Automatic Mapping Clinical Notes to Medical Terminologies

Jon Patrick, Yefeng Wang and Peter Budd
School of Information Technologies
University of Sydney
NSW 2006, Australia
{jonpat,ywang1,pbud3427}@it.usyd.edu.au

Abstract

Automatic mapping of key concepts from clinical notes to a terminology is an important task to achieve for extraction of the clinical information locked in clinical notes and patient reports. The present paper describes a system that automatically maps free text into a medical reference terminology. The algorithm utilises Natural Language Processing (NLP) techniques to enhance a lexical token matcher. In addition, this algorithm is able to identify negative concepts as well as performing term qualification. The algorithm has been implemented as a web based service running at a hospital to process real-time data and demonstrated that it worked within acceptable time limits and accuracy limits for them. However broader acceptability of the algorithm will require comprehensive evaluations.

1 Introduction

Medical notes and patient reports provide a wealth of medical information about disease and medication effects. However a substantial amount of clinical data is locked away in non-standardised forms of clinical language which could be usefully mined to gain greater understanding of patient care and the progression of diseases if standardised. Unlike well written texts, such as scientific papers and formal medical reports, which generally conform to conventions of structure and readability, the clinical notes about patients written by a general practitioners, are in a less structured and often minimal grammatical form. As these notes often have little if any formal organisation, it is difficult to extract information systematically. Nowadays there is an increased interest in the automated processing of clinical notes by Natural Language Processing (NLP) methods which can exploit the underlying structure inherent in language itself to derive meaningful information (Friedman et al., 1994).

In principle, clinical notes could be recorded in a coded form such as SNOMED CT (SNOMED International, 2006) or UMLS (Lindberg et al., 1993), however, in practice notes are written and stored in a free text representation. It is believed that the encoding of notes will provide better information for document retrieval and research into clinical practice (Brown and Sönksen, 2000). The use of standard terminologies for clinical data representation is critical. Many clinical information systems enforce standard semantics by mandating structured data entry. Transforming findings, diseases, medication procedures in clinical notes into structured, coded form is essential for clinician research and decision support system. Using concepts in domain specific terminology can enhance retrieval. Therefore, converting free text in clinical notes to terminology is a fundamental problem in many advanced medical information systems.

SNOMED CT is the most comprehensive medical terminology in the world and it has been adopted by the Australia government to encode clinical disease and patient reports. The doctors want a system to develop a standard terminology on SNOMED CT for reporting medical complaints so that their information is exchangeable and semantically consistent for other practitioners, and permit automatic extraction of the contents of clinical notes to compile statistics about diseases and their treatment. Translate medical concepts in free text into standard medical terminology in coded form is a hard problem, and cur-

Proceedings of the 2006 Australasian Language Technology Workshop (ALTW2006), pages 75–82.

rently mostly solved by employing human coders trained both in medicine and in the details of the classification system. To increase the efficiency and reduce human cost, we are interested to develop a system that can automate this process. There are many researchers who have been working on mapping text to UMLS (The Unified Medical Language System), however, there is only a little work done on this topic for the SNOMED CT terminology. The present work proposes a system that automatically recognises medical terms in free text clinical notes and maps them into SNOMED CT terminology. The algorithm is able to identify core medical terms in clinical notes in real-time as well as negation terms and qualifiers. In some circles SNOMED CT is termed an ontology, however this paper only covers its role as a terminology so we will use that descriptor only.

2 Related Work

2.1 Concept Mapping in Medical Reports

There has been a large effort spent on automatic recognition of medical and biomedical concepts and mapping them to medical terminology. The Unified Medical Language System Metathesaurus (UMLS) is the world's largest medical knowledge source and it has been the focus of much research. Some prominent systems to map free text to UMLS include SAPHIRE (Hersh et al., 1995), MetaMap (Aronson, 2001), IndexFinder (Zou et al., 2003), and NIP (Huang et al., 2005). The SAPHIRE system automatically maps text to UMLS terms using a simple lexical approach. IndexFinder added syntactic and semantic filtering to improve performance on top of lexical mapping. These two systems are computationally fast and suitable for real-time processing. Most of the other researchers used advanced Natural Language Processing Techniques combined with lexical techniques. For example, NIP used sentence boundary detection, noun phrase identification and parsing. However, such sophisticated systems are computationally expensive and not suitable for mapping concepts in real time.

