Fast Tagging of Medical Terms in Legal Text

Christopher Dozier
Thomson
610 Opperman Drive
Eagan MN 55123
chris.dozier@thomson.com

Ravi Kondadadi
Thomson
610 Opperman Drive
Eagan MN 55123

Khalid Al-Kofahi
Thomson
610 Opperman Drive
Eagan MN 55123

Mark Chaudhary
Thomson
610 Opperman Drive
Eagan MN 55123

Xi Guo
Thomson
610 Opperman Drive
Eagan MN 55123

ABSTRACT
Medical terms occur across a wide variety of legal, medical, and news corpora. Documents containing these terms are of particular interest to legal professionals operating in such fields as medical malpractice, personal injury, and product liability. This paper describes a novel method of tagging medical terms in legal, medical, and news text that is very fast and also has high recall and precision. To date, most research in medical term spotting has been confined to medical text and has approached the problem by extracting noun phrases from sentences and mapping them to a list of medical concepts via a fuzzy lookup. The medical term tagging described in this paper relies on a fast finite state machine that finds within sentences the longest contiguous sets of words associated with medical terms in a medical term authority file, converts word sets into medical term hash keys, and looks up medical concept ids associated with the hash keys. Additionally, our system relies on a probabilistic term classifier that uses local context to disambiguate terms being used in a medical sense from terms being used in a non-medical sense. Our method is two orders of magnitude faster than an approach based on noun phrase extraction and has better precision and recall for terms pertaining to injuries, diseases, drugs, medical procedures, and medical devices. The methods presented here have been implemented and are the core engines for a Thomson West product called the Medical Litigator. Thus far, we used this method to tag over 165 million instances of some 164,000 unique medical concepts in over 100 million documents from case law, jury verdicts and settlements documents, briefs, jurisprudence and other types of analytical documents, legal newspapers, general newspapers, medical encyclopedia and an array of biomedical journals.

1. INTRODUCTION
Medical terminology is important to legal practitioners who specialize in areas of law such as medical malpractice, product liability, intellectual property, and personal injury among others. Attorneys specialized in these areas perform legal research to, for example, (1) learn more about a medical concept, (2) find relevant litigation pertaining to a medical concept in general or within a specific legal or factual context, and (3) determine the legal liability of involved parties (e.g., should a doctor have known about a drug-drug interaction because it has been published in specialized proceedings). To satisfy these needs, we developed a novel medical term tagging method that tags terms pertaining to injuries, diseases, drugs, medical procedures, and medical devices over a large and diverse set of documents. Our medical term tagger is very fast and has high recall and precision. Thus far, we used this method to tag over 165 million instances of some 164,000 unique medical concepts in over 100 million documents from case law, jury verdicts and settlements documents, briefs, jurisprudence and other types of analytical documents, legal newspapers, general newspapers, medical encyclopedia and an array of biomedical journals.

Categories and Subject Descriptors
H.3.1 [Information Storage and Retrieval]: Content Analysis and Indexing - Dictionaries, Indexing methods, Linguistic processing, Thesauruses; I.2.7 [Artificial Intelligence]: Natural Language Processing - Language models, Language parsing and understanding, Text analysis; J.3 [Life and Medical Sciences]: Medical information systems

General Terms
Algorithms, Performance, Design

Keywords
Medical litigation, medical terms, UMLS, MetaMap, medical text, context disambiguation

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

ICAI’07, June 4-8, 2007, Palo Alto, CA USA
Copyright 2007 ACM 978-1-59593-680-6 ...$5.00.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

ICAI’07, June 4-8, 2007, Palo Alto, CA USA
Copyright 2007 ACM 978-1-59593-680-6 ...$5.00.
Plaintiff’s decedent had peripheral vascular disease, diabetes, asthma, bronchitis, obstructive pulmonary disease and two myocardial infarctions. A cardiac catheterization revealed two areas of blockage in a single vessel. Defendant cardiologist attempted to perform a balloon angioplasty procedure, but was unable to reach the blockages due to twisting of the arteries.

Figure 1: An excerpt from a medical malpractice case, with the terms tagged by the Medical Litigator system highlighted.

with medical terms in an authority file, converts the word sets into medical term hash keys, and looks up medical concept ids associated with the hash keys. Additionally our system relies on a probabilistic term classifier that uses local context to disambiguate terms being used in a medical sense from terms being used in a non-medical sense.

