Analysis of Semantic Classes in Medical Text for Question Answering

Yun Niu and **Graeme Hirst**
Department of Computer Science
University of Toronto
Toronto, Ontario M5S 3G4
Canada
yun@cs.toronto.edu, gh@cs.toronto.edu

Abstract

To answer questions from clinical-evidence texts, we identify occurrences of the semantic classes — disease, medication, patient outcome — that are candidate elements of the answer, and the relations among them. Additionally, we determine whether an outcome is positive or negative.

1 Motivation

The published medical literature is an important source to help clinicians make decisions in patient treatment (Sackett and Straus, 1998; Straus and Sackett, 1999). Clinicians often need to consult literature on the latest information in patient care, such as side effects of a medication, symptoms of a disease, or time constraints in the use of a medication. For example:[1]

> **Q:** In a patient with a suspected MI does thrombolysis decrease the risk of death if it is administered 10 hours after the onset of chest pain?

The answer to the question can be found in Clinical Evidence (CE) (Barton, 2002), a regularly updated publication that reviews and consolidates experimental results for clinical problems:

> **A:** Systematic reviews of RCTs have found that prompt thrombolytic treatment (within 6 hours and perhaps up to 12 hours and longer after the onset of symptoms) reduces mortality in people with AMI and ST elevation or bundle branch block on their presenting ECG.

The goal of the EpoCare project ("Evidence at Point of Care") at the University of Toronto is to develop methods for answering questions automatically with CE as the source text. (We do not look at primary medical research text.) Currently, the system accepts keyword queries in PICO format (Sackett et al., 2000). In this format, a clinical question is represented by a set of four fields that correspond to the basic elements of the question:

> **P:** *a description of the patient (or the problem)*;
> **I:** *an intervention*;
> **C:** *a comparison or control intervention (may be omitted)*;
> **O:** *the clinical outcome*.

For example, the question shown above can be represented in PICO format as follows:

> **P:** myocardial infarction
> **I:** thrombolysis
> **C:** —
> **O:** mortality

Our work in the project is to extend the keyword retrieval to a system that can answer questions expressed in natural language.

In our earlier work (Niu et al., 2003), we showed that current technologies for factoid question answering (QA) are not adequate for clinical questions, whose answers must often be obtained by synthesizing relevant context. To adapt to this new characteristic of QA in the medical domain, we exploit *semantic classes* and *relations* between them in medical text. Semantic classes are important for our task because the information contained in them is often a good candidate for answering clinical questions. In the example above, PICO elements correspond to three semantic classes: DISEASE *(medical problem of the patient)*, INTERVENTION *(medication applied to the disease)* and the CLINICAL OUTCOME. They together constitute a SCENARIO of treatment. Similarly, a diagnosis scenario often includes SYMPTOMS, TESTING PROCEDURE, and HYPOTHESIZED DISEASES. To understand the semantics of medical text and find answers to clinical questions, we need to know how these classes relate to each other in a specific scenario. For example, is

[1] All the examples in this paper are taken from a collection of questions that arose over a two-week period in August 2001 in a clinical teaching unit at the University of Toronto.

this medication a special type of another one; is this medication applied to this disease? These are the kind of relations that we are interested in. In this work, we use a cue-word–based approach to identify semantic classes in the treatment scenario and analyze the relations between them. We also apply an automatic classification process to determine the polarity of an outcome, as it is important in answering clinical questions.

2 Identifying Semantic Classes in Medical Text

2.1 Diseases and Medications

The identification of named entities (NEs) in the biomedical area, such as PROTEINS and CELLS, has been extensively explored; e.g., Lee et al. (2003), Shen et al. (2003). However, we are not aware of any satisfactory solution that focuses on the recognition of semantic classes such as MEDICATION and DISEASE. To straightforwardly identify DISEASE and MEDICATION in the text, we use the knowledge base Unified Medical Language System (UMLS) (Lindberg et al., 1993) and the software MetaMap (Aronson, 2001).

