Classification of Semantic Relations in Different Syntactic Structures in Medical Text using the MeSH Hierarchy

by

Neha Bhooshan

Submitted to the Department of Electrical Engineering and Computer Science in Partial Fulfillment of the Requirements for the Degrees of Bachelor of Science in Computer Science and Engineering and Master of Engineering in Electrical and Computer Science at the Massachusetts Institute of Technology

February 4, 2005

Author____
Department of Electrical Engineering and Computer Science
February 4, 2005

Certified by____
Peter Szolovits
Thesis Supervisor

Accepted by____
Arthur C. Smith
Chairman, Department Committee on Graduate Theses

Classification of Semantic Relations in Different Syntactic Structures in Medical Text using the MeSH Hierarchy

by

Neha Bhooshan

Submitted to the Department of Electrical Engineering and Computer Science

February 4, 2005

In Partial Fulfillment of the Requirements for the Degrees of Bachelor of Science in Computer Science and Engineering and Master of Engineering in Electrical and Computer Science at the Massachusetts Institute of Technology

ABSTRACT

Two different classification algorithms are evaluated in recognizing semantic relationships of different syntatic compounds. The compounds, which include noun-noun, adjective-noun, noun-adjective, noun-verb, and verb-noun, were extracted from a set of doctors' notes using a part of speech tagger and a parser. Each compound was labeled with a semantic relationship, and each word in the compound was mapped to its corresponding entry in the MeSH hierarchy. MeSH includes only medical terminology so it was extended to include everyday, non-medical terms. The two classification algorithms, neural networks and a classification tree, were trained and tested on the data set for each type of syntactic compound. Models representing different levels of MeSH were generated and fed into the neural networks. Both algorithms performed better than random guessing, and the classification tree performed better than the neural networks in predicting the semantic relationship between phrases from their syntactic structure.

Thesis Supervisor: Peter Szolovits
Title: Director, Clinical Decision-Making Group

Table of Contents

1 Introduction

Medical text parsing (editing, indexing, storing, and retrieving medical expressions within records) has been an important area of research in medical informatics as more and more healthcare institutions are using electronic medical records for patients. Most records currently have unstructured medical text ranging from doctor's notes and prescriptions to lab results and discharge summaries. The capability to extract the key concepts and relationships in the medical text will allow the system to properly grasp the content and knowledge embedded in the medical text. The information gathered can be used for data mining applications, organizing information, double-checking information, use in clinical research, etc.

Given medical text, a computer first has to be able to parse the text. This is mostly done with syntactic parsers as a way to break up the text into manageable chunks. This can be difficult given the non-grammatically correct nature of medical text. The computer then autocodes the sentence, which involves tagging different parts of the sentence with codes from a medical ontology. These codes can map to the definition of the word or its semantic type, thereby giving the computer the foundation for understanding what the text is about. Once the computer has the free text in a standardized representation, it can then determine relationships between different words and try to classify text based on the codes or relationships.

2 Problem Statement

The method for semantic mapping is motivated by work done by Barbara Rosario and Marti Hearst [1]. They developed a classification algorithm for identifying semantic relationships between two-word noun compounds. The medical ontology that they used was MeSH (Medical Subject Headings) lexicon hierarchy, which is part of UMLS. They looked specifically at biomedical article headings and mapped nouns to MeSH, taking advantage of MeSH's hierarchical structure for generalization across classes of nouns. They then used machine learning classification algorithms to classify relationships between two-word noun compounds and carry out semantic labeling. This thesis hopes to extend Rosario and Hearst's work in two directions: application to clinical data and extension to other syntactic structures.

Rosario collected noun-noun compounds from titles and abstracts found on MEDLINE. Since MeSH is indexed for MEDLINE, the compounds were probably technical in nature. In contrast, my corpus of data is a set of doctors' notes from a hospital's emergency department, and thereby clinical in nature.

Rosario also only looked at compounds with just two nouns. This thesis will explore if similar machine classification algorithms can be applied to other two-word compounds like adjective-noun compounds and more complex constituent groups like noun-verb phrases.

In order to accomplish these two tasks, the follow procedure was followed:

1. Extract different syntactic structures from medical text corpus using a parser
2. Label each phrase within a particular syntactic structure with a semantic relationship
3. Map each term in the phrase to its corresponding entry in the MeSH hierarchy
4. Create models that predict the semantic relationship between phrases from their syntactic structure, using different levels of the MeSH hierarchy
5. Train neural networks and a classification tree algorithm on the different models
6. Assess and compare performance of the models

3 Related Work

There have been a number of parsers developed to address the problem of information extraction for medical free-texts. MEDPARSE [2] is a pathology parsing project translator that maps concept terms in surgical pathology reports to SNOMED or UMLS. It uses Reverse Backus Naur Form (RBNF), which is a collection of sentence-construction rules. MEDPARSE starts with a large sentence and reduces it in steps until it reaches a null sentence by looking up each word in a lexicon table to determine its part of speech and UMLS identifier. Another type of parser, a Morpho-Semantic parser [3], decomposes each word in a sentence to single, elementary concepts that it represents.