We constructed a medical term authority file using two resources: The National Institute of Health’s (NIH) Unified Medical Language Systems (UMLS) \(^1\) Metathesaurus \(^2\) of medical terms, and the Red Book \(^3\) database, a product of Micromedix (a sister Thomson company).

The Metathesaurus was used as the source for diseases, injuries, medical devices and medical procedures. It gets its name from the fact that its vocabulary is built from the vocabularies of a number of source thesauri used in patient care, clinical, and health services research. The source thesauri for the Metathesaurus include SNOMED, ICD9, and MeSH.

The Red Book database was used as the source for drug terms (including such additional information as brand and generic names).

Our method is two orders of magnitude faster than a tagging approach based on noun phrase extraction and medical term matching using the NIHs MetaMap program \([4]\). In addition, our method has better recall and precision than MetaMap for the medical terms our method targets. Because of its speed, and accuracy, the medical term tagger has been put into production and is the underlying technology for the Medical Litigator, a Thomson-West product for attorneys who practice in areas of law relevant to medicine.

The paper is organized as follows. The following section describes the Medical Litigator application. Section three describes previous approaches to the problem of tagging medical terms in text. Section four describes our medical term tagging algorithms. Section five describes the performance of the system in terms of recall, precision and throughput. Section six discusses our research plans in this area.

2. MEDICAL TERM TAGGING

As discussed in the previous section, medical terminology is important to legal practitioners who specialize in areas of law such as medical malpractice, product liability, intellectual property and personal injury, among others. The “Medical Litigator” system is a product from Thomson-West that aims at providing those attorneys with better research tools. The system provides attorneys interested in learning about a medical term (e.g., a medical procedure or a disease) with direct access to medical encyclopedia from the search results. It also provides the legal researcher with a product that cuts across traditional content silos such as legal, news and medical databases.

The system is a significant improvement over free text search because medical terms are tagged as meta data and are indexed as such, therefore enabling users to efficiently execute queries containing medical and non-medical language (e.g., “heart attacks and car accidents”). To improve recall, the system expands user queries by conflating medical terms with their synonyms (e.g., “heart attack” and “myocardial infarction”). To improve precision, the system distinguishes between medical and non-medical usage of a medical term (e.g., “miscarriage”) by incorporating the local context. To maintain efficient search time, both term conflation and term disambiguation are done at tagging and indexing time, as opposed to runtime.

Thus far, the system has processed over 100 million documents, generated approximately 165 million tags representing about 164,000 unique medical terms. Figure 1 shows an excerpt from a medical malpractice case that has been processed by the Medical Litigator system. Notice that medical terms are hyperlinked. By clicking on a link, users can retrieve medical definitions and encyclopedia articles about a term as well as explore the term’s occurrences in the corpus.

Figure 2 shows the screen that Medical Litigator displays when a user views text and clicks on the term colonoscopy or any of its synonyms (endoscopic examination of colon, endoscopy of colon, lower gastrointestinal tract examination). On the right side of the screen, a medical encyclopedia page describing colonoscopy is displayed. On the left side, a panel is displayed that shows the counts of the concept colonoscopy in various tagged corpora.

3. PREVIOUS APPROACHES TO THE TAGGING PROBLEM

The problem of tagging medical terms listed in the UMLS is a subset of the larger problem of tagging terms in text that correspond to a list of normalized terms contained in an ontology. Previous approaches to this problem have included systems that parse noun phrases from sentences and map them to a list of medical concepts. Examples of such systems include MicroMesh \([8]\), Metaphrase \([13]\), KnowledgeMap \([7]\), PhraseX \([12]\), and MetaMap \([5]\).

The MetaMap program is freely downloadable from the UMLS web site and is specifically engineered to match noun phrases to UMLS concepts. For these reasons, we chose MetaMap as a baseline system against which to measure the performance of our system. MetaMap works by extracting noun phrases from text with a shallow parser and matching the noun phrases to UMLS concept terms using a statistical scoring function.