UMLS contains three knowledge sources: the Metathesaurus, the Semantic Network, and the Specialist Lexicon. Given an input sentence, MetaMap separates it into phrases, identifies the medical concepts embedded in the phrases, and assigns proper semantic categories to them according to the knowledge in UMLS. For example, for the phrase *immediate systemic anticoagulants*, MetaMap identifies *immediate* as a TEMPORAL CONCEPT, *systemic* as a FUNCTIONAL CONCEPT, and *anticoagulants* as a PHARMACOLOGIC SUBSTANCE. More than one semantic category in UMLS may correspond to MEDICATION or DISEASE. For example, either a PHARMACOLOGIC SUBSTANCE or a THERAPEUTIC OR PREVENTIVE PROCEDURE can be a MEDICATION; either a DISEASE OR SYNDROME or a PATHOLOGIC FUNCTION can be a DISEASE.

We use some training text to find the mapping between UMLS categories and the two semantic classes in the treatment scenario. The training text was tagged for us by a clinician to mark DISEASE and MEDICATION. It was also processed by MetaMap. After that, the annotated text was compared with the output of MetaMap to find the corresponding UMLS categories. Medical text containing these categories can then be identified as either MEDICATION or DISEASE. In the example above, *anticoagulants* will be taken as a MEDICATION. The problem of identification of medical terminology is still a big challenge in this area. MetaMap does not provide a full solution to it. For cases in which the output of MetaMap is not consistent with the judgment of the clinician who annotated our text, our decisions rely on the latter.

2.2 Clinical Outcome

The task of identifying clinical outcomes is more complicated. Outcomes are often not just noun phrases; instead, they usually are expressed in complex syntactic structures. The following are some examples:

(1) Thrombolysis *reduces the risk of dependency*, but *increases the risk of death*.

(2) *The median proportion of symptom free days improved more* with salmeterol than with placebo.

In our analysis of the text, we found another type of outcome which is also very important: the outcome of clinical trials:

(3) Several small comparative RCTs [randomized clinical trials] have found sodium cromoglicate to be *less effective* than inhaled corticosteroids *in improving symptoms and lung function*.

(4) In the systematic review of calcium channel antagonists, indirect and limited comparisons of intravenous versus oral administration found *no significant difference in adverse events*.

We treat these as a special type of clinical outcome. For convenience, we refer to them as "results" in the following description when necessary. A "result" might contain a clinical outcome within it, as results often involve a comparison of the effects of two (or more) interventions on a disease.

In medical text, the appearance of some words is found often to be a signal of the occurrence of an outcome, and usually several words signal the occurrence of one single outcome. The combination approach that we applied for identifying outcomes is based on this observation. Our approach does not extract the whole outcome at once. Instead, it tries to identify the different parts of an outcome that may be scattered in the sentence, and then combines them to form the complete outcome.

2.2.1 Related work

Rule-based methods and machine-learning approaches have been used for similar problems.

Gildea and Jurafsky (2002) used a supervised learning method to learn both the identifier of the semantic roles defined in FrameNet such as theme, target, goal, and the boundaries of the roles (Baker et al., 2003). A set of features were learned from a large training set, and then applied to the unseen data to detect the roles. The performance of the system was quite good. However, it requires a large training set for related roles, which is not available in many tasks, including tasks in the medical area.

Rule-based methods are explored in information extraction (IE) to identify roles to fill slots in some pre-defined templates (Català et al., 2003). The rules are represented by a set of patterns, and template role identification is usually conducted by pattern matching. Slots indicating roles are embedded in these patterns. Text that satisfies the constraints of a pattern will be identified, and the contents corresponding to the slots are extracted. This approach has been proved to be effective in many IE tasks. However, pattern construction is very time-consuming, especially for complicated phrasings. In order to select the roles and only the roles, their expression has to be customized specifically in patterns. This results in increasing difficulties in pattern construction, and reduces the coverage of the patterns.