The parser then uses this list of concepts and a set of pre-defined rules to build a parse tree of the sentence. Naomi Sager's Linguistic String Projec t[4] uses a sublanguage grammar and a word classification scheme tailored for the medical lexicon to extract and store information in a database. Carol Friedman [5] also developed a medical parser called MedLEE, which tokenizes, tags words and phrases with syntactic and semantic types, and then determines the structure of the sentence and interprets the relationships among the elements in the sentence.

4 UMLS

The Unified Medical Language System (UMLS) was used extensively in this project so a description of its structure and features is provided.

4.1 Overview

UMLS was implemented by the National Library of Medicine (NLM) to facilitate development of computer systems seeking to "understand" medical data and language [6]. There are three main components in the UMLS: Metathesaurus, Semantic Network, and SPECIALIST Lexicon. The Metathesaurus integrates over 100 different medical terminologies like ICD-10, which is International Statistical Classification of Diseases and Related Health Problems maintained by the World Health Organization, and SNOMED, which is the Systemized Nomenclature of Medicine database maintained by the American College of Physicians. Each of these vocabulary sources addresses a certain branch of medicine like Diagnostic and Statistical Manual of Mental Disorders (DSM) for mental health and Current Dental Terminology (CDT) for dentistry. The Metathesaurus serves to connect these heterogeneous databases together to create a single interface for the user. The basic unit of the Metathesaurus is the "concept"; all of the different names in the different vocabulary sources that mean the same thing are mapped to one concept. In addition to supplying unique identifiers, source vocabulary, and lexical variants for each concept, the Metathesaurus also assigns it one or more semantic types which are provided in the Semantic Network so that relationships between concepts can also be identified. There are currently 135 semantic types and 54 relationships. The SPECIALIST Lexicon contains the syntactic, morphologic, and orthographic information for each concept.

4.2 MeSH Hierarchy

MeSH, which stands for Medical Subject Headings, is one of the source vocabularies used in the UMLS and it is maintained by NLM. Its primary purpose is to support indexing, cataloguing, and retrieval of the 10 million plus medical literature articles stored in the NLM MEDLINE database [7]. After an article is submitted to MEDLINE, a NLM personnel reads the article and decides what descriptors to assign to the article to indicate what it is about to MEDLINE users.

MeSH is composed of three main units: descriptors, concepts, and terms. A descriptor class is a class of concepts that are closely related to each other. Each descriptor class has a preferred concept which best defines it and the preferred term for that preferred concept is the descriptor or heading. Each subordinate concept has a synonymous, broader, narrower, or related relationship to the preferred concept, and each one also has a preferred term. This means that synonymous concepts as well as concepts that differ slightly in meaning are grouped together under the same descriptor class. If a user wants to index an article with a particular term, he can look it up to see if it is a descriptor in which case it would be indexed directly with the descriptor. If the term is not a descriptor, it is an entry term of a subordinate concept listed under a descriptor. The article would then be indexed with that descriptor. The MeSH Browser is an online tool which looks up this information for users. For example, let's say that the user wants to index an article on *Pneumonitis*. The MeSH Browser would output:

```
MeSH Heading: Pneumonia
Tree Number: C08.381.677, C08.730.610
Entry Term: Experimental Lung Inflammation, Lung Inflammation,
Pneumonitis, Pulmonary Inflammation
```

In this case, all of the subordinate concepts are synonyms for the descriptor *Pneumonia*, and the article would be indexed with that descriptor.

The descriptor classes, also known as headings, make up the MeSH hierarchy. There are 15 high-level categories in MeSH: A for Anatomy, B for Organisms, C for Diseases, and so on. Each category is then divided into subcategories: A01 for Body Regions, A02 for Musculoskeletal System, A03 for Digestives System, and so on. Within each subcategory, descriptors are arranged hierarchically from general to specific in up to 11 levels. Each descriptor is given a unique MeSH tree number which represents its place in the hierarchy.

```
Cells;A11
   Antibody-Producing Cells;A11.063
      B-Lymphocytes;A11.063.438
         B-Lymphocyte Subsets;A11.063.438.450
         Plasma Cells;A11.063.438.725
      Antigen-Presenting Cells;A11.066
         Dendritic Cells;A11.066.270
            Langerhans Cells;A11.066.270.500
```

A descriptor can have more than one MeSH tree number. For example, *eye* has two MeSH tree numbers:

```
Body Regions;A01
   Head;A01.456
      Face;A01.456.505
         Eye;A01.456.505.420

Sense Organs;A09
   Eye;A09.371
```

There are currently approximately 43,000 MeSH headings in the MeSH hierarchy.

4.3 MetaMap Transfer

NLM implemented the MetaMap Transfer (MMTx) tool to facilitate usage of the UMLS [8]. The basic purpose of MMTx is to map text or find the best coverage to a UMLS concept. It can process a document or a single term. In processing a document, it first tokenizes the text with the Xerox part-of-speech tagger into noun phrases. It then generates variants using the SPECIALIST lexicon that includes derivational variants, acronyms, synonyms, abbreviations, and inflectional and spelling variants. Each variant is looked up in all of the Metathesaurus strings to retrieve its corresponding candidate. All the candidates are then evaluated using a weighted average measuring centrality, variation, coverage, and cohesiveness. The candidates are ordered based on their evaluation score. Mapping is completed upon joining all candidates of disjoint phrases into sentences and paragraphs, and the mapping with the highest score is MMTx's best interpretation of the document.