The MetaMap scoring function assigns a match score between a noun phrase and UMLS concept using four criteria: coverage, cohesiveness, centrality, and involvement. The coverage criteria measures how many of the noun phrase tokens match the candidate concepts tokens. The cohesiveness criteria measures how many contiguous tokens from the noun phrase match the concept phrase. The centrality

---


---
Figure 2: A display of the reference page for the term colonoscopy or any of its synonyms (endoscopic examination of colon, endoscopy of colon, lower gastrointestinal tract examination). The left pane shows the counts for the concept colonoscopy in various tagged corpora.

criteria checks whether the head noun of the noun phrase matches the head noun of the concept phrase. And the involvement criteria “combines coverage and cohesiveness and takes into account word order variation”.

The Metamap match score is given as a positive integer. Metamap gives higher scores to concepts that it considers better matches for a particular noun phrase submitted to it.

In [11], Jacquemin describes his term spotting system FASTR. FASTR is a system that finds in a given text collection occurrences of terms and term variants specified in a term list. FASTR addresses the same problem as our system. The input to the system is a list of terms and a text collection. The output of the system is a mapping of string segments in the collection to terms in the list. FASTR differs from our system in that it applies to a wider range of problems than just the medical term domain and in that it makes use of more complex and time consuming techniques such as shallow parsing. The unique property of our system is that it obtains high precision and recall using a high speed look up technique governed by the terms in the list and a few simple processing rules.

In [9], Hanisch et al. describe a system that identifies protein names using terms from the list and delimiter tokens to quickly identify protein terms in text.

In [14], Zou et al. describe a system called IndexFinder that is similar to our system in that it maps medical term words to numbers and creates order independent hash keys to look up UMLS medical terms. The difference between IndexFinder and our tagger is that we use stopping conditions to find boundaries of medical terms and search for the longest medical term within the boundary. Indexfinder searches for all permutations of discovered medical terms within a fixed text segment. So, IndexFinder would search for permutations of [bone, fracture, lung, cancer] over the text segment: “Patient had a bone fracture and lung cancer”. This would yield “bone cancer” as a matching concept. Like our system, IndexFinder has very fast throughput (43 kilobytes per second). Unlike our system, IndexFinder is targeted exclusively at medical text and does not include a term disambiguation component when tagging non-medical text. Unfortunately, the authors do not report the precision and recall of their approach, so we are unable to compare our two systems on medical text.

4. DESCRIPTION OF METHODS

Figure 3 is an illustration that shows the resources and the processing modules that comprise the medical term tagger, as well as their interdependencies. Closed boxes denote processing components, while open boxes denote resources (e.g., UMLS Metathesaurus). The dotted line separates those components needed at runtime from those needed during preprocessing. In this section, we describe the four main processing components of the system. Namely, we describe (1) the methods used to construct the authority file, (2) the methods used for labeling authority terms as ambiguous or unambiguous, (3) the term tagger, and (4) the term disambiguator. Notice that the first two components are run offline during preprocessing, while the later two are executed at runtime.
Figure 3: The figure illustrates the processing components and the resources comprising our tagging methods as well as their interdependencies. Closed boxes denote processing components and open boxes denote resources. The dotted line separates pre-processing components from those executed at runtime.

4.1 Creation of Medical Term Authority File

The Authority File contains five types of medical terms: injuries, diseases, medical procedures, medical devices, and drugs. It is constructed from the UMLS Metathesaurus and the Red Book drug reference database.

The UMLS Metathesaurus is constructed by the NIH and is the largest thesaurus in the biomedical domain. Medical terms in the UMLS are mapped to a concept number, classified by semantic type, and placed in hierarchical and non-hierarchical relationships with other medical concepts. The total UMLS thesaurus contains about 2,000,000 medical concepts and the terms in the thesaurus come from a diverse set of sub-thesauri including SNOMED and ICD09. UMLS is not designed to support a specific purpose; rather it is intended as a resource from which a subset of medical terms can be utilized for particular applications.

The RED BOOK Drug Reference contains information on over 200,000 prescription and over-the-counter drugs and bulk chemicals as well as brand-name and generic drug information.

For the purposes of our application, we excluded terms whose primary sense is not medical. For example, terms like “vacuum cleaner” and “barbed wire” are listed as medical devices in the Metathesaurus. While this is a valid sense, it is not the primary one for these terms. Such exclusion, while not necessary, reduces the number of terms requiring sense disambiguation without affecting the overall quality of the product. We constructed a rule-based classifier to identify and remove such terms from the authority. The classifier relies on three features: the string probability as estimated by a character-based language model, the length of the term, and the inverse document frequency (idf) of the words comprising the term. To ensure quality, in-house medical experts reviewed and verified the excluded term list.