2.2.2 A combination approach

Different pieces of an outcome are identified by various cue words. Each occurrence of a cue word suggests a portion of the expression of the outcome. Detecting all of them will increase the chance of obtaining the complete outcome. Also, different occurrences of cue words provide more evidence of the existence of an outcome.

The first step of the combination approach is to collect the cue words. Two sections of CE (stroke management, asthma in children) were analyzed for detection of outcome. The text was annotated by a clinician in the EpoCare project. About two-thirds of each section (267 sentences in total) was taken as the analysis examples for collecting the cue words, and the rest (156 sentences) as the test set. Some words we found in the analysis are the following:

Nouns: *death, benefit, dependency, outcome, evidence, harm, difference*.

Verbs: *improve, reduce, prevent, produce, increase*.

Adjectives: *beneficial, harmful, negative, adverse, superior*.

After the cue words are identified, the next question is what portion of text each cue word suggests as the outcome, which determines the boundary of the outcome. The text was pre-processed by the Apple Pie parser (Sekine, 1997) to obtain the part-of-speech and phrase information. We found that for the noun cues, the noun phrase that contains the noun will be part of the outcome. For the verb cue words, the verb and its object together constitute one portion of the outcome. For the adjective cue words, often the corresponding adjective phrase or the noun phrase belongs to the outcome. Cue words for the results of clinical trials are processed in a slightly different way. For example, for *difference* and *superior*, any immediately following prepositional phrase is also included in the results of the trial.

Our approach does not rely on specific patterns, it is more flexible than pattern-matching techniques in IE systems, and it does not need a large training set. A limitation of this approach is that some connections between different portions of an outcome may be missing.

2.2.3 Evaluation and analysis of results

We evaluated the cue word method of detecting the outcome on the remaining one-third of the sections of CE. (The test set is rather small because of the difficulty in obtaining the annotations.) The outcome detection task was broken into two sub-tasks, each evaluated separately: to identify the outcome itself and to determine its textual boundary. The result of identification is shown in Table 1. Eighty-one sentences in the test set contain either an outcome or result, which is 52% of all the test sentences. This was taken as the baseline of the evaluation: taking all sentences in the test set as positive (i.e., containing an outcome or result). By contrast, the accuracy of the combination approach is 83%.

There are two main reasons why some outcomes were not identified. One is that some outcomes do not have any cue word:

(5) *Gastrointestinal symptoms* and *headaches* have been reported with both montelukast and zafirlukast.

The other reason is that although some outcomes contained words that might be regarded as cue words, we did not include them in our set; for example, *fewer* and *higher*. Adjectives were found to have the most irregular usages. It is normal for them to modify both medications and outcomes, as shown in the following examples:

(6) ... children receiving *higher* dose inhaled corticosteroids ...

Table 1: Results of identifying outcomes in CE

Method	Correct	False Positives	False Negatives	Precision%	Recall%	Accuracy%
Baseline	81	75	0	52 (81/156)	100	52
Combination approach	67	14	14	83 (67/81)	83	82

Table 2: Results of boundary detection of correctly identified outcomes in CE. A: Identified fragments; B: true boundary.

Type of Overlap	Number	Percentage
Exact match	26	39
A entirely within B	19	28
B entirely within A	13	19
Each partially within the other	8	12
No match	1	1

(7) ... mean morning PEFR was 4% *higher* in the salmeterol group.

Other adjectives such as *less*, *more*, *lower*, *shorter*, *longer*, and *different* have similar problems. If they are taken as identifiers of outcomes then some false positives are very likely to be generated. However, if they are excluded, some true outcomes will be missed. There were 14 samples of false positives. The main cause was sentences containing cue words that did not have any useful information:

(8) We found that the balance between *benefits* and *harms* has not been clearly established for the evacuation of supratentorial haematomas.

(9) The third systematic review did not evaluate these *adverse outcomes*.