5 Data Processing

5.1 Corpus

The medical text corpus consists of approximately 2,000 discharge notes from the Emergency Department at Children's Hospital in Boston. The CHED notes are written by a physician when a patient first enters the hospital. The following is a de-identified note.

```
896776962
  . 15mo girl with vomiting for the past three hours.  Suddenly started
vomiting and briefly 'felt weak.' Dad called ambulance. Vomited X 7 total.  No
fever or diarrhea.
CURRENT MEDICATIONS:  - Synthroid.   PMH:  Medical History: Positive for
congential hypothyroidism.   ALLERGIES: No known medication allergies.
IMMUNIZATIONS: Up to date.
PE:   APPEARANCE: Active.  Alert.  Playful.  Smiling.  . VS: BP: 99/50.  HR:
122.  RR: 24.  Temp: 36.0 C.   HEAD: Atraumatic, normocephalic. EARS: Canals
clear.  TMs with nl light reflex and mobile. THROAT: No exudate, erythema or
tonsillar enlargement. CHEST: Lungs clear to auscultation. CARDIOVASCULAR:
Heart rate and rhythm regular.  Without murmurs. ABDOMEN: Bowel sounds normal.
Soft, nontender. No masses or  organomegaly.
DISPOSITION/PLAN: Discharged in good condition.
ASSESSMENT:  1.  Infectious gastroenteritis.  009.0.
CPT-4:
Level of service: 99283.  Status:  Urgent.
```

This example note begins with a description of what happened with general comments. It then lists medications, and previous medical history, and allergies. A condensed physical examination description follows covering basic organs. It ends with an assessment of the illness.

5.2 Parsing

5.2.1 Problem

Looking at the example discharge note in Figure 1, the lack of correct grammar is evident. The first sentence's verb, *vomit*, is in the wrong tense. Two sentences are missing subjects: *Suddenly started* and *Vomited*. The rest of the note has been structured with labels, but even the sentences used within are grammatically incorrect; the sentence, *TMs with nl light reflex and mobile*, is missing a verb. The combination of medical-specific terminology and incorrect grammar makes medical text difficult to parse.

I looked at three different parsers to determine the best way to extract the desired compounds: Link Grammar parser, Probabilistic Context-Free Grammar parser, and Bottom-up parser.

5.2.2 Link Grammar Parser

5.2.2.1 Overview

The Link Grammar parser is a syntactic parser using link grammar, and it was developed at Carnegie Mellon University [9]. Link grammar looks at a word as a block with connectors pointing to the right or to the left. A left-pointing connector connects with a right-pointing connector of the same type on another word. Words have rules about how their connectors can be connected up, i.e. rules about what would constitute a valid use of that word. A valid sentence is one in which all the words present are connected correctly and in accordance with the rules. The two global rules are that links cannot cross and that all words must be indirectly connected with each other.

The parser has a dictionary of about 60,000 word forms. A dictionary entry consists of a word and its connector assignment. In addition to the dictionary, there are word files, which contain words in a certain category (e.g. all proper nouns) and thus, have the same connector assignment. The following is an example of such a dictionary entry – word1 : A+ & B-. This means that 'word' must make an "A" link to the right and a "B" link to the left in a valid sentence.

There are 107 different kinds of connectors. For example, S connects subject-nouns to finite verbs and A connects pre-noun adjectives to nouns. Many categories have subscripts. For example, Ss denotes a singular subject relationship like "The dog is nice" while Sp denotes a plural subject relationship like "The dogs were nice". The broad range of categories allows the parser to be very specific on what word can connect with another word. The Parser can handle unknown vocabulary by making intelligent guesses from the open connectors of its surrounding words.

Once the parser has determined the set of valid links for words in the sentence, the post-processor divides the sentence into domains using the links and double-checks that the links are used correctly. The final output is the syntactic structure of the sentence as a set of labeled links connecting pairs of words. The constituent representation showing noun phrases, verb phrases, etc. can also be derived from this set of links.

5.2.2.2 Performance

When I used the original Link Parser on the discharge notes, it produced valid link structures for 9% out of 400 sentences. This was due to the fact that the Parser was designed to parse strictly grammatically correct sentences. However, as noted above, many sentences in medical text are either non-grammatically correct or just fragments. For example, the sentence "21 year old male injured his right knee" could not parse because there was no determiner "The" at the front of the sentence and the phrase "year old" was not recognized as an adjective phrase.

