

# Automatically Identifying Candidate Treatments from Existing Medical Literature

Catherine Blake and Wanda Pratt

Information and Computer Science  
University of California, Irvine  
Irvine, California 92697-3425  
{cblake, pratt}@ics.uci.edu

## Abstract

The scientific literature available to researchers continues to increase at an alarming rate. Despite the wealth of knowledge within the published literature, the quantity and unstructured nature of those texts make it difficult to use them for answering research questions. Thus, many potentially useful connections among the documents go unnoticed. To address this problem, we developed three approaches to detect such connections automatically. The simplest approach used only words in document titles. The other two approaches used knowledge from an existing terminology model; one used the knowledge base to transform the titles to known medical concepts, and the other applied additional semantic constraints to prune concepts. To determine effectiveness, we compared each approach on the task of identifying a set of now-known but previously implicit connections in the biomedical literature, which suggest magnesium would be effective in treating migraines. The concept representation improved precision from 8.3 to 9.8% and recall from 22.7 to 30.1% when compared with using word features. Applying additional semantic constraints improved precision (22.3%) with only a small degradation in recall (19.4%).

## Introduction

The goal of many biomedical researchers is to answer, for a given medical condition, the question: “*What new promising treatment strategies should we investigate?*” Researchers use a variety of information sources to answer this question, including results from their own research and findings reported by other experts. Unfortunately, the quantity of information available makes it difficult to keep up with developments even within their own narrow field of expertise. The primary source of medical citations, MEDLINE, currently contains more than eleven million entries and grows by about 8,000 citations each week (NLM 2001). Our goal is to help researchers identify new plausible treatment strategies for a medical condition by automating the process of mining new treatments from MEDLINE.

To be useful to researchers, the answer to this question should be more than a list of all possible treatment strategies. The answer should also indicate which treatments are more likely to be effective and why. We use the physiological characteristics that co-occur with a medical condition to order and prune treatment options. Researchers can also use these characteristics to gain insight into the underlying mechanisms of a candidate treatment.

This goal is not new; Swanson and Smalheiser have shown that such text mining is indeed possible (Swanson 1988; Swanson and Smalheiser 1997), and other researchers have expanded on their idea (Lindsay and Gordon 1999; Weeber 2000). For example, Swanson identified eleven logical connections that suggest magnesium would be effective in treating migraines that were implicit in the biomedical literature before 1988 (Swanson 1988). The semi-automated system, ARROW-SMITH, has supported the discovery of seven other new treatment hypotheses, such as using fish oil to treat Raynaud’s disease (Swanson and Smalheiser 1997). Medical scientists are now conducting traditional clinical trials to verify each of the new proposed treatments.

The physiological connections that suggest magnesium would be effective in treating migraines span many areas of research. For example, experts in mental health were probably aware of the co-occurrence between migraines and depression or other brain function problems. Experts in blood probably knew about the relationship between migraines and blood platelet activity. It was only after Swanson considered this distributed knowledge, together with magnesium properties that he was able to infer the new candidate treatment. The journal *Magnesium* had identified a connection between magnesium and migraines in 1984 and 1985, however it is unlikely that clinicians that treat migraines would read this highly specialized journal unless prompted from a more widely read source.

Previous work has yielded valuable insights into new treatments, however substantial manual intervention has been required to reduce the number of possible connections. In contrast, we explore and evaluate fully automated approaches to this problem. Our approach replaces manual ad-hoc pruning by using an existing

knowledge-based. Our use of an intermediate set of physiological conditions to partition the medical condition from the proposed set of treatments helps to manage the sizable branching factor. We used three levels of semantic processing: no semantic processing (i.e., using title words only), transforming words to known concepts, and pruning those concepts based on the semantic constraints. To compare effectiveness, we calculated precision and recall of each approach in detecting the eleven connections between magnesium and migraines, which were implicit in the medical literature prior to 1988 (Swanson 1988).

## Approach

We designed an experiment to explore the effect of alternative representations and semantic pruning in identifying known physiological conditions suggesting magnesium as an effective treatment for migraines. The system will identify these connections automatically using bibliographically disjoint biomedical literature.

The next section describes the process of making inferences between bibliographically disjoint medical literature, independently of the chosen representation. We then describe how the system generates the word and concept representations and the process that we used to select semantic types to prune the medical concepts.