Table 2 shows the result of boundary detection for those outcomes that were correctly identified. The true boundary is the boundary of an outcome that was annotated manually. The *no match* case means that there is a true outcome in the sentence but the program missed the correct portions of text and marked some other portions as the outcome. The program identified 39% of the boundaries exactly the same as the true boundaries. In 19% of the samples, the true boundaries were entirely within the identified fragments. The spurious text in them (the text that was not in the true boundary) was found to be small in many cases, both in terms of number of words and in terms of the importance of the content. The average number of words correctly identified was 7 for each outcome and the number of spurious words was 3.4. The most frequent content in the spurious text was the medication applied to obtain the outcome. In the following examples, text in "⟨⟩" is the outcome (result) identified automatically, and text in "{}" is spurious.

(10) The RCTs found ⟨no significant adverse effects {associated with salmeterol}⟩.

(11) The second RCT ... also found ⟨no significant difference in mortality at 12 weeks {with lubeluzole versus placebo}⟩ ...

Again, adjectives are most problematic. Even when a true adjective identifier is found, the boundary of the outcome is hard to determine by an unsupervised approach because of the variations in the expression. In the following examples, the true boundaries of outcomes are indicated by "[ ]", adjectives are highlighted.

(12) Nebulised ... , but [⟨serious **adverse** effects⟩ are rare].

(13) Small RCTs ... found that [... was ⟨**effective**⟩, with ...].

The correctness of the output of the parser also had an important impact on the performance, as shown in the following example:

(14) RCTs found no evidence that lubeluzole improved clinical outcomes in people with acute ischaemic stroke.
(S ... (NPL (DT that) (JJ lubeluzole) (JJ improved) (JJ clinical) (NNS outcomes)) ...)

In this parse, the verb *improve* was incorrectly assigned to be an adjective in a noun phrase. Thus *improve* as a verb cue word was missed in identifying the outcome. However, another cue word *outcomes* was matched, so the whole noun phrase of *outcomes* was identified as the outcome. On the one hand, the example shows that the wrong parsing output

directly affects the identification process. On the other hand, it also shows that missing one cue word in identifying the outcome can be corrected by the occurrence of other cue words in the combination approach.

3 Analysis of Relations

Recognition of individual semantic classes is not enough for text understanding; we also need to know how different entities in the same semantic class are connected, as well as what relations hold between different classes. Currently, all these relations are considered at the sentence level.

3.1 Relations within the same semantic class

Relations between different medications are the focus of this sub-section, as a sentence often mentioned more than one medication. Relations between diseases can be analyzed in a similar way, although they occur much less often than medications. Text from CE was analyzed manually to understand what relations are often involved and how they are represented. The text for the analysis is the same as in the class-identification task discussed above. As with classes themselves, it was found that these relations can be identified by a group of cue words or symbols. For example, the word *plus* refers to the COMBINATION of two or more medications, the word *or*, as well as a comma, often suggests the ALTERNATIVE relation, and the word *versus* (or *v*) usually implies a COMPARISON relation, as shown in the following examples:

(15) The combination of aspirin *plus* streptokinase significantly increased mortality at 3 months.

(16) RCTs found no evidence that calcium channel antagonists, lubeluzole, aminobutyric acid (GABA) agonists, glycine antagonists, *or* N-methyl-D-aspartate (NMDA) antagonists improve clinical outcomes in people with acute ischaemic stroke.

(17) One systematic review found no short or long term improvement in acute ischaemic stroke with immediate systemic anticoagulants (unfractionated heparin, low molecular weight heparin, heparinoids, *or* specific thrombin inhibitors) *versus* usual care without systemic anticoagulants.

It is worth noting that in CE, the experimental conditions are often explained in the description of the outcomes, for example:

(18) ... receiving higher dose inhaled corticosteroids *(3.6cm, 95% CI 3.0 to 4.2 with double dose beclometasone v 5.1cm, 95% CI 4.5 to 5.7 with salmeterol v 4.5cm, 95% CI 3.8 to 5.2 with placebo)*.

(19) It found that ... oral theophylline ... versus placebo increased the mean number of symptom free days *(63% with theophylline v 42% with placebo; P=0.02)*.