Prof. Szolovits added a medical lexicon to the Link Parser's dictionary, containing the connector assignments for medical terms. These assignments were mapped from the lexical definitions in the UMLS SPECIALIST lexicon [7]. The success rate did not improve significantly when the same 400 sentences were parsed using the Link Parser with the added medical lexicon. As mentioned earlier, the Link Parser can guess intelligently what type of speech an unknown word is by its placement within the sentence. So although the original Link Parser did not contain *hyperphagia* in its dictionary, it could infer from where it was placed that it was a noun.

In order to increase the parser's success rate, I relaxed connector assignments for different word groups and created new connectors to allow parsing of non-grammatically correct sentences. Going back to the first example, the noun "male" originally had a connector assignment of determiner connecter and subject connector. I modified the assignment so that the determiner connecter was optional. A second modification made "year old" an adjective phrase so then the sentence was parsed successfully. The modified Link Parser was able to produce valid link structures for 63% of the same 400 sentences from the discharge notes.

Although the link structures are valid, the generated constituent representation were increasingly inaccurate. This was due to the new links that I had created. I tried to modify the constituent post-processing code to ensure that the correct constituent representation is outputted, but was unable to do so successfully.

5.2.3 Probabilistic Context-Free Grammar Parser

I then looked at Probabilistic Context-Free parsers as another method for parsing. The Stanford Lexicalized PCFG parser does a product-of-experts model of plain PCFG parsing and lexicalized dependency parsing [10]. In principle, it is possible to add lexical entries by modifying the main lexicon file, but in practice it is not very feasible because it the lexicon is defined in terms of weighted rewrites which have undergone smoothing and renormalization. Grammar rules would be even more difficult to add by hand. Since a medical lexicon could not be added to the PCFG parser, the best choice seemed to pre-tag the medical texts with syntactic tags and then run the tagged text through the parser since the parser will honor the tag assignments.

5.2.3.1 Part of Speech Taggers

Pre-tagging is accomplished by a part of speech (POS) tagger which is a tool that assigns part of speech tags like noun or adverb to words. Some words can have different syntactic functionality since words have different syntactic categories in different contexts. For example, “sleep” can act as a noun as in “She needs some sleep” or as an infinitive verb as in “She needs to sleep”.

The Penn Treebank tag-set is a commonly used tag set that has 36 tags for different syntactic categories[11]. It differentiates between different verb tenses such as VBD for past tense verbs like “She was ill” and VBG for gerunds or past participles like “She is mowing the lawn”, and it includes tags for punctuation. The Penn Treebank project has two annotated corpora that were manually tagged to serve as lexical references: the Wall Street Journal corpus and the Brown corpus, which consists of 500 texts, each consisting of 2,000 words [12].

The BrillTagger is a tool that tags words according to their part of speech [13]. It uses the Penn Treebank syntax tags like NN for noun and JJ for adjective. In the first stage, each word in the text is assigned a tag in isolation to other words using the LEXICON file, which contains a lexical entry for every word. Each entry has an ordered list of tags with the most likely tag appearing first. The RULEFILE rules are then used to correct a tag by looking at the word’s closest neighbors and applying its rules. An unknown word is assumed to be a noun and a proper noun if it is capitalized, and RULEFILE is utilized to attempt to correct the tag. The CONTEXTUALRULEFILE file is used to improve accuracy. The BrillTagger’s LEXICON, RULEFILE, and CONTEXTUALRULEFILE were created using the Brown and Wall Street Journal (WSJ) corpus. Although these two corpora were huge and cover a wide range of English words, they do not cover medical terminology. Also, due to the irregular grammatical nature of medical text, the rules in RULEFILE and CONTEXTUALRULEFILE might not be able to tag the medical terms correctly. The result is that the BrillTagger is not well suited for tagging medical text.

To fix this problem, Jenny Shu and Margaret Douglas, two MIT graduate students, created a medical lexicon to be added to the BrillTagger’s lexicon to increase its knowledge of medical terms [14]. Their corpus was a set of 24 nursing notes with a total of 4,991 words which was manually tagged. They used the LRAGR table in the UMLS Specialist Lexicon which gives the syntactic category and tense/agreement information for a word. Since the UMLS Specialist Lexicon and the Penn Treebank use different categories, Shu and Douglass created a mapping to translate UMLS tags to Penn tags and counted the relative frequencies of each word in the tagged text to incorporate the ordering of the parts of speech for the corpus words. The performance was evaluated using the kappa metric. Kappa is the measurement of agreement between two observers taking into account agreement that could occur by chance (expected agreement); the higher the kappa, the better the tagger performed. Running the BrillTagger with the default files on the nursing notes corpus resulted in a kappa of 0.7363 while running the BrillTagger with the additional medical lexicon resulted in a kappa of 0.8839. They also ran compared the performance of a statistical tagger, a 2-stage and 3-stage hidden Markov model. Using the default lexicon, the statistical tagger's kappa was 0.8314 while using the additional medical lexicon, the performance improved to a kappa of 0.8648. Thus, I decided to use the BrillTagger with the added medical lexicon to tag my CHED data.

Tokenization was performed using a Perl script which separated punctuation from words. For example, "She is late." is tokenized to "She is late .". I then ran the BrillTagger on the tokenized CHED data. However, the CHED notes were not manually tagged so it was not possible to calculate the BrillTagger's accuracy in tagging every word.