MetaMap has the capacity to code free text to a controlled terminology of UMLS. The MetaMap program uses a three step process started by parsing free-text into simple noun phrases using the Specialist minimal commitment parser. Then the phrase variants are generated and mapping candidates are generated by looking at the UMLS source vocabulary. Then a scoring mechanism is used to evaluate the fit of each term from the source vocabulary, to reduce the potential matches. The MetaMap program is used to detect UMLS concepts in e-mails to improve consumer health information retrieval (Brennan and Aronson, 2003).

The work done by (Hazelhurst et al., 2005) is on taking free text and mapping it into the classification system UMLS (Unified Medical Language System). The basic structure of the algorithm is to take each word in the input, generate all synonyms for those words and find the best combination of those words which matches a concept from the classification system. This research is not directly applicable to our work as it does not run in real time, averaging 1 concept matched every 20 seconds or longer.

2.2 Negation and Term Composition

Negation in medical domains is important, however, in most information retrieval systems negation terms are treated as stop words and are removed before any processing. UMLS is able to identify propositions or concepts but it does not incorporate explicit distinctions between positive and negative terms. Only a few works have reported negation identification (Mutalik et al., 2001; Chapman et al., 2001; Elkin et al., 2005).

Negation identification in natural languages is complex and has a long history. However, the language used in medical domains is more restricted and so negation is believed to be much more direct and straightforward. Mutalik et al (2001) demonstrated that negations in medical reports are simple in structure and syntactic methods are able to identify most occurrences. In their work, they used a lexical scanner with regular expressions and a parser that uses a restricted context-free grammar to identify pertinent negatives in discharge summaries. They identify the negation phrase first then identify the term being negated.

In the work of (Chapman et al., 2001), they used a list of negation phrases derived from 1,000 sentences of discharge summaries. The text is first indexed by UMLS concepts and a rule base is then applied on the negation phrases to identify the scope of the negation. They concluded that medical text negation of clinical concepts is more restricted than in non-medical text and medical narrative is a sublanguage limited in its purpose, so therefore may not require full natural language understanding.

3 SNOMED CT Terminology

The Systematized Nomenclature of Medicine Clinical Terminology (SNOMED CT) is developed and maintained by College of American Pathologists. It is a comprehensive clinical reference terminology which contains more than 360,000 concepts and over 1 million relationships. The concepts in SNOMED CT are organised into a hierarchy and classified into 18 top categories, such as *Clinical Finding*, *Procedure*, *Body Part*, *Qualifier* etc. Each concept in SNOMED CT has at least three descriptions including 1 preferred term, 1 fully specified name and 1 or more synonyms. The synonyms provide rich information about the spelling variations of a term, and naming variants used in different countries. The concepts are connected by complex relationship networks that provide generalisation, specialisation and attribute relationships, for example, "*focal pneumonia*" is a specialisation of "*pneumonia*". It has been proposed for coding patient information in many countries.

4 Methods

4.1 Pre-processing

Term Normalisation

The clinical notes were processed at sentence level, because it is believed that the medical terms and negations do not often cross sentence boundaries. A maximum entropy model based sentence boundary detection algorithm (Reynar, and Ratnaparkhi, 1996) was implemented and trained on medical case report sentences. The sentence boundary detector reports an accuracy of 99.1% on test data. Since there is a large variation in vocabulary written in clinical notes compared to the vocabulary in terminology, normalisation of each term is necessary. The normalisation process includes stemming, converting the term to lower case, tokenising the text into tokens and spelling variation generation (haemocyte vs. hemocyte). After normalisation, the sentence then is tagged with POS tag and chunked into chunks using the GENIA tagger (Tsuruoka et al., 2005). We did not remove stop words because some stop words are important for negation identification.

