic Type
laboratory or test result
clinical attribute
fully formed anatomical structure (and 5 sub-types)
substances (and 23 sub-types except for materials, diagnostic aids, or hazardous substances)
organ or tissue function
organism function (and 1 sub-type)
pathologic function (and 3 subtypes except for neoplastic process or experimental model of disease)

3.4 Finding Correlations

A key part of our text-mining approach is the process of identifying associated or correlated concepts. This process produces both the linking concepts and the target concepts. To identify the correlated concepts, LitLinker uses Apriori, the reference algorithm for identifying association rules.[20] Specifically, we used ARtool, a free, association rule mining application that was written in Java and distributed under the GNU General Public License.[21] For our approach to text-mining, a starting literature already provides the context for associations; thus, we only care about the first component of Apriori, which generates frequent itemsets. This data-mining approach requires a specific level of support before generating the frequent itemsets. The **level of support** indicates the probability that a term will occur in the document collection. For our approach, level of support is calculated as the number of titles with the concept divided by the total number of titles.

We decided that associations were likely to be spurious unless the concept occurred in at least five titles, which resulted in a support level of 0.002 for identifying the linking concepts in our migraine example. The initial number of identified linking concepts for this example was 167. To identify the target concepts for each of the linking concepts, the support level was set in the same manner, requiring at least five occurrences in the titles. In an interactive version, it could be helpful to allow users to specify such thresholds of support.

Initially, the concept for the literature search is the starting concept and the correlated concepts that are found in this step act as the linking concepts. LitLinker repeats the same process, using each linking concept to initiate a separate literature search. The resulting correlated concepts for each linking concept create the total set of target concepts.

3.5 Grouping Concepts

This step of grouping related concepts occurs before the step to find correlations, but aspects of the finding correlations step are needed to understand the importance of grouping concepts. For text mining, we are looking for high-level but strong patterns among terms, and those patterns could go undetected if similar terms are not grouped together. It does not make sense to make subtle distinctions of meaning in this process. For example, distinguishing among the many ways to measure magnesium in the body is unimportant; the key concept is *magnesium*. This problem of subtly different concepts is particularly acute because a minimum level of support is needed to find a correlation. However, if one central concept is represented by many subtly different words or phrases, the level of support might be too low to be noticed.

Initially, we assumed that the synonymy links in the UMLS knowledge base would help LitLinker group together such subtly different concepts. However, those synonymy links are only among terms that have **exactly** the same meaning. For example, *migraine* and *migrainous headache* are synonyms in the UMLS, but *common migraine* and *migraine* are not considered synonyms because common migraine is just one of several kinds of migraines. Although for many biomedical applications such a fine level of distinction is appropriate, for text mining, we need to consider a much coarser level of synonymy that would allow us to ignore such subtle distinctions.

Because the UMLS could not provide us with a clear method for determining this coarse level of similarity, we created a combined semantic and lexical approach to address this problem. The basic intuition behind this approach is to cluster related concepts into a group labeled with the shortest concept name that is a common subset of all the terms in a group. For example, *migraine* and *common migraine* would be grouped into a cluster named *migraine.* The goal is that the resulting group name would represent a cluster of similar terms that would be used to identify target concepts.

To accomplish this grouping goal, LitLinker first transforms each multi-word concept into all possible adjacent word subsets. For example, LitLinker converts the UMLS concept *blood magnesium level* into six different forms: *blood*, *magnesium*, *level*, *blood magnesium*, *magnesium level*, and *blood magnesium level*. Some forms, such as *level*, are concepts that are too general to be informative. As with the pruning step, LitLinker eliminates all such general terms automatically by looking at their level in the UMLS hierarchy. It also removes all words or phrases that are not valid terms in the UMLS. For example, this grouping process produces a cluster called *magnesium*, which includes 25 similar, but not strictly synonymous terms, such as *magnesium deficiency*, *magnesium ion*, *urine magnesium*, and *blood magnesium level*.

If we consider Swanson's *migraine-magnesium* example, *epilepsy* is one of the linking concepts, but LitLinker would never identify it as a linking concept without grouping. The main reason for this oversight is that the word *magnesium* has a very low support value in the *epilepsy* literature. However, three other concepts that are closely related to *magnesium* (*serum magnesium*, *magnesium level*, and *magnesium deficiency*) also appear in the epilepsy literature. Individually, those related concepts also have insufficient support values, but with grouping, the sum of their support values passes the support test and correctly designates *epilepsy* as a linking concept that connects *migraine* and *magnesium*. Without this grouping step, many potentially interesting concepts would have been overlooked.