Table 1: The table shows the IDs of 3 UMLS concepts and their associated strings.

<table>
<thead>
<tr>
<th>Concept id</th>
<th>Concept strings</th>
</tr>
</thead>
<tbody>
<tr>
<td>C0827051</td>
<td>Heart attack, myocardial infarction</td>
</tr>
<tr>
<td>C0242379</td>
<td>Cancer of lung, lung cancer, pulmonary cancer</td>
</tr>
<tr>
<td>C0345904</td>
<td>Cancer of liver, hepatic cancer, liver cancer</td>
</tr>
</tbody>
</table>

4.2 Classifying Term Ambiguity Level

Recall that some medical terms can have non-medical senses in non-medical text. For this purpose, we assigned each term in the authority one of three labels. Terms that are nearly always used in a medical sense regardless of their contexts are labeled “unambiguous”, terms that are used in a medical sense in non-medical text collections only some of the time are labeled “ambiguous”, and terms that are rarely used in a medical sense in non-medical collections are labeled “problematic”. An example of a term that is used only in a medical sense regardless of context is “myocardial infarction”. An example of a term that may or may not be
A feature based on the average number of WordNet \([11]\) senses per word in medical term. This feature is based on the idea that the more senses the words comprising a medical term have the more ambiguous the term is likely to be.

\[
Score = 0.5 \frac{P(M)}{P(N)} + 0.5 \frac{P(M)}{P(L)}
\]

The model was trained on a random sample of cases, news documents and all UMLS terms.

### 4.3 Fast Lookup of Medical Terms

The fast lookup finds medical terms in sentences by finding the longest spans of medical words in the sentence that

<table>
<thead>
<tr>
<th>Table 2: The table lists non-stop words from Table 1 and the corresponding (arbitrary) IDs</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Word id</strong></td>
</tr>
<tr>
<td>1:3</td>
</tr>
<tr>
<td>2:4</td>
</tr>
<tr>
<td>3:6</td>
</tr>
<tr>
<td>4:7</td>
</tr>
<tr>
<td>5:9</td>
</tr>
<tr>
<td>5:8</td>
</tr>
</tbody>
</table>

used in a medical sense in non-medical collections is “miscarriage”, as in “miscarriage of justice” in legal documents. And an example of a term that is almost always used in a non-medical sense in non-medical collections is the term “section”.

When processing medical collections (e.g., medical journals), the system tags all terms, regardless of their labels. However, when processing non-medical collections, the system skips over problematic terms, and tags ambiguous terms only if they have a medical context. Hence, correct labeling significantly impacts the quality of the solution.

We used a support vector machine to classify UMLS terms in the authority file. Each term is represented using a set of approximately a dozen features. Some of the features are the following.

A boolean feature indicating whether the medical is comprised of common words. If each word in the term is commonly found in news corpora, the feature is set to 1 otherwise it is set to zero.

A character based language model score indicating how probable the term is to occur in a legal or news collection rather than in a medical collection. The idea behind this is that the medical terms containing character n-grams that occur more frequently in legal texts than in medical texts are more ambiguous. A character language based model each for legal, news and medical texts is built and are applied on a given UMLS string to decide its ambiguity. If \(P(N)\), \(P(L)\) and \(P(M)\) are probabilities of a term belonging to news, legal and medical genres respectively, the ambiguity score is defined as

\[
Score = 0.5 \frac{P(M)}{P(N)} + 0.5 \frac{P(M)}{P(L)}
\]

\(\mu(\text{currKey})\) is a function that returns medical concept id associated with hash key.

Figure 4. Pseudo-code for Medical Term Tagging Algorithm
can be converted into hash keys matching hash keys in a medical term hash table. By span of medical term, we mean medical words in the sentence that are contiguous or separated only by filler words such as “of”, “the”, and “his”. By medical words, we mean the individual non-filler words found in the UMLS and drug terms we used to create the medical term hash keys. Converting a span of medical words into a hash key involves converting each word into a unique number through a hash lookup in the medical word table, ordering the numbers from smallest to largest, and concatenating the numbers with “:” separators.

We perform minimal pre-processing before we apply the tagging algorithm. Namely, sentence splitting, tokenization and normalization (e.g., removing possessive constructs and using singular forms). The tagging algorithm is listed in Figure 4.