Administration Entity Identification

Entities such as *Date*, *Dosage* and *Duration* are useful in clinical notes, which are called administration entities. A regular expression based named entity recognizer was built to identify administration units in the text, as well as quantities such as *5 kilogram*. SNOMED CT defined a set of standard units used in clinical terminology in the subcategory of *unit* (*258666001*). We extracted all such units and integrated them into the recognizer. The identified quantities are then assigned the SNOMED CT codes according to their units. Table 1 shows the administration entity classes and examples.

Entity Class	Examples
Dosage	40 to 45 mg/kg/day
Blood Pressure	105mm of Hg
Demography	69 year-old man
Duration	3 weeks
Quantity	55x20 mm

Table 1: Administration Entities and Examples.

4.2 SNOMED CT Concept Matcher

Augmented SNOMED CT Lexicon

The Augmented Lexicon is a data structure developed by the researchers to keep track of the words that appear and which concepts contain them in the SNOMED CT terminology. The Augmented Lexicon is built from the Description table of SNOMED CT. In SNOMED CT each concept has at least three descriptions, preferred term, synonym term and fully specified name. The fully specified name has the top level hierarchy element appended which is removed. The description is then broken up into its atomic terms, i.e. the words that make up the description. For example, *myocardial infarction (37436014)* has the atomic word *myocardial* and *infarction*. The UMLS Specialist Lexicon was used to normalise the term. The normalisation process includes removal of stop words, stemming, and spelling variation generation. For each atomic word, a list of the Description IDs that contain that word is stored as a linked list in the Augmented Lexicon. An additional field is stored alongside the augmented lexicon, called the "Atomic term count" to record the number of atomic terms that comprise each description. The table is used in determining the accuracy of a match by informing the number of tokens needed for a match. For example, the atomic term count for *myocardial infarction (37436014)* is *2*, and the accuracy ratio is *1.0*. Figure 1 contains a graphical representation of the Augmented SNOMED CT Lexicon.

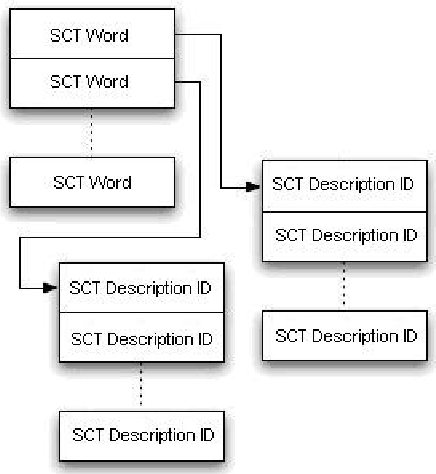

Figure 1: Augmented SCT Lexicon

Token Matching Algorithm

The token matching algorithm takes unstructured text and pre-processes it using the same techniques as are applied to the concepts when generating the augmented lexicon. It then attempts to find each SNOMED CT Description which is contained in the input sentence. For each word, the algorithm looks up the Augmented Lexicon, retrieving a list of the descriptions which contain the word. Figure 2 gives a graphical representation of the data structure used in the algorithm. The elements of the matrix are n-grams from the input sentence with the diagonal line sequence runs of two words. The remainder of the matrix is the cell to the left of it with the next word appended onto it. In this way the algorithm covers every possible sequence of sequential tokens