I also came across the LT CHUNK tool, which is a syntactic chunker[15]. It uses the LT POS tagger[16]. The LT POS first tokenizes the text, breaking text into words and sentences. The tokens are then sent to a morphological classifier which looks up the token in its lexicon and assigns it all possible tags. Finally, the token-tags are sent to the morphological disambiguator which chooses a single tag based on the context. Once all of the words in the text have been tagged, LT CHUNK uses context-sensitive grammar rules to recognize simple noun and verb groups. For example, for the sentence, "The young boy has injured his right knee", it would output "[The young boy] <has injured> [his right knee]". However, there was not a way to add the medical lexicon to LT CHUNK's lexicon. Also it is a coarse parser in that it does not distinguish if an adjective is included in the noun phrase. Since I needed a finer parser, I did not use the LT CHUNK tool.

5.2.3.2 Performance

Once the medical text was tagged, it was run through the Stanford Lexicalized PCFG parser. The initial set of noun-noun compounds had 1,740 compounds of which 1,118 were correctly identified compounds, giving a precision of 64% . The incorrectly classified compounds included proper names of hospitals, abbreviations, and incorrect punctuation.

5.2.4 Bottom-up Parser

An alternative to the PCFG parser was a bottom-up parser. In order to extract compounds from the text, Rosario used a POS tagger and a program that recognized sequences of units tagged as nouns. I implemented a similar scheme in which noun, adjective, and verb phrases were found by looking at tags sequentially. These phrases were then grouped and the parser ran through the data again, this time looking for more complex structures like noun-verb and adjective-preposition-noun phrases. It did not use any grammatical rules.

The bottom-up parser generated 2,167 compounds of which 1,299 were correctly identified, yielding a precision of 62%. The incorrect compounds were misclassified due to the same problems mentioned above for the PCFG parser.

I decided to use the bottom-up parser since it was easier to extract different kinds of compounds like adjective-noun and adjective-noun-noun compounds; the PCFG parser would have required an extensive top-down parser since it outputted the sentence with all the tags and constituents marked. The performances were similar in terms of precision and the bottom-up parser yielded more noun-noun compounds.

Based on the number of compounds in each category, I chose the following categories: Noun-Noun, Adjective-Noun, Noun-Adjective, Noun-Verb, and Verb-Noun, Once I had my different syntactic compounds, the next step was annotating the words in the compound with their corresponding MESH code.

5.3 Annotation with MeSH

The MESH hierarchy has approximately 43,000 words, all of which are nouns. There are no adjectives or verbs listed. It also does not have medical abbreviations. This was problematic since clinical data uses all syntactical categories and medical abbreviations. More significantly, it uses more common, everyday terms than what are found in medical literature. In the 1299 noun-noun compounds that were extracted from the CHED corpus, there are 1241 unique nouns. Of these 1241 nouns, only 346 were present in the MESH hierarchy which left 895 nouns with no MESH mapping.

5.3.1 Lexical Variance

The nouns in the MeSH hierarchy are either in their singular or plural sense. For example, the hierarchy has Fractures, but not Fracture, and it has Ear, but not Ears. Out of the 895 unmapped nouns, 78 were different agreements of the corresponding descriptor. Rather than manually adding all the tenses of each noun, I used the Lexical Lookup feature in the MetaMap Transfer API. The Lexical Lookup is the feature that generates variants using the SPECIALIST lexicon to include derivational variants, acronyms, synonyms, abbreviations, and inflectional and spelling variants[8]. The morphological changes associated with singular/plural agreements (adding *–s* or *–es*) appear first followed by other variants. Thus, the lexical variants are looked up for each word, and the first word to have an entry in the MeSH hierarchy is used.

```
LexicalElement|lungs|
     Variant|0|lung|noun|0|3|0|4|0|0|0|1|i
     Variant|0|pulmonary|adj|0|3|0|4|0|0|0|2|s
     Variant|0|pneumonias|noun|0|3|0|4|0|0|0|6|ids
     Variant|0|pneumonia|noun|0|3|0|4|0|0|0|5|ds
     Variant|0|pneumonic|adj|0|3|0|4|0|0|0|2|s
     Variant|0|pulmonic|adj|0|3|0|4|0|0|0|2|s
     Variant|0|pul|adj|0|3|0|4|0|0|0|4|as
     Variant|0|pulmonary artery|noun|0|3|0|4|0|0|0|5|ds
     Variant|0|pulmonary arteries|noun|0|3|0|4|0|0|0|6|ids
     Variant|0|pneumoniae|noun|0|3|0|4|0|0|0|6|ids
```

The word “lungs” has 10 lexical variants. The first one is “lung” which can be found in MESH. So if “lungs” occurs in a note, it will automatically map to “lung” in MESH and not “pneumonia” which is also considered a lexical variant. The Lexical Lookup also proved helpful with adjectives derived from nouns. For example, *adenoidal* is derived from *adenoids* (which is present in MeSH) by appending the suffix like *–al* to the end of the noun.

```
Phrase: adenoidal
LexicalElement|adenoidal|
      Variant|0|adenoids|noun|0|8|0|9|0|0|0|3|d
      Variant|0|adenoid|noun|0|8|0|9|0|0|0|3|d
      Variant|0|adenoidal|adj|0|8|0|9|0|0|0|0|
```