In Figure 4, S is a vector of tokens w taken from a sentence to be tagged. wi denotes a single word token at position i in the sentence. T denotes the set of unique tokens that comprise terms in the medical term list. C is the set of hash keys that correspond to a valid medical term. So, C might include, for example, the hash key “1:3” which maps to the concept for “heart attack”. D is the set of delimiter tokens that signal the end of a run of medical terms. Delimiter tokens include semi colons, end of sentence markers, and words which are both outside the T set of medical word tokens and outside the set of filler tokens such as “of” and “the”. currConcept is the id of the concept associated with a given medical term. For example, the id C0012324 is associated with the medical terms “heart attack” and “myocardial infarction”. tokenStack is the stack of tokens from a sentence representing a sequence of w valid tokens in sentence S. For example, tokenStack will contain [lung, cancer] for the sequence “cancer of the lung” in sentence S [He, has, cancer, of, the, lung, eos]. indexStack is the stack of position indices corresponding to w tokens in tokenStack. For the [lung, cancer] tokenStack, the indexStack is [6,3].

\[ \sigma(\text{tokenstack}) \] is a function that returns the hash key associated with token terms w in tokenStack. For example if the token “cancer” is assigned number 309 and token “lung” is assigned 851, the hash key for both “lung cancer” and “cancer of the lung” is “309:851”. Tokens in keys are order independent (except for medical device terms and drugs) and do not include filler words such as “of” and “the”.

\[ \mu(\text{currKey}) \] is a function that returns a medical term concept id associated with a hash key. If no concept is associated with a hash key, no concept number is returned. Multiple hash keys can be associated with a single concept id. But a single hash key can be associated with only one concept id.

Our performance testing showed that the tagging algorithm as described in Figure 4 has a very high precision and good recall. To improve performance further, we extended the algorithm in the following ways.

We included code to deal with structures in which the head noun of the medical phrase was modified by multiple adjectives in a parallel structure. By this, we mean phrases such as “throat, liver, and kidney cancer”.

Some organizations share names with medical terms such the drug “Criticare” in Criticare Systems Inc and the disease “cancer” in “American Cancer Association”. In order to avoid tagging medical terms embedded in organization names, we checked the words in the immediate context of the medical term for capitalization. If the term is capitalized, is surrounded by capitalized words, and is proximate to an organization term such “Inc” or “Association”, we assume the term is part of an organization name and do not tag it as a medical term.

We applied a part of speech mask over sentences to eliminate phrases having problematic constructs like those ending with adjectives. For example, consider the phrase “atherosclerosis in the cerebral vessels”. Since “in” and “the” are filler words, the substring “atherosclerosis in the cerebral” is linked to “cerebral atherosclerosis”. In order to avoid this, we use a rule-based approach to prevent tagging phrases ending with adjectives/adverbs.

4.4 Disambiguator - Context Checking for Ambiguous Terms

Some terms in the authority have medical and non-medical senses, especially when used in non-medical corpora such as legal and news documents. For example, the term “miscarriage” in legal documents is often included in the non-medical phrase “miscarriage of justice”. To deal with this problem, we constructed a classifier that determines the sense of a term based on its context (i.e., the words before and after the term). This classifier is used whenever an “ambiguous” authority term (see section 4.2) is found in a non-medical collection (e.g., a case law document).

When an authority term is actually used in a medical sense, we mark the words surrounding it as comprising a medical context. Analogously, when an authority term takes on a non-medical sense, we mark the words surrounding it as comprising a non-medical context. The classifier compares the likelihood of a context word conditioned on a medical context with its likelihood when conditioned on a non-medical context.

Mathematically, this is written as follows:

\[ P(T_{isMed}) \prod_{i=0}^n P(w_i|M) \geq P(T_{notMed}) \prod_{i=0}^n P(w_i|\neg M) \]

(1)

where:

- \( w_i \) denotes a context word (there are n of them).
- \( P(T_{isMed}) \) is the prior probability that a term is assuming a medical sense.
- \( P(T_{notMed}) \) is the prior probability that a term is not assuming a medical sense.
- \( P(w_i|M) \) is the conditional probability of the word \( w_i \) occurring in a medical context.
- \( P(w_i|\neg M) \) is the conditional probability of the word \( w_i \) occurring in a non-medical context.