3.6 Assembling Target Concepts

Assembling target concepts from all the linking concepts requires multiple processing steps. First, LitLinker merges the lists of target concepts from each of the linking concepts. It also retains a connection to each of the linking concepts that connect that target to the starting concept.

Second, because we are interested in only novel connections, LitLinker must prune previously known connections from the list of target concepts. We decided that any co-occurrence with the starting concept constituted a known connection. Thus, LitLinker checked each candidate target term against the entire set of UMLS concepts that were extracted for the starting literature. If a candidate target term was an element of this set, LitLinker eliminated it as a potential target term.

The final result of the assembly process is to list the target concepts together with their linking concepts. LitLinker ranks the target concepts according to the number of linking concepts that connect that target concept to the original start concept. Such a list should provide enough information to help researchers evaluate and explore these possible correlations and determine which seem worthy of further investigation.

4. RESULTS FOR MIGRAINE EXAMPLE

To explore how well our new approach to text mining works, we ran LitLinker with *migraine* as a starting concept and compared the results to Swanson's example that finds a connection between *migraine* and *magnesium*. We compare LitLinker's ability to find Swanson's specified linking connections as well as its ability to identify *magnesium* as a correlated and potentially causally related target term.

4.1 Results for Linking Concepts

LitLinker identified a total of 118 linking concepts that met our level of support for the *migraine* literature, but only 29 of those concepts linked *migraine* to *magnesium*. In Swanson's work, he identified eleven valid connections between *migraine* and *magnesium*, and LitLinker was able to identify five of those.

Four of the linking concepts that LitLinker did not identify (*inflammation, substance p, brain hypoxia*, and *vascular tone and reactivity*) were too infrequent in the *migraine* literature to satisfy our level of support threshold. For example, only three titles in the *migraine* literature include *substance p* and none include *inflammation*.

One other missed linking concept, *spreading cortical depression*, could not be retrieved as a linking concept because MetaMap never identified it as a medical concept. The MetaMap lexical analyzer breaks *spreading cortical depression* into two separate phrases, *spreading* and *cortical depression*. The program then eliminates *spreading* because it is not a concept in the UMLS and maps *cortical depression* to two separate terms, *cortex* and *depression*. This problem illustrates how important the step of identifying biomedical concepts is.

For the final missed concept, *personality*, LitLinker did include it in the total set of 118 linking concepts, but it was not in the 29 correlated with *magnesium* because it did not occur in any titles from the *magnesium* literature. However, LitLinker did identify an additional 24 linking terms that could help a biomedical researcher understand how *magnesium* and *migraines* are related. In Table 2, we have listed the 29 concepts that link *migraine* to *magnesium*. In future work, it will be interesting to get input from migraine researchers on the legitimacy and usefulness of our larger list of linking concepts. Swanson used his own expertise to choose his 11 linking concepts, but they were never verified with experts.

4.2 Results for Magnesium as Target

To identify target terms for *migraine*, LitLinker analyzed a total of 223,413 titles that contained the starting concept or any of the 118 linking concepts. After the pruning and ranking stage, *magnesium* is tied at a rank of 11 in the total list of 528 target concepts. Figure 3 illustrates the top 12 target concepts and the number of concepts that link them to the starting concept, *migraine*.

Table 2 – The 29 linking concepts for connecting *migraine* and *magnesium*. * denotes the concepts found by Swanson.

Linking Concepts	
prolactin	*heparin*
propranolol	*histamine*
*epilepsy**	*hydroxytryptamine*
tryptophan	*estradiol*
aspirin	*progesterone*
reserpine	*caffeine*
cerebrospinal fluid	*calcium antagonist**
*serotonin**	*lithium*
*prostaglandins**	*nifedipine*
hemodynamics	*antagonists*
blood platelets	*free fatty acids*
substances	*clinical aspects*
*platelet aggregation**	*arteries*
cerebral cortex	*catecholamines*
muscle contraction	

Although the point of the pruning step is to eliminate general, uninformative terms, that process is not perfect. In the list of top 12 ranked target terms in Figure 3, there remain two such overly general terms: *liquid substance* and *enhancer*. In addition, some potentially interesting target concepts are pruned from the original list. For example, concepts such as *sodium* and *rheumatoid arthritis*, were pruned but seem like reasonable target concepts. LitLinker eliminated those two examples because the number of titles that they appear in exceeded our threshold for too frequent terms, which was meant to and did eliminate many of the overly general terms. Thus, future experiments should examine the effect of this threshold in all the pruning phases. Another possible solution to the pruning problem would be to create an interactive version of LitLinker. An interactive feature would allow a researcher to have control over both the support thresholds and the pruning heuristics to balance the task of sorting through extraneous target terms against the possibility of missing important target concepts.