However, the Lexical Lookup feature only seems useful for simple morphological changes. Although it also outputs synonyms or closely related words to the word which could be found in MeSH, sometimes it is the incorrect word.

```
LexicalElement|pulmonary|
      Variant|0|pulmonary|adj|0|8|0|9|0|0|0|0|
      Variant|0|pneumonias|noun|0|8|0|9|0|0|0|7|ssd
      Variant|0|pneumonic|adj|0|8|0|9|0|0|0|4|ss
      Variant|0|pneumonia|noun|0|8|0|9|0|0|0|7|ssd
      Variant|0|pulmonic|adj|0|8|0|9|0|0|0|4|ss
      Variant|0|pn|noun|0|8|0|9|0|0|0|9|ssda
      Variant|0|pul|adj|0|8|0|9|0|0|0|2|a
      Variant|0|pns|noun|0|8|0|9|0|0|0|9|ssda
      Variant|0|lungs|noun|0|8|0|9|0|0|0|2|s
      Variant|0|pas|noun|0|8|0|9|0|0|0|9|ssda
      Variant|0|pa|noun|0|8|0|9|0|0|0|9|ssda
      Variant|0|pneumoniae|noun|0|8|0|9|0|0|0|7|ssd
      Variant|0|lung|noun|0|8|0|9|0|0|0|2|s
```

The word *pulmonary*, which means related to the lung, also lists *lung* in its output but it also outputs *pneumonia* first so then *pulmonary* would be mapped to *pneumonia.* It is unclear if there is a reason behind the order of the lexical variants. In any case, the Lexical Lookup feature helps in finding MeSH descriptors that are morphologically different from the given word.

5.3.2 Extending MeSH

The rest of the words had to be manually added to MeSH either by mapping a word to already present MeSH descriptor or directly adding the word to MeSH underneath an already present MeSH descriptor or creating a new category. After adding unmapped nouns, adjectives, and verbs, the current extended MeSH hierarchy for nouns increased by 3.51%, from 42,610 entries to 44,105 entries. There are also 150* mappings.

5.3.2.1 Brand-Name Medications

There were also 32 unknown words that were brand names for drugs such as Vanceril and Atrovent. I looked up each drug to find its generic name and manually mapped the brand name to this generic name, which was found in MESH. For example, Vanceril is the brand name for Beclomethasone and Atrovent is the brand name for Ipratropium. I used the MeSH browser to look up the drugs.

5.3.2.2 Abbreviations

There are also 115 abbreviations in the set of 932 unmapped nouns. When I looked up "BP", a common abbreviation used to denote blood pressure in clinical notes, in the Lexical Lookup, it outputted:

```
LexicalElement|bp|
      Variant|0|bp|noun|0|1|0|2|0|0|0|0|
      Variant|0|bps|noun|0|1|0|2|0|0|0|1|i
      Variant|0|bereitschaftspotential|noun|0|1|0|2|0|0|0|2|e
      Variant|0|bereitschaftspotentials|noun|0|1|0|2|0|0|0|2|e
```

So it was not helpful in finding a mapping to MESH. I then tried using the MMTX processing module to determine if it could output the correct term. However it outputted blood pressure, arterial pressure, base pairing, and boiling point as all equally likely possibilities for *BP*. So I manually mapped each abbreviation to its most likely interpretation. Since the corpus was limited to clinical data, the meanings of the abbreviations were pretty straightforward. It would be rare for *BP* to stand for boiling point or base pairing in a doctor's note; its most likely interpretation is blood pressure. To find the meaning, I looked up the abbreviations at an online medical reference site called www.pharma-lexicon.com/ and also got assistance from Dr. Kenneth Mandl at

Boston's Children's Hospital. I then used the MeSH Browser to find its corresponding MeSH descriptor. For example, URI is an abbreviation for Upper Respiratory Infection. Inputting this into the MeSH Browser resulted in URI being mapped to Respiratory Tract Infection since Upper Respiratory Infection is a subordinate concept of Respiratory Tract Infection. However, if *URI* is looked up in the MeSH Browser, it does not have it. It did not have *orchiopexy*, which is the surgical fixation of a testis, as a descriptor nor as an entry term. So the coverage of the MeSH Browser was limited.

5.3.2.3 Medical terminology

There were even some medical terms that were just absent. Some words are probably considered too common place to be used in medical literature; *yeast* was not in MeSH but its genus name, *Saccharomyces*, was in MeSH. Another reason might be that there are not articles written on certain medical terms like *Knuckles* or *Eardrums*, thus they are not indexed and included in MeSH.