The conditional probabilities \( P(w_i|M) \) and \( P(w_i|\neg M) \) are computed as follows:

\[ P(w_i|M) = \frac{CM(w_i)}{\sum_{w_i} CM(w_i)} \]

\[ P(w_i|\neg M) = \frac{CnotM(w_i)}{\sum_{w_i} CnotM(w_i)} \]

Where \( CM(w_i) \) is the frequency of the word \( w_i \) in a me-
dical context, and $CnotM(w_i)$ is the frequency of the word $w_i$ in a non-medical context.

We use an Expectation Maximization (EM) algorithm [6] to estimate the prior probabilities of the terms and the conditional probabilities. We iteratively compute the medical word frequencies and the priors to maximize the likelihood of a given term being medical or non-medical in a given context.

The E-Step consists of estimating the probability of a word being in a medical context, given the prior probabilities of all medical terms. The priors are used in classifying each medical term instance as medical or non-medical and the word counts in the medical contexts are used to compute the medical word probabilities. We start the first iteration of the E-step with initial conditional probabilities of some words based on hand tagged medical terms in a small hand tagged corpus.

The M-Step consists of using the medical word probabilities computed in the E-Step to estimate the probability of a medical term in a given context being medical or not according to equation 1. The E-M iteration is continued until the number of terms that switch the class (from medical to non-medical or vice-versa) is zero. The final contextual model is stored to be used later online during tagging.

5. EVALUATION

We evaluated our system against a system constructed from the CASS parser [2] and the NIH Metamap medical term matching program. The Metamap plus CASS system identifies medical terms in text by finding noun phrases in a text sentence with the CASS shallow parser and mapping the noun phrases to medical concepts in UMLS using Metamap’s comparison function. We did not include drug names in our experiment because we used Micromedex Redbook rather than UMLS as our source for drug names.

We tested the taggers using four different document collection sets: news, caselaw, jury verdicts, and medline abstracts. The test document collection sets were randomly selected from larger collection sets containing the same type of document. The only restriction we placed on the documents was that the document needed to contain at least one medical word from our medical word index. We placed this constraint on the document selection criteria to avoid forcing our editors to read documents that very probably contained no medical terms. The larger news and caselaw collections were especially prone to having medical concepts contained no medical terms. The larger news and caselaw collections were especially prone to having medical concepts contained no medical terms. The larger news and caselaw collections were especially prone to having medical concepts contained no medical terms. The larger news and caselaw collections were especially prone to having medical concepts contained no medical terms. The larger news and caselaw collections were especially prone to having medical concepts contained no medical terms.

We placed this constraint on the document selection criteria to avoid forcing our editors to read documents that very probably contained no medical terms. The larger news and caselaw collections were especially prone to having medical concepts contained no medical terms. The larger news and caselaw collections were especially prone to having medical concepts contained no medical terms.

The processing speed per document for the Metamap tagging method versus the fast lookup tagger is given in Table 5. For all four collections, the overall throughput for the Metamap method was 0.46 kilobytes per second. The overall throughput for the Fast Lookup Tagger was 63.8 kilobytes per second. This represents a 139 fold increase in processing speed.

The remarkable speed increase in the fast lookup tagger is due to the efficiency with which it finds the boundaries of medical terms within text and the efficiency with which it looks up terms in the UMLS term to find their pertinent medical concepts.

The precision and recall of the two taggers is shown below in Table 4.

The superior precision of the fast tagger on spotting medical terms in medical text is in part due to the fact that the Metamap program was overly aggressive in mapping noun phrases to medical concepts. For example, in the following sentence, Metamap maps “chronic and severely ill patients” to C0008679, the concept id corresponding to severe illness.

The superior precision of the fast tagger on spotting medical terms in medical text is in part due to the fact that the Metamap program was overly aggressive in mapping noun phrases to medical concepts. For example, in the following sentence, Metamap maps “chronic and severely ill patients” to C0008679, the concept id corresponding to severe illness.

The facts are not in dispute and it is agreed that no further hearing is necessary.

Many recall errors made by Metamap plus CASS and not made by the fast tagger in both medical and non-medical context are due parsing errors. For example, Metamap failed to map “head and neck injuries” to its proper concept C0011053 because the CASS parser chunked “head” separately from “neck injuries”.