### Process Overview

The inference process described below has been adapted from earlier work (Swanson and Smalheiser 1997). This text-mining process begins by downloading MEDLINE titles related to a medical condition, in this case, migraines. Using a search strategy that requires *migraine* to appear in the title, we retrieved 2571 citations which were published before January 1 1988. This citation set (referred to as the **C-literature**) contains 225 more titles than the initial study because we included the second half of 1987.

- 
- (1) C-literature  $\leftarrow$  citations from MEDLINE containing the word 'migraine'
  - (2) B-terms  $\leftarrow$  words or medical concepts generated for each title in the C-literature
  - (3) B-literature  $\leftarrow$  citations from MEDLINE containing any of the B-terms
  - (4) A-terms  $\leftarrow$  words or medical concepts generated for each title in the B-literature
- 

**Figure 1 The text-mining process used to identify logical inferences from the medical literature. Adapted from (Swanson and Smalheiser 1997).**

After collecting the C-literature, the system automatically generates a list of physiological conditions associated with the medical condition: the **B-terms**. We considered using various parts of the citation text in the C-literature, such as the abstract, however basing B-terms on titles only enables our automated approach to be compared with the initial ARROWSMITH system. Our system represents B-terms at

three different semantic levels: words (no semantics), concepts (any medical concepts which occurs in the knowledge base) and pruned concepts (only medical concepts of a particular semantic type).

Unlike the initial study, we did not manually prune any of the B-terms (word or medical concepts). We did however remove **stopwords**, that is words which have little meaning such as *and* and *the*. We used a generic set of 417 words developed independently for the purposes of information retrieval (Sanderson 1999). We added to this numbers, days of the week and month names to this list. We also used a second set of 31 stopwords, such as *study* and *test* that are not meaningful in the medical domain. We developed these terms for our previous work on generating medical concepts directly from biomedical text (which we describe briefly in the following section) (Blake and Pratt 2001). The system removes each of the 448 stopwords from both the word and medical concept representations. Our experiment employs a much smaller stopword list than the 5,000 stopwords developed during the initial study for the purposes of pruning B-terms.

In addition to a larger stopword list, the initial study manually constrained B-terms to exogenous agents such as deficiencies, dietary factors, toxins, and drugs or environmental factors. Our approach replaces this manual pruning phase with an automated knowledge-based approach, which prunes medical concepts based on their semantic type (described in the feature representation section below). Our system orders B-terms in the same way as the initial study; in decreasing order of frequency.

During the third phase, the system retrieves citations from MEDLINE, which relate to each B-term. The system repeats the search strategy from step (1), however instead of searching for a medical condition; the system retrieves all documents related to the physiological characteristics associated with the medical condition. We refer to the set of citations retrieved using B-terms as **B-literature**. It is from the B-literature that the system generates a candidate set of treatments: the **A-terms**. The system orders A-terms based on their frequency and on the number of physiological characteristics that link the treatment to the medical condition (i.e. the number of B-terms). We had considered weighting A-terms by term frequency and inverse document frequency (tfidf), however previous results based on tfidf weighting were inconclusive (Lindsay and Gordon 1999). It is also unclear if a system should assign a higher weight to a treatment (an A-term) which is effective at distinguishing one citation from another (tfidf weighting) without considering the physiological characteristics of the medical condition.

### Feature Representation

Previous approaches to identifying treatments from medical literature represent the text as words or phrases (Lindsay and Gordon 1999; Weeber 2000). We define a word as a set of characters separated by a space (i.e. we did not use hyphens). Our system removes each of the 448 stopwords (417 generic stopwords and 31 specific to

medicine) in addition to terms that are synonymous with *migraine*, such as *migrainous* and *headache*. All other words are included in the word feature set.

Although the concept representation is semantically richer than words, we do not claim that it is a substitute for deep natural language understanding. Unlike many natural language systems, our system does not use a part-of-speech tagger to identify candidate phrases. Instead, it uses a simple heuristic, the stopwords, to partition each sentence into a set of clauses. Each clause is then mapped to one of the 800,000 concepts and 1.9 million concept names within the **Metathesaurus**, a component of the **Unified Medical Language System (UMLS)** (NLM 2000). For example, the clause *5-Hydroxytryptamine* and *brain serotonin relevance* are both mapped to the medical concept *Serotonin*. Serotonin is a neurotransmitter that has been associated with depression, obsessive compulsive disorder, aggressive behaviors and perception.