(20) Studies of ... inhaled steroid *(see salmeterol v high dose inhaled corticosteroids under adult asthma)*.

These descriptions are usually in parentheses. They are often phrases and even just fragments of strings that are not represented in a manner that is uniform with the other parts of the sentence. Their behavior is more difficult to capture and therefore the relations among the concepts in these descriptions are more difficult to identify. Because they usually are examples and data, omission of them will not affect the understanding of the whole sentence in most cases.

Six common relations and their cue words were found in the text which are shown in Table 3. Cue words and symbols between medical concepts were first collected from the training text. Then the relations they signal were analyzed. Some cue words are ambiguous, for example, *or*, *and*, and *with*. *Or* could also suggest a comparison relation although most of the time it means alternative, *and* could represent an alternative relation, and *with* could be a specification relation. It is interesting to find that *and* in the text when it connects two medications often suggests an alternative relation rather than a combination relation (e.g., the second *and* in example 5). Also, compared with *versus*, *plus*, etc., *and* and *with* are weak cues as most of their appearances in the text do not suggest a relation between two medications.

On the basis of this analysis, an automatic relation analysis process was applied to the test set, which was the same as in outcome identification. The test process was divided into two parts: one took parenthetical descriptions into account (case 1) and the other one did not (case 2). In the evaluation, for sentences that contain at least two medications, "correct" means that the relation that holds between the medications is correctly identified. We do not evaluate the relation between any two medications in a sentence; instead, we only considered two medications that are related to each other by a cue word or symbol (including those connected by cue words

Table 3: Cue words/symbols for relations between medications

Relation(s)	Cue Words/Symbols
comparison	superior to, more than, versus, or, compare with, between ... and ...
alternative	or, ",", and
combination	plus, add to, addition of ... to ..., combined use of, and, with, "("
specification	with, "("
substitute	substitute, substituted for
preference	rather than

Table 4: Results of relation analysis

	Correct	Wrong	Missing	False Positive
Case 1	49	7	10	9
Case 2	48	7	3	6

other than the set collected from the training text). The results of the two cases are shown in Table 4.

Most errors are because of the weak indicators *with* and *and*. As in the outcome identification task, both the training and test sets are rather small, as no standard annotated text is available.

Some of the surface relationships in Table 3 reflect deeper relationships of the semantic classes. For example, COMPARISON, ALTERNATIVE, and PREFERENCE imply that the two (or more) medications have some common effects on the disease(s) they are applied to. The SPECIFICATION relation, on the other hand, suggests a hierarchical relation between the first medication and the following ones, in which the first medication is a higher-level concept and the following medications are at a lower level. For example, in example 17 above, *systemic anticoagulants* is a higher-level concept, *unfractionated heparin, low molecular weight heparin*, etc., are examples of it that lie at a lower level.

3.2 Relations between different semantic classes

In a specific domain such as medicine, some default relations often hold between semantic classes. For example, a CAUSE–EFFECT relation is strongly embedded in the three semantic classes appearing in a sentence of the form: "*medication ... disease ... outcome*", even if not in this exact order. This default relation helps the relation analysis because in most cases we do not need to depend on the text between the classes to understand the whole sentence. For instance, the CAUSE–EFFECT relation is very likely to express the idea that applying the intervention on the disease will have the outcome. This is another reason that semantic classes are important, especially in a specific domain.

4 The polarity of outcomes

Most clinical outcomes and the results of clinical trials are either positive or negative:

(21) *Positive*: Thrombolysis reduced the risk of death or dependency at the end of the studies.

(22) *Negative*: In the systematic review, thrombolysis increased fatal intracranial haemorrhage compared with placebo.

Polarity information is useful for several reasons. First of all, it can filter out positive outcomes if the question is about the negative aspects of a medication. Secondly, negative outcomes may be crucial even if the question does not explicitly ask about them. Finally, from the number of positive or negative descriptions of the outcome of a medication applying to a disease, clinicians can form a general idea about how "good" the medication is. As a first step in understanding opposing relations between scenarios in medical text, the polarity of outcomes was determined by an automatic classification process.