Another scenario was that although the term did occur in the MeSH hierarchy, it was always paired with another term. For example, *ectopia* is an abnormal position for an organ or body part. It is present in 18 entries where it is grouped with the particular body part where the abnormality occurs. For example, *Ectopic lentis* is ectopia in the eye and *Pregnancy, Ectopic* is one where the fertilized ovum is implanted outside the uterus. However, it does not have its own entry. Looking up *ectopia* in the MeSH browser only yielded pointers to those descriptors; it is not listed as a subordinate concept for either descriptor. Again, since the MeSH hierarchy is built off of the articles that it indexes, some medical terms can be overlooked and not included.

5.3.2.4 Common Words and Syntaxes

Common, everyday words like *assistance, center, uptake*, and *elevation* are not used to index medical literature. For the noun-noun compounds, after mapping the drug names and abbreviations, there were still 702 nouns that had no corresponding entry in MESH. Some words are present in MeSH; the word *tests* is part of 89 MeSH entries like

“Liver Function Tests” and “Breath Tests; every body part and disease has some kind of test. But the word itself does not have its own entry.

So these words were manually and directly added to MeSH. The medical terminology was relatively easy to add. The definition was looked up online and the word was placed using common sense. For example, *Eardrum* was placed under *Ear* and *Knuckle* was placed under *Finger Joint*. The MeSH Browser was useful in mapping a common term to the appropriate MeSH descriptor. For example, *OTC* stands for *over the counter*, meaning over the counter drugs. When given *over the counter*, the Browser pointed to as an entry term for *Non-Prescription Drugs* since *Over-the-Counter Drugs* is an entry term for that descriptor. For some words, I had to add a completely new branch. For example, words like *wall*, *duct*, *canal*, *flank*, and *extremity* were grouped as body structures and placed under a new descriptor, *Body Parts;A01.990*.

I also used the MRHIER table in the UMLS Metathesaurus. The MRHIER table gives the complete path from the concept to the top of the hierarchical context, thus providing a complete list of parents for the concept. It consists mainly of the SNOMED concepts. An alternative was the MRREL table; however that table includes concepts from all of the terminologies which makes it difficult to navigate and also only offers distance-1 relationships: immediate parents, children, and siblings. When I looked up *perforation*, as in *bowel perforation*, online, the definition was the act of perforating or punching holes through a material. That did not seem to make sense in the context. I looked up the parents of *perforation* in MRHIER which outputted the following:

```
WORD Perforating C0549099
PARENT LIST
Systematized Nomenclature of Medicine. 2nd ed.
Morphology Axis
Mechanical Abnormality
Miscellaneous Structural Abnormality
PARENT LIST
SNOMED CT Concept
Body structure
Morphologically altered structure
Morphologically abnormal structure
Mechanical abnormality
PARENT LIST
SNOMED CT Concept
Qualifier value
Additional values
Descriptors
```

```
Appearances
Causal appearances
```

It turned out that *perforation* is also an abnormal opening in a hollow organ, caused by a rupture or injury; this definition makes sense within the *bowel perforation* context. So I added *perforation* under the *Pathological Conditions, Anatomical* descriptor.

Figuring out how to add the non-medical words was more difficult. I wanted to keep the hierarchical structure of MeSH in organizing the non-medical words. I initially tried designing my hierarchy, but the number of words and possible categories was overwhelming. I then used WordNet, an online semantic network of words developed at the Cognitive Science Laboratory at Princeton University to better organize the words.

5.3.2.5 WordNet

WordNet consists of only nouns, verbs, adjectives, and adverbs. Words that refer to the same concept are grouped into synonym sets or synsets. These synsets are then connected to other synsets by semantic relationships. These semantic relationships differ somewhat among the syntactic categories.

WordNet addresses the polysemy of words (i.e. words have different meanings in different contexts) by providing the different “senses” of each word. For example, the noun *break* has 15 different senses including an interruption, a lucky break, geographical fault, and an rupture. It provides example sentences to make the meaning of each sense clear.

The WordNet 1.6 edition has about 94,000 noun forms, 10,000 verb forms, 20,000 adjective forms, and 4,500 adverb forms [17], and it is available for use online.

5.3.2.5.1 Noun Classification

The primary semantic relationship used for nouns is hyponymy or the is-a relationship. This results in a fine stratification of noun categories, up to 12 levels from general to very specific. The relationship is transitive. Another semantic relationship is meronym or the part-whole relationship with three semantic types: separable parts, members of a group, and substances. This relationship is not transitive. For example, *head* has the following hypernyms (*head* is a kind of…):

```
Sense 1
head -- (the upper part of the human or animal body ; "he turned his head")
      => external body part -- (any body part visible externally)
        => body part -- (any part of an organism such as an organ or extremity)
          => part, piece -- (a portion of a natural object; "he needed a piece of granite")
            => thing -- (a separate and self-contained entity)
              => entity -- (that which is perceived to have its own distinct existence)

Sense 4
head, chief, top dog -- (a person who is in charge; "the head of the whole operation")
      => leader -- (a person who rules or guides or inspires others)
        => person, individual -- (a human being; "there was too much for one person to do")
          => organism, being -- (a living thing that can develop) the ability to act or function independently)
            => living thing, animate thing -- (a living (or once living) entity)
              => object, physical object -- (a tangible and visible entity; "it was full of  other objects")
                => entity -- (that which is perceived  to have its own distinct existence)
```