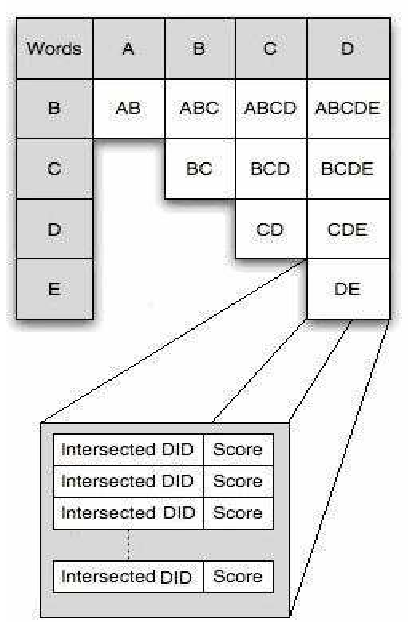

Figure 2: Matching Matrix example

The data stored in each cell is a list of Description IDs (DID) that are in all the tokens that comprise the cell, i.e. the intersection of each set of DID of each word. The score is then calculated using the "atomic term count", which stores the number of tokens that make up that description. The score is the number of tokens in the current cell that have the DID in common divided by the number of tokens in the full description, i.e.:

$$\text{Score} = \frac{\#\text{of Tokens in Sequence}}{\#\text{of Tokens in Full Description}}$$

The algorithm itself is shown here in Figure 3 as pseudo-code. Step 1 is building the Matching Matrix. Step 2 is using the Matching Matrix to find the best combination of sequences that gives the highest score. This final score is dependant on the number of tokens used to make the match divided by the total number of tokens in the input stream, i.e.:

$$\text{Score} = \frac{\#\text{of Tokens used in all matches}}{\#\text{of Tokens in total input stream}}$$

STEP 1

for each word in list:
    add entry to the Matching Matrix
    for new column:
        Intersect new word with
        cell from matching table

Sort the matching array in descending order based off
    the scores
for each row in the matrix:
    start at the right most cell

STEP 2

if the top score for the cell is 1.0
    add cell details to current best match list,
    update current match score.
    recursively call STEP 2 on cell
    (row=column+2, column=right)
else:
    move one column left to the next cell
    or
    the right-most cell of the next row if left cell
    empty

repeat STEP 2 until visited all cells

Figure 3: Matching Algorithm Pseudo-code

Adding Abbreviations

Different sub-domains have different definitions of abbreviations. In medical domain, the abbreviations are highly ambiguous, as (Liu et al., 2002) show that 33% of abbreviations in UMLS are ambiguous. In different hospitals, they have their own convention of abbreviations, and the abbreviations used are not the same cross the sections in the same sub-domain. This creates difficulties for resolving the abbreviation problem. As we are processing clinical data in the RPAH (Royal Prince Alfred Hospital) ICU (Intensive Care Unit), we believe that the abbreviations used in their reports are restricted to a sub-domain and not that ambiguous. We use a list of abbreviations provided by the ICU department, and integrated them into the Augmented Lexicon. The abbreviations are manually mapped to SNOMED CT concepts by two experts in RPAH. The list consists of 1,254 abbreviations, 57 of them are ambiguous (4.5%). We decided not to disambiguate the abbreviations in the token matching, but return a list of all possible candidates and leave it for later stage to resolve the ambiguity.

4.3 Negation Identification

In our system, we aim to identify the negation phrases and the scope of the negation. Two kinds of negations are identified, the pre-coordianted SNOMED CT concepts and concepts that are explicitly asserted as negative by negation phrases. A pre-coordinated phrase is a term that exists in SNOMED CT terminology that represents a negative term, for example *no headache*.

SNOMED CT contains a set of per-coordinated negative terms under the *Clinical Finding Absent (373572006)* category that indicate the absence of findings and diseases. However, SNOMED CT is not an exhaustive terminology, it is not able to capture all negated terms. Moreover clinical notes have many negation forms other than “absence”, such as “denial of procedures”. For a negative term that has a pre-coordinated mapping in SNOMED CT, we mark up this term using the SNOMED CT concept id (CID), for other negations, we identify the negation phrases and the SNOMED CT concepts that the negation applies on. The following examples show the two different negations:

no headache (Pre-coordinated negation term)
"absent of"
CID: 162298006
no headache (context-dependent category)

no evidence of neoplasm malignant
(Explicitly asserted negation)
negation phrase: "no evidence of"
CID: 363346000
malignant neoplastic disease (disorder)