We use support vector machines (SVMs) to distinguish positive outcomes from negative ones. SVMs have been shown to be efficient in text classification tasks (Joachims, 1998). Given a training sample, the SVM finds a hyperplane with the maximal margin of separation between the two classes. The classification is then just to determine which side of the hyperplane the test sample lies in. We used the SVM^{light} package (Joachims, 2002) in our experiment.

4.1 Training and test examples

The training and test sets were built by collecting sentences from different sections in CE; 772 sentences were used, 500 for training (300 positive, 200 negative), and 272 for testing (95 positive, 177 negative). All examples were labeled manually.

4.2 Evaluation

The classification used four different sets of features. The first feature set includes every unigram that appears at least three times in the whole training set. To improve the performance by attenuating

the sparse data problem, in the second feature set, all names of diseases were replaced by the same tag *disease*. This was done by pre-processing the text using MetaMap to identify all diseases in both the training and the test examples. Then the identified diseases were replaced by the *disease* tag automatically. As medications often are not mentioned in outcomes, they were not generalized in this manner.

The third feature set represents changes described in outcomes. Our observation is that outcomes often involve the change in a clinical value. For example, after a medication was applied to a disease, something was *increased* (*enhanced, more,* ...) or *decreased* (*reduced, less,* ...). Thus the polarity of an outcome is often determined by how change happens: if a bad thing (e.g., mortality) is reduced then it is a positive outcome; if the bad thing is increased, then the outcome is negative. We try to capture this observation by adding context features to the feature set. The way they were added is similar to incorporating the negation effect described by Pang et al. (2002). But instead of just finding a "negation word" (*not, isn't, didn't,* etc.), we need to find two groups of words: those indicating *more* and those indicating *less*. In the training text, we found 9 words in the first group and 7 words in the second group. When pre-processing text for classification, following the method of Pang et al., we attached the tag *_MORE* to all words between the *more*-words and the following punctuation mark, and the tag *_LESS* to the words after the *less*-words.

The fourth feature set is the combination of the effects of feature set two and three. In representing each sentence by a feature vector, we tested both presence (feature appears or not) and frequency (count the number of occurrences of the feature in the sentence).

The accuracy of the classification is shown in Table 5. The baseline is to assign a random class (here we use negative, as they are more frequent in the test set) to all test samples.

The *presence* of features performs better than *frequency* of features in general. Using a more general category instead of specific diseases has a positive effect on the presence-based classification. We speculate that the effect of this generalization will be bigger if a larger test set were used. Pang et al. (2002) did not compare the result of using and not using the negation context effect, so it is not clear how much it improved their result. In our task, it is clear that the *_MORE*/*_LESS* feature has a significant effect on the performance, especially for the *frequency* features.

Table 5: Results of outcome polarity classification

Features	Presence (%)	Frequency (%)
Baseline	65.07	65.07
Original unigrams	88.97	87.87
Unigrams with *disease*	90.07	88.24
Unigrams with *_MORE/_LESS* tag	91.54	91.91
Unigrams with *disease* and *_MORE/_LESS* tag	92.65	92.28

5 Conclusion

We have described our work in medical text analysis by identifying semantic classes and the relations between them. Our work suggests that semantic classes in medical scenarios play an important role in understanding medical text. The scenario view may be extended to a framework that acts as a guideline for further semantic analysis.

Semantic classes and their relations have direct applications in medical question answering and query refinement in information retrieval. In question answering, the question and answer candidates will contain some semantic classes. After identifying them on both sides, the question can be compared with the answer to find whether there is a match. In information retrieval, relations between semantic classes can be added to the index. If the query posed by the user is too general, the system will ask the user to refine the query by adding more concepts and even relations so that it will be more pertinent according to the content of the source. For example, a user may search for a document describing the comparison of aspirin and placebo. Instead of just using *aspirin* and *placebo* as the query terms, the user can specify the comparison relation as well in the query.