John William Van Vrancken IV, 17, was in stable condition, suffering from head and neck injuries.

A large number of the CASS parsing errors in the Medline corpus occurred on non-standard text. For example, CASS interpreted “Gastrointestinal Hemorrhage/Surgery” as a single noun phrase rather than as two noun phrases separated by the “/” character. A few simple changes to the CASS parser could probably fix these problems. If such changes to the parser were made, the precision and recall of the Metamap and CASS system on the Medline corpus would be close to the precision and recall of the fast tagger.

6. DISCUSSION

The speed and accuracy of the fast term tagger result from using words in the medical term list itself along with some relatively simple stopping conditions to find both the boundaries of the medical terms in the text and the constituent words of the terms such that the words can be converted to hash keys for medical concept lookup.

The fast term tagger described here is an example of a system in which a simpler approach outperforms an approach based on more complex processing. This is analogous to the discovery made during the MUC [1] competitions that the FASTUS [3] system, based on cascaded finite state machines, did a better job extracting information from text than systems based on deep parsing. Our method relies on the fact that the UMLS Metathesaurus is a robust term list. By using words in the medical term list itself along with some relatively simple stopping conditions to find both the boundaries of the medical terms in the text and the constituent words of the terms such that the words can be converted to hash keys for medical concept lookup.

By robust list, we mean a list that contains many of the term variants corresponding to the semantic concepts covered by the list. In the case of UMLS, this means that semantically related terms such as “hepatic cancer” and “liver cancer” are already loaded into the term list. And this means we do not have to rely on semantic transforms such as those described in FASTR to expand the list to discover these variants.

Services offered by the CMHCs became less comprehensive and more oriented toward chronic and severely ill patients.

John William Van Vrancken IV, 17, was in stable condition, suffering from head and neck injuries.
Table 4: Precision and Recall of Medical Taggers on the Four Different Text Collections

<table>
<thead>
<tr>
<th>Collection Type</th>
<th>Metamap Precision</th>
<th>Metamap Recall</th>
<th>Fast Tagger Precision</th>
<th>Fast Tagger Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>news</td>
<td>0.62</td>
<td>0.63</td>
<td>0.94</td>
<td>0.89</td>
</tr>
<tr>
<td>caselaw</td>
<td>0.19</td>
<td>0.60</td>
<td>0.91</td>
<td>0.95</td>
</tr>
<tr>
<td>Jury verdicts and settlements</td>
<td>0.83</td>
<td>0.53</td>
<td>0.95</td>
<td>0.79</td>
</tr>
<tr>
<td>medline</td>
<td>0.74</td>
<td>0.73</td>
<td>0.94</td>
<td>0.84</td>
</tr>
</tbody>
</table>

Table 5: Tagger performance on different collections

<table>
<thead>
<tr>
<th>Collection Type</th>
<th>Collection size</th>
<th>Metamap Tagging speed in kbps</th>
<th>Fast tagger speed in kbps</th>
</tr>
</thead>
<tbody>
<tr>
<td>News</td>
<td>4572 KB</td>
<td>0.6</td>
<td>114.3</td>
</tr>
<tr>
<td>Caselaw</td>
<td>3800 KB</td>
<td>0.5</td>
<td>84.4</td>
</tr>
<tr>
<td>Jury verdicts and settlements</td>
<td>256 KB</td>
<td>0.21</td>
<td>7.3</td>
</tr>
<tr>
<td>Medline Abstracts</td>
<td>1900 KB</td>
<td>0.31</td>
<td>42.2</td>
</tr>
</tbody>
</table>

For future applications, we plan to investigate whether we can mine robust term lists from legal and non-legal corpora using a small initial seed set of terms. And whether we can in turn build fast tagging systems using these mined term lists. We also intend to address recall errors produced by the fast term tagger when non-filler terms are embedded within medical term noun phrases.

7. CONCLUSION

This paper describes a novel method of tagging medical terms in legal, medical, and news text that is very fast and also has high recall and precision. We have shown that an effective tagger for medical terms related to diseases, injuries, drugs, medical devices, and medical procedures can be built using words from a robust medical term list along with a probabilistic term classifier that uses local context to disambiguate terms being used in a medical sense from terms being used in a non-medical sense. The system produces more accurate tags and is over 100 times faster than systems based on shallow parsing and complicated noun phrase matching.

8. REFERENCES