Figure 4: Examples of Negations

To identify explicitly asserted negation, we implemented a simple-rule based negation identifier similar to (Chapman et al, 2001; Elkin et al, 2005). At first the SNOMED CT concept id is assigned to each medical term, the negation phrases then are identified using a list of negation phrases in (Chapman et al, 2001). Then a rule base is applied on the negation phrase to check at its left and right contexts to see if any surrounding concepts have been negated. The algorithm is able to identify the negation of the form:

negation phrase … (SNOMED CT phrase)*
(SNOMED CT phrase)* … negation phrase

The contexts can up to 5 non-stopwords long, which allow identification of negation of coordination structure, for example in the following sentence segment:

… and pelvis ***did not*** reveal **retroperitoneal lymphadenopathy** or **mediastinal lymphadenopathy** …

In this sentence segment, the terms, *retroperitoneal lymphadenopathy* and *mediastinal lymphadenopathy* are negated.

Whenever there is a overlapping between Pre-coordinated negation and explicitly asserted negation, we identify the term as pre-coordinated negation. For example, the term *no headache (162298006)* will not be identified as the negation of *headache (25064002)*.

4.4 Qualification and Term Composition

In medical terminology a term may contain an atomic concept or composition of multiple concepts, for example the term pain is an atomic concept and *back pain* represents composition two atomic concepts back and pain. Some composite concepts appear as single concepts in medical terminology, for example *back pain* is a single concept in SNOMED CT. Such concept is called pre-coordinated concept. However, the medical terms can be composed by adding adjective modifiers to form new terms, for example, the add qualifiers to the concept *pain* can have *back pain*, *chronic back pain*, *chronic low back pain* etc. It is impossible to pre-coordinate combinations of all qualifiers into a terminology, because it will lead to term explosion. Term composition allows user to create new composite concepts using two or more single or composite concept. It is a solution to so called content completeness problem.

The SNOMED CT terminology has a subclass of terms called qualifier values. The qualifier values are used to qualify core concepts. The SNOMED CT defined qualifying relationship adds additional information about a concept without changing its meaning. In most cases, the qualifier is an adjective. There are also some nouns classified as qualifiers, such as *fractions (278277004)*.

The purpose of the qualifier matching is to perform term composition. We separate the qualifiers apart from the Augmented Lexicon when performing concept matching, and build another lexicon that contains only qualifiers. Another reason for treating the qualifier differently is that the qualifier values always conflict with commonly used English words, for example, the unit qualifier *day (258703001)*, side qualifier *left (7771000)*, technique qualifier *test (272394005)*. Such qualifiers cause noise when mapping text to concepts, and they should be refined by looking at their context.

The Concept Matchers runs at first to identify any SNOMED CT concepts and qualifiers. A search then is run to look at the qualifiers' surroundings using the following rules to identify the scope of qualification. A concept can have multiple qualifiers to modify it.

(Qualifier / JJ|NN)* … (Concept / NN)*
(Concept / NN)* … (Qualifier / JJ|NN)*

The first rule aims identify left hand side qualifications, for example in the following sentence segment:

… She had severe **lethargy** and ***intermittent right upper*** **abdominal discomfort** …

The second rule aims to identify right hand side qualification, for example:

... **autoimmune screening** were ***normal*** …

If no concepts are found with in a context window, the qualifier then is not considered as a modifier to any medical concepts, thus removed to reduce noise.

5 Results and Discussion

The token matching algorithm has been implemented as a web-based service named TTSCT (Text to SNOMED CT) that provides web interfaces for users to submit clinical notes and respond with SNOMED CT codes in real-time. The system is able to encode SNOMED CT concepts, qualifiers, negations, abbreviations as well as administration entities. It has been developed as the first step to the analysis and deep understanding of clinical notes and patient data. The system has been installed in RPAH (Royal Prince Alfred Hospital) ICU (Intensive Care Unit) aiming to collect bedside patient data. The web interface has been implemented in several clinical form templates the RPAH, allowing data to be captured as the doctors fill in these forms. A feedback form has been implemented allowing clinicians to submit comments, identify terms that are missed by the system and submit corrections to incorrectly labelled terms. Figure 5 shows the concepts that have been identified by the TTSCT system and Figure 6 shows the responding SNOMED CT codes.