If we were to evaluate the results without the final pruning step, *magnesium* would be at a tied ranking of 23rd out of a total of 564 concepts. Thus, the pruning did help make *magnesium* more visible to a researcher searching for connections, but even without the final pruning stage, *magnesium* still would have been an obvious connection. The other target concepts from LitLinker's results could provide interesting insights to migraine researchers as well, but we have not yet asked experts to evaluate the list.

5. CONCLUSION

With the explosion of the scientific literature, text mining systems such as LitLinker will become critical for helping researchers make discoveries across distinct portions of the biomedical literature. We have shown that our combination of a data-mining approach and a knowledge-based approach can be incorporated into an effective text-mining system. For Swanson's famous *migraine-magnesium* example, LitLinker automatically identified

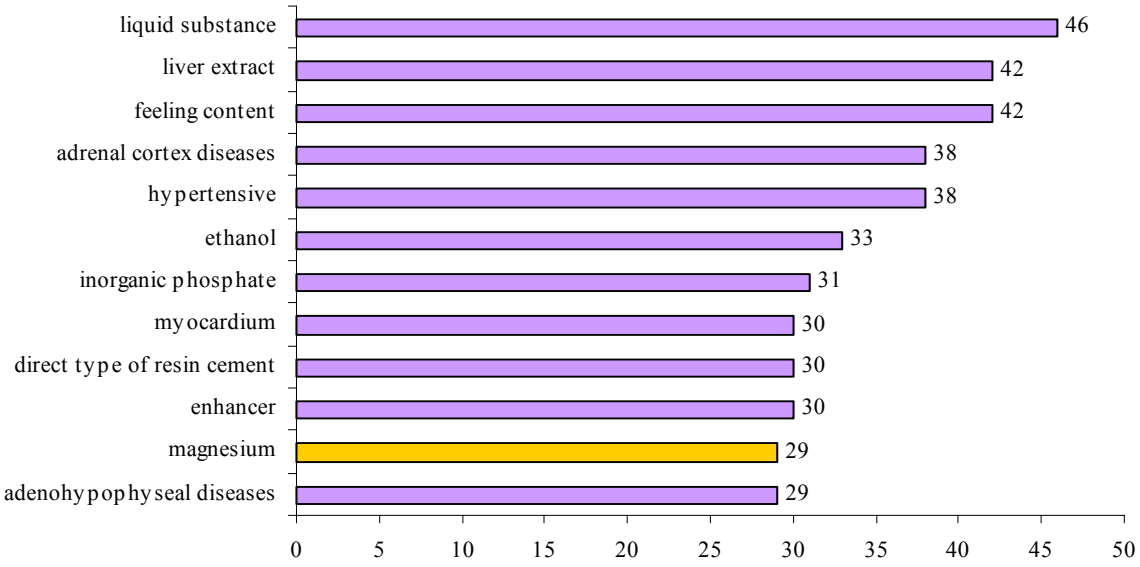

Figure 3 – The First 12 Target Concepts and the Number of Concepts that Link them to Migraine. The chart shows the number of linking concepts out of 118 possible linking concepts. The only target concepts listed are the first 12 out of the 528 target concepts identified. Magnesium is ranked as tied in 11th place.

magnesium in a short list of concepts that could be causally linked to *migraine*.

LitLinker also identified a large set of linking terms that connect *magnesium* to *migraine*. Although the results did not include all of Swanson's 11 linking terms, LitLinker detected an additional 24 plausible links that previous approaches have never identified.

Evaluating knowledge discovery systems is a fundamentally challenging task because if they are successful, by definition they are capturing new knowledge that has yet to proven useful. Thus, we evaluated LitLinker on the most widely used example case for text mining. Unlike previous systems which explored this example, LitLinker was able to make this connection in a purely open-ended discovery process. Other researchers have focused mainly on replicating the linking terms in Swanson's examples, and none have provided a ranking of possible target terms.