---

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)

---

**Table 1 The UMLS semantic hierarchy branches used to filter medical concepts.**

In addition to the metathesaurus, the UMLS contains 132

high-level, hierarchically organized concepts called semantic types. Each medical concept is associated a subset of semantic types. For example, Serotonin is an *Organic Chemical*, a *Pharmacologic Substance* and a *Neuroreactive Substance or Biogenic Amine*. Our system generates the third feature set by constraining medical concepts to the semantic types displayed in table 1.

## Results

We measured the precision and recall performance of each feature representation using the known physiological connections related to migraines identified by Swanson as our gold standard (Swanson 1988). We focus on steps 1 and 2 in the text-mining process: the C to B connections.

### Determining Relevance

Our universe of documents consists of the 2571 MEDLINE titles published before 1988 with the word *migraine* in the title. We calculated the number of relevant documents by identifying synonymous terms for each connection using Lindsay and Gordon's earlier work and the UMLS (see table 2). For example, titles represented as the concept *Leao depression*, or containing either the word *leao* or *depression* would be considered potentially relevant to the *spreading cortical depression* connection (table 2, item 4). The first author then manually inspected each of the 461 MEDLINE titles (this was necessary because not all titles with *calcium* refer to *calcium channel blockers*). Of the 451 titles (we were unable to classify ten titles), 366 referred to at least one of the known connections.

| B-term                           | Synonyms  | C-Liter<br>ature | Word | Medical<br>Concept | Semantic<br>Pruning |
|----------------------------------|---|------------------|------|--------------------|---------------------|
| 1. Type A personality            | None listed   | 26               | 3    | 7                  | 3                   |
| 2. Vascular tone and reactivity  | Vascular resistance, vascular responses   | 26               | 3    | 7                  | 4                   |
| 3. Calcium channel blockers      | Calcium antagonists, exogenous calcium blockaders, exogenous calcium inhibitors, exogenous calcium antagonist   | 36               | 14   | 9                  | 6                   |
| 4. Spreading cortical depression | Leao depression   | 15               | 5    | 5                  | 4                   |
| 5. Epilepsy                      | Seizure disorder, seizure syndrome, epileptic disorder, epileptic, epileptic convulsions, epilepsy epileptic attack, epileptic seizures, epileptic fits | 40               | 9    | 8                  | 2                   |
| 6. Serotonin                     | 5-HT, 5-hydroxytryptamine, 5-hydroxytryptamine, serotonergic, enteramine, serotonergic, 5-hydroxytryptamine, 5HT hippophaine,                           | 93               | 17   | 11                 | 10                  |
| 7. Platelet activity             | Platelet aggregation, platelet aggregability  | 111              | 8    | 15                 | 13                  |
| 8. Inflammation                  | Inflammatory reaction, inflammatory infiltration, leukocytic infiltrate, inflammatory cell infiltration,  | 1                | 3    | 3                  | 2                   |
| 9. Prostaglandins                | prostanoids, ketoprostanoglandin, PG-prostanoglandin,   | 14               | 3    | 3                  | 3                   |
| 10. Substance P                  | Euler-Gaddum substance P, SP(1-11), stress SP-substance P, neurokinin 1   | 1                | 4    | 10                 | 3                   |
| 11. Brain hypoxia                | Cerebral anoxia, anoxia of brain, cerebral hypoxia  | 3                | 14   | 32                 | 21                  |

**Table 2: Synonymous terms for each of the valid connections identified by Swanson and Smallheiser.**

Despite the disparity of the number of titles that supported a connection between migraines and each of the physiological conditions identified by Swanson and migraines, the medical literature successfully identified all of the logical connections (see table 2). The most supported connections were between migraines and *serotonin* (93 titles) and *platelet activity* (111 titles). Interestingly, only one title referred to *substance P* before 1988, however three studies have since verified that patients suffering from migraines have higher levels of substance P than control groups (Marukawa, Shimomura et al. 1996; May and Goadsby 2001; Nakano, Shimomura et al. 1993).

We considered the rank of each connection separately which we summarize in table 3. The rank of the semantically pruned concepts was higher than either the word or the concept representations in eight of the eleven connections. Although all representations identified all connections, we postulate that in practice a researcher would only look at the more highly ranked terms. If we restrict the number of terms considered by a researcher to the top fifty, then the word, concept and semantic pruned representations would have missed seven, five and two of the implicit connections respectively.