No neoplasm malignant $_{\text{negation}}$ seen.

Sections confirm **CRANIOPHARYNGIOMA** $_{\text{concept}}$ with **small** $_{\text{qualifier}}$ **fragments** $_{\text{qualifier}}$ of **adjacent** $_{\text{qualifier}}$ **brain tissue** $_{\text{concept}}$.

The **slides** $_{\text{concept}}$ show degenerate **atypical** $_{\text{qualifier}}$ urothelial **cells** $_{\text{concept}}$ occurring in **sheets** $_{\text{qualifier}}$ and singly with **hyperchromatic** $_{\text{qualifier}}$ **enlarged** $_{\text{qualifier}}$ **irregular** $_{\text{qualifier}}$ **nuclei** $_{\text{concept}}$.

Figure 5: A Sample Clinical Note

SNOMED CT Concept	SCT Concept ID	SCT Fully Specified Name
CRANIOPHARYNGIOMA	40009002	Craniopharyngioma (morphologic abnormality)
	189179009	Craniopharyngioma (disorder)
Brain tissue	256865009	Brain tissue (substance)
Cells	4421005	Cell structure (cell structure)
	362837007	Entire cell (cell)
hyperchromatic	9767008	Hyperchromatism (morphologic abnormality)

Qualifiers	SCT Concept ID	SCT Fully Specified Name	Scope of Qualification
Small	255507004	Small (qualifier value)	
	263796003	Lesser (qualifier value)	
Fragments	29140007	Fragment of (qualifier value)	
Adjacent	18769003	Juxta-posed (qualifier value)	brain tissue
Atypical	112231000	Atypical (qualifier value)	Cells
Sheets	255292000	Sheets (qualifier value)	
Enlarged	260376009	Enlarged (qualifier value)	
Irregular	49608001	Irregular (qualifier value)	
Fragments	29140007	Fragment of (qualifier value)	Tissue

Negation	Negation Phrase	Negative Term
no neoplasm malignant	No	neoplasm malignant (86049000)

Figure 6: Concepts, Qualifiers and Negations Identified From the Sample Note

We are currently collecting test data and evaluating the accuracy of our method. We plan to collect patient reports and cooperate with the clinicians in the RPAH to identify correct mappings, missing mappings and incorrect mappings. Although the algorithm hasn't been comprehensively evaluated on real data, we have collected some sample patient reports and a few feedback from some clinicians. Preliminary results demonstrate that the algorithm is able to capture most of the terms within acceptable accuracy and response time.

By observation, missing terms and partially identified terms are mainly due to the incompleteness in SNOMED CT. In the above example, the *atypical urothelial cells* is only partially matched, because neither *atypical urothelial cell* is present in SNOMED CT as a single term nor *urothelial* can be found as a qualifier in SNOMED CT. However the qualified term *moderate urothelial cell atypia* can be found in SNOMED CT. This raises the question of term composition and decomposition because the terms in the terminology have different levels of composition and the qualification can be written in a different order with morphological transformation (*urothelia cell atypia* vs. *atypical urothelial cell*). The qualifier ontology and term relationships must be addressed to make sure term composition is done in a reliable manner.

Restricting the concept mapping to noun phrase chunkers can rule out many false positives and also increase the speed of processing, however many pre-coordinated terms and qualifications cross noun phrase boundaries, for example the term "*Third degree burn of elbow (87559001)*" will be broken into two terms "*Third degree burn (403192003)*" and "*elbow (76248009)*" and their relationship not preserved.

6 Conclusions

In conclusion, we propose an algorithm to code free text clinical notes to medical terminology and implemented it as a web-service system. The algorithm utilised NLP techniques to enhance lexical concept mappings. A qualifier identifier and negation identifier have been implemented for recognising composite terms and negative concepts, which can then create more effective information retrieval and information extraction. The system is yet to be fully evaluated, nevertheless the test on sample data shows it is already meeting expectations. In the future, we will perform comprehensive evaluation for the algorithm on real clinical data, and compare the system with some well known term indexing algorithms.