We will continue working on the second level of the semantic analysis, to explore the relations on the scenario level. A complete scenario contains all three semantic classes. One scenario may be the explanation or justification of the previous scenario(s), or contradictory to the previous scenario(s). Detecting these relationships will be of great help for understanding-based tasks, such as context-related question answering, topic-related summarization, etc. As different scenarios might not be adjacent to each other in the texts, classical rhetorical analysis cannot provide a complete solution for this problem.

Acknowledgements

The EpoCare project is supported by grants from Bell University Laboratories at the University of Toronto. Our work is also supported by a grant from the Natural Sciences and Engineering Research Council of Canada and an Ontario Graduate Scholarship. We are grateful to Sharon Straus, MD, and other members of the EpoCare project for discussion and assistance.

References

Alan R. Aronson. 2001. Effective mapping of biomedical text to the UMLS metathesaurus: The MetaMap program. In *Proceedings of American Medical Informatics Association Symposium*, pages 17–21.

Collin F. Baker, Charles J. Fillmore, and Beau Cronin. 2003. The structure of the Framenet database. *International Journal of Lexicography*, 16(3):281–296.

Stuart Barton. 2002. *Clinical Evidence*. BMJ Publishing Group, London.

Neus Català, Núria Castell, and Mario Martín. 2003. A portable method for acquiring information extraction patterns without annotated corpora. *Natural Language Engineering*, 9(2):151–179.

Daniel Gildea and Daniel Jurafsky. 2002. Automatic labeling of semantic roles. *Computational Linguistics*, 28(3):245–288.

Thorsten Joachims. 1998. Text categorization with support vector machines: Learning with many relevant features. In *Proceedings of the European Conference on Machine Learning (ECML)*, pages 137–142.

Thorsten Joachims. 2002. SVM^{light} homepage. In *http://svmlight.joachims.org/*.

Ki-Joong Lee, Young-Sook Hwang, and Hae-Chang Rim. 2003. Two-phase biomedical NE recognition based on SVMs. In *Proceedings of 41st annual meeting of the Association for Computational Linguistics, Workshop on Natural Language Processing in Biomedicine*, pages 33–40.

Donald A. B. Lindberg, Betsy L. Humphreys, and Alexa. T. McCray. 1993. The Unified Medical Language System. *Methods of Information in Medicine*, 32(4):281–291.

Yun Niu, Graeme Hirst, Gregory McArthur, and Patricia Rodriguez-Gianolli. 2003. Answering clincal questions with role identification. In *Proceedings of 41st annual meeting of the Association for Computational Linguistics, Workshop on Natural Language Processing in Biomedicine*, pages 73–80.

Bo Pang, Lillian Le, and Shivakumar Vaithyanathan. 2002. Thumbs up? Sentiment classification using machine learning techniques. In *Proceedings of 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 79–86.

David L. Sackett and Sharon E. Straus. 1998. Finding and applying evidence during clinical rounds: The "evidence cart". *Journal of the American Medical Association*, 280(15):1336–1338.

David L. Sackett, Sharon E. Straus, W. Scott Richardson, William Rosenberg, and R. Brian Haynes. 2000. *Evidence-Based Medicine: How to Practice and Teach EBM*. Harcourt Publishers Limited, Edinburgh.

Satoshi Sekine. 1997. Apple pie parser homepage. In *http://nlp.cs.nyu.edu/app/*.

Dan Shen, Jie Zhang, Guodong Zhou, Jian Su, and Chew-Lim Tan. 2003. Effective adaptation of hidden Markov model–based named entity recognizer for biomedical domain. In *Proceedings of 41st annual meeting of the Association for Computational Linguistics, Workshop on Natural Language Processing in Biomedicine*, pages 49–56.

Sharon E. Straus and David L. Sackett. 1999. Bringing evidence to the point of care. *Journal of the American Medical Association*, 281:1171–1172.