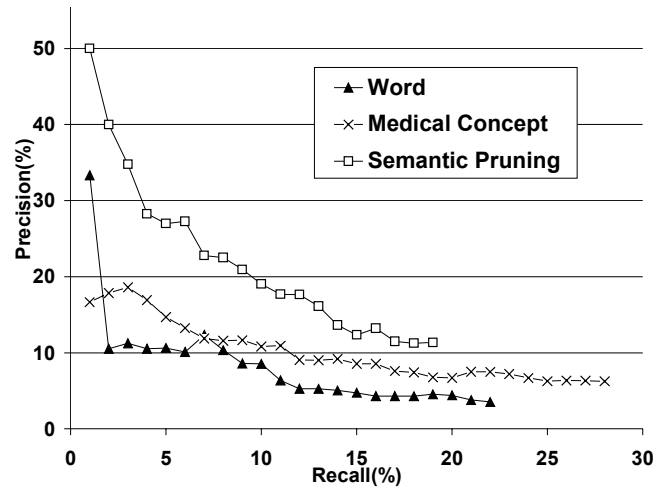
| Connection               | Rank      |                 |                  |
|--------------------------|-----------|-----------------|------------------|
|                          | Word      | Medical Concept | Semantic Pruning |
| Personality Type         | 79        | 61              | <b>23</b>        |
| Vascular tone            | <b>46</b> | 65              | 81               |
| Calcium channel blockers | 68        | 28              | <b>10</b>        |
| Cortical depression      | 93        | 145             | <b>45</b>        |
| Epilepsy                 | 56        | 19              | <b>8</b>         |
| Serotonin                | 14        | 12              | <b>5</b>         |
| Platelet activity        | 5         | 7               | <b>2</b>         |
| Inflammation             | 822       | 402             | <b>170</b>       |
| Prostaglandins           | 382       | 100             | <b>42</b>        |
| Substance P              | 172       | <b>38</b>       | 44               |
| Brain hypoxia            | <b>6</b>  | 26              | 19               |

**Table 3 The semantically pruned medical concepts were more highly ranked than the word representation or the complete set of medical words (best rank indicated in bold).**

Rather than calculating precision for each known connection and then averaging the results, we calculated precision based on any of the connections. That is we ranked each word (concept) in descending order of document frequency, then used any B-term or synonymous B-term word (concept) from table 2 to indicate relevance. Our motivation for this approach was that some of the connections, such as *inflammation* and *substance P* had very few relevant documents.

Improving the semantic quality of features used to represent the title corresponded to an improvement in precision. The average precision for word, concept and

semantically pruned concept representations was 8.3, 9.8 and 22.3 percent respectively. The interpolated precision for the semantic pruning representation was consistently higher than either the word or the complete set of medical concepts.



**Figure 2 Interpolated precision and recall for the eleven physiological linkages associated with migraines. The average interpolated was 8.3, 9.8 and 22.3% using word, medical concept and semantically pruned medical concept features respectively.**

Recall performance did not correlate directly with the semantic quality of features. The word representation identified 83 of the 366 connections while concepts identified 110 connections improving recall from 22.7 to 30.1 percent. Although the semantically pruned features were more highly ranked, fewer connections were identified (71) resulting in 19.4 percent recall.

In addition to precision and recall, we measured the dimensionality required to represent the problem space at each semantic level (see table 4). A medical concept representation required 50.7 percent fewer features than words. The semantically constrained feature set required 61% fewer features than concepts and 80.9% less than the word representation. This drastic dimensionality reduction is critical if text mining processes the MEDLINE database, which already holds over eleven million citations.

|  | Word  | Medical Concept | Semantic Pruning |
|--|-------|-----------------|------------------|
| <b>Total terms to represent MEDLINE titles</b> |       |                 |                  |
| Distinct terms                                 | 2732  | 1811            | 618              |
| Total terms                                    | 10827 | 5330            | 2067             |
| <b>Average terms per citation</b>              |       |                 |                  |
| Distinct terms                                 | 1.1   | 0.70            | 0.24             |
| Total terms                                    | 4.2   | 2.1             | 0.80             |

**Table 4 Each semantic level requires a different number of features to represent the 2,571 MEDLINE citations containing the word *migraine* (excluding 448 stopwords).**

## Future Work

We quantified the effect of alternative semantic pruning with respect to identifying known C-B connections indicating that magnesium would be effective in treating migraines. We plan to extend the medical concept and semantic pruning approaches to link the physiological conditions associated with a medical problem to an effective treatment (i.e., the B-A connections). We are also planning to repeat these experiments on other connections identified by Swanson and Smalheiser (Swanson and Smalheiser 1997).