References

Aronson, A. R. (2001). *Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program*. Proc AMIA Symp 17: 21.

Brennan, P. F. and A. R. Aronson (2003). *Towards linking patients and clinical information: detecting UMLS concepts in e-mail*. Journal of Biomedical Informatics 36(4/5): 334-341.

Brown, P. J. B. and P. Sönksen (2000). *Evaluation of the Quality of Information Retrieval of Clinical Findings from a Computerized Patient Database Using a Semantic Terminological Model.* Journal of the American Medical Informatics Association 7: 392-403.

Chapman, W. W., W. Bridewell, et al. (2001). *A Simple Algorithm for Identifying Negated Findings and Diseases in Discharge Summaries.* Journal of Biomedical Informatics 34(5): 301-310.

Chute, C. G. and P. L. Elkin (1997). *A clinically derived terminology: qualification to reduction*. Proc AMIA Annu Fall Symp 570: 4.

Elkin, P. L., S. H. Brown, et al. (2005). *A controlled trial of automated classification of negation from clinical notes*. BMC Medical Informatics and Decision Making 2005(5): 13.

Friedman, C., P. O. Alderson, et al. (1994). *A general natural-language text processor for clinical radiology.* Journal of the American Medical Informatics Association 1(2): 161-174.

Friedman, C., L. Shagina, et al. (2004). *Automated Encoding of Clinical Documents Based on Natural Language Processing*. J Am Med Inform Assoc 11: 392-402.

Hazlehurst, B., H. R. Frost, et al. (2005). *MediClass: A System for Detecting and Classifying Encounter-based Clinical Events in Any Electronic Medical Record*, American Medical Informatics Association.

Hersh, W. R. and D. Hickam (1995). *Information retrieval in medicine: The SAPHIRE experience.* Journal of the American Society for Information Science 46(10): 743-747.

Huang, Y., H. J. Lowe, et al. (2005). *Improved Identification of Noun Phrases in Clinical Radiology Reports Using a High-Performance Statistical Natural Language Parser Augmented with the UMLS Specialist Lexicon*, American Medical Informatics Association.

Lindberg, D. A., B. L. Humphreys, et al. (1993). *The Unified Medical Language System.* Methods Inf Med 32(4): 281-91.

Liu H., Johnson S. B. and Friedman C. (2002), *Automatic Resolution of Ambiguous Terms Based on Machine Learning and Conceptual Relations in the UMLS.* Journal of the American Medical Informatics Association, Vol. 9, No. 6, Pages 621-636.

Mutalik, P. G., A. Deshpande, et al. (2001). *Use of General-purpose Negation Detection to Augment Concept Indexing of Medical DocumentsA Quantitative Study Using the UMLS*, Journal of the American Medical Informatics Association 2001.

Nadkarni, P., R. Chen, et al. (2001). *UMLS Concept Indexing for Production Databases*. Journal of the American Medical Informatics Association 8: 80-91.

Reynar, J. C. and A. Ratnaparkhi (1997). *A maximum entropy approach to identifying sentence boundaries.* Proceedings of the Fifth Conference on Applied Natural Language Processing: 16–19.

SNOMED International. (2006). *SNOMED International: The Systematized Nomenclature of Medicine*. http://www.snomed.org. Last accessed 08-09-2006.

Stearns, M. Q., C. Price, et al. (2001). *SNOMED clinical terms: overview of the development process and project status.* Proc AMIA Symp 662(6). Association for Computing Machinery. 1983. *Computing Reviews*, 24(11):503-512.

Tsuruoka, Y., Y. Tateishi, et al. *Developing a robust part-of-speech tagger for biomedical text*. Lecture notes in computer science: 382-392.

Zou, Q., W. W. Chu, et al. (2003). *IndexFinder: A Method of Extracting Key Concepts from Clinical Texts for Indexing*. Proc AMIA Symp 763: 7.