The system that we described operated automatically in a batch mode to identify potential new target terms. Our plan is to build on this base and extend LitLinker into the type of interactive system that researchers need to identify, capture, and explore new discoveries in the vast biomedical literature.

6. ACKNOWLEDGEMENTS

This material is based upon work supported by the National Science Foundation under Grant IIS-0133973.

7. REFERENCES

1. NLM. *MEDLINE Fact Sheet*. 2003. http://www.nlm.nih.gov/pubs/factsheets/medline.html
2. Swanson, D.R., *Medical literature as a potential source of new knowledge.* Bulletin of the Medical Library Associaton, 1990. **78**(1): p. 29-37.
3. Swanson, D.R., *Online search for logically-related non-interactive medical literatures: A systematic trial-and-error strategy.* Journal of the American Society for Information Science, 1989. **40**(5): p. 356-358.
4. Swanson, D.R., *Migraine and magnesium: eleven neglected connections.* Perspect. Biol. Med., 1988. **31**: p. 526-557.
5. Swanson, D.R. and N.R. Smalheiser, *An interactive system for finding complementary literatures: a stimulus to scientific discovery.* Artificial Intelligence, 1997. **91**: p. 183-203.
6. Lindsay, R.K. and M.D. Gorden, *Literature-based discovery by lexical statistics.* Journal of the American Society for Information Science, 1999. **50**(7): p. 574-587.
7. Gordon, M.D. and S. Dumais, *Using latent semantic indexing for literature based discovery.* Journal of the American Society for Information Science, 1998. **49**(8): p. 674-685.
8. Blake, C. and W. Pratt. *Automatically Identifying Candidate Treatments from Existing Medical Literature*. in *AAAI Spring Symposium on Mining Answers from Texts and Knowledge Bases*. 2002. Palo Alto, California.
9. Weeber, M., et al., *Using concepts in literature-based discovery: simulating Swanson's Raynaud-fish-oil and migraine-magnesium discoveries.* Journal of the American Society for Information Science, 2001. **52**(7): p. 548-557.
10. Feldman, R., I. Dagan, and H. Hirsh, *Mining text using keyword distributions.* Journal of Intelligent Information Systems: Integrating Artificial Intelligence and Database Technologies, 1998. **10**(3): p. 281-300.
11. Ahonen, H., et al. *Applying Data Mining Techniques for Descriptive Phrase Extraction in Digital Document Collections*. in *Proceedings of the IEEE Forum on Research and Technology Advances in Digital Libraries*. 1998. 2-11.
12. Nahm, U.Y. and R.J. Mooney. *Using Information Extraction to Aid the Discovery of Prediction Rules from Text*. in *KDD-2000 Workshop on Text Mining*. 2000. Boston.
13. Clifton, C., et al. *TopCat: data mining for topic identification in a text corpus*. in *Principles of Data Mining and Knowledge Discovery. Third European Conference, PKDD'99*. 1999. Prague, Czech Republic. 174-83.
14. Feldman, R., et al. *Text mining at the term level*. in *Proceedings Principles of Data Mining and Knowledge Discovery. Second European Symposium, PKDD '98*. 1998. 65-73.
15. Srinivasan, P. and T. Rindflesch. *Exploring text mining from MEDLINE*. in *Proc AMIA Symp*. 2002. 722-6.
16. Pratt, W., M.A. Hearst, and L.M. Fagan. *A Knowledge-Based Approach to Organizing Retrieved Documents*. in *AAAI '99: Proceedings of the Sixteenth National Conference on Artificial Intelligence*. 1999. Orlando, FL. p. 80-85.
17. NLM. *The Enrez Database System*. 2003. http://www.ncbi.nlm.nih.gov/Web/Search/entrezfs.html
18. NLM. *The UMLS Fact Sheet*. 2003. http://www.nlm.nih.gov/pubs/factsheets/umls.html
19. Aronson, A. *Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program*. in *Proc AMIA Symp*. 2001. 17-21.
20. Agrawal, R., et al., *Fast Discovery of Association Rules*, in *Advances in Knowledge Discovery and Data Mining*, U. Fayyad, et al., Editors. 1995, AAAI/MIT Press.
21. Cristofor, L. *ARtool*. 2002. http://www.cs.umb.edu/~laur/ARtool/