The semantically pruned concept representation drastically reduced the number of features required to represent this problem space. While this has speed implications, it also means that representing other citation text becomes feasible. We plan to extend the current 'title only' strategy to other parts of the citation, such as the abstract and the entire document. Our empirical work with 91 biomedical articles, indicates 76 features will be required per citation if represented as words, compared to eight when using medical concepts when both the title and abstract are used (Blake and Pratt, 2001). Our automated concept and semantically pruned concept approaches appear well suited to problems in the medical domain where large quantities of text are available.

## Conclusion

Our experiment shows that increasing the semantic quality of features used to represent text improves system performance with respect to automatically identifying implicit connections from biomedical literature. Accurately identifying these connections is an important first step towards building tools that will provide researchers with an answer to the question: "What new promising treatment strategies should we investigate?" Unlike previous approaches that require manual ad-hoc pruning, our knowledge-based approach automatically generates the set of physiological characteristics associated with a medical condition.

Increasing the semantic quality of automatically generated B-terms from no semantics (words), to medical concepts and finally to semantically pruned concepts resulted in precision of 8.3 to 9.8 and 22.3 percent respectively. Changes in recall were not as large and did not correspond to the increase in semantic quality. Concepts had the highest recall of 30.1%, followed by words (22.3%) and then semantically pruned concepts (19.4).

The rank of semantically pruned features was higher than either the word or complete set of medical concepts in nine of the eleven connections associated with migraines. If researchers only consider the top fifty ranked terms then the word representation would only have identified three of the eleven connections as compared with six for the medical concept.

In addition to improving precision, the semantically pruned feature space required 80.9 percent fewer features than the word representation. This dramatic dimensionality reduction suggests that this approach would be suitable for large medical corpus.

The results indicate that a knowledge-based approach, which incorporates semantic pruning, is effective in automatically identifying new promising treatments from the medical literature.

## References

- Blake, C. and Pratt, W. 2001. Better rules, fewer features: A semantic approach to selecting features from text. In *Proceedings of the IEEE Data Mining Conference*, San Jose, California, IEEE Press:59-66.
- Lindsay, R. K. and Gordon, M. D. 1999. Literature-Based Discovery by Lexical Statistics. *Journal of the American Society for Information Science* 50(7): 574-587.
- Marukawa, H., Shimomura, T., and Takahashi, K. 1996. Salivary substance P, 5-hydroxytryptamine, and gamma-aminobutyric acid levels in migraine and tension-type headache. *Headache* 36(2): 100-104.
- May, A. and Goadsby, P. 2001. Substance P receptor antagonists in the therapy of migraine. *Expert Opin Investig Drugs* 10(4): 673-678.
- Nakano, T., T. Shimomura, K. Takahashi and S. Ikawa 1993. Platelet substance P and 5-hydroxytryptamine in migraine and tension-type headache. *Headache* 33(10): 528-532.
- NLM 2000. The SPECIALIST Lexicon. Available at: [www.nlm.nih.gov/pubs/factsheets/umlslex.html](http://www.nlm.nih.gov/pubs/factsheets/umlslex.html)
- NLM 2001. National Library of Medicine. Available at: [www.nlm.nih.gov](http://www.nlm.nih.gov)
- Sanderson, M. 1999. Stop word list. Available at: [http://www.dcs.gla.ac.uk/idom/ir\\_resources/linguistic\\_uts/](http://www.dcs.gla.ac.uk/idom/ir_resources/linguistic_uts/ils/)
- Swanson, D. R. 1988. Migraine and magnesium: eleven neglected connections. *Perspect. Biol. Med.* 31: 526-557.
- Swanson, D. R. and Smalheiser, N. R. 1997. An interactive system for finding complementary literatures: a stimulus to scientific discovery. *Artificial Intelligence* 91: 183-203.
- Weeber, M., Klein, H., Aronson, A.R., Mork, J.G, Jongvan den Berg, L., Vos, R. 2000. Text-Based Discovery in Biomedicine: The Architecture of the DAD-system. In *Proceedings of the American Medical Informatics Association Symposium*, Los Angeles: 903-7.