It then has the following hyponyms (…is a kind of head):

```
Sense 1
head -- (the upper part of the human or animal body ; "he turned his head")
      => human head -- (the head of a human being)

Sense 4
head, chief, top dog -- (a person who is in charge; "the head of the whole operation")
      => administrator, executive -- (someone who manages a government agency or department)
      => department head -- (the head of a department)
      => don, father -- (the head of an organized crime family)
      => general, superior general -- (the head of a religious order or congregation)
      => general manager -- (the highest ranking manager)
      => head of household -- (the head of a household or family or tribe)
```

Finally, it has the following meronyms(parts of head):

```
Sense 1
head -- (the upper part of the human or animal body ; "he turned his head")
         HAS PART: muzzle -- (forward projecting part of the head of certain animals)
         HAS PART: ear -- (the sense organ for hearing and equilibrium)
         HAS PART: brain -- (that part of the central nervous system that includes the higher nervous centers)
         HAS PART: skull -- (the bony skeleton of the head of vertebrates)
         HAS PART: face -- (the front of the human head from the forehead to the chin and ear to ear)
         HAS PART: temple -- (the flat area on either side of the forehead; "the veins in his temple throbbed")
```

After looking over my set of unmapped words, I decided that I would add two more categories to the MeSH hierarchy; one branch for properties and one branch for activity. They seemed to provide the broadest fit for the words that needed to be mapped. I started off with the hyponyms for *property*.

```
property --
 (a basic or essential attribute shared by all members of a class; "a study of the physical properties of atomic
 particles")
      => connectivity -- (the property of being connected or the degree to which something has connections)
      => heredity, genetic endowment -- (the total of inherited attributes)
      => age -- (how long something has existed; "it was replaced because of its age")
      => manner, mode, style, fashion -(how something is done or how it happens; "her dignified manner")
      => constitution, composition, makeup -- (the way in which someone or something is composed)
      => consistency -- (the property of holding together and retaining its shape)
      => disposition -- (a natural or acquired habit or characteristic tendency in a person or thing)
      => tactile property, feel -- (a property perceived by touch)
      => visual property -- (an attribute of vision)
      => olfactory property, smell, aroma, odor, scent -- (any property detected by the olfactory system)
      => sound property -- (an attribute of sound)
      => taste property -- (a property appreciated via the sense of taste)
      => edibility, edibleness -- (the property of being fit to eat)
      => physical property -- (a property used to characterize physical objects)
      => chemical property -- (a property used to characterize materials in reactions)
      => sustainability -- (the property of being sustainable)
      => strength -- (the property of being physically or mentally strong; "fatigue sapped his strength")
      => concentration -- (the strength of a solution)
      => weakness -- (the property of lacking physical or mental strength; liability to failure under pressure)
      => temporal property -- (a property relating to time)
      => viability -- ((of living things) capable of normal growth and development)
      => spatial property, spatiality -- (any property relating to or occupying space)
      => magnitude -- (the property of relative size or extent)
      => size -- (the property resulting from being one of a series of graduated measurements ")
      => analyticity -- (the property of being analytic)
       => selectivity -- (the property of being selective)
```

I then looked up the hyponyms for each of the given properties like *magnitude*.

```
magnitude -- (the property of relative size or extent; "they tried to predict the magnitude of the explosion")
      => proportion, dimension -- (magnitude or extent; "a building of vast proportions")
      => order of magnitude -- (a degree in a continuum of size or quantity; "it was on the order of a mile")
      => dimension -- (the magnitude of something in a particular direction)
      => degree, grade, level --
 (a position on a scale of intensity or amount or quality; "a moderate degree of intelligence ")
      => degree -- (the seriousness of something; "murder in the second degree"; "a second degree burn")
      => amplitude -- (greatness of magnitude)
      => multiplicity -- (the property of being multiple)
      => size -- (the physical magnitude of something; "a wolf is about the size of a large dog")
```

=> size -- (a large magnitude; "he blanched when he saw the size of the bill")
=> bulk, mass, volume -- (the property of something that is great in magnitude)
=> muchness -- (greatness of quantity or measure or extent)
=> amount -- (how much of something is available; "an adequate amount of food for four people")
=> extent --
(the distance or area or volume over which something extends; "the vast extent of the desert"; "an orchard o
f considerable extent")

So phrases like *feeding volume*, *dehydration level*, and *stool consistency* could now be mapped to MeSH descriptors. Similarly, I looked up *activity* and used its hyponyms as starting points, including *change of state*, to further branch out the category.

change of state -- (the act of changing something into something different in essential characteristics)
=> aeration -- (the act of charging a liquid with a gas making it effervescent)
=> passage, transition -- (the act of passing from one state or place to the next)
=> meddling, tampering -- (the act of altering something secretly or improperly)
=> transfer, transference --
(the act of transferring something from one form to another; "the transfer of the music from record to tape suppressed much of the background noise")
=> termination, ending, conclusion -- (the act of ending something; "the termination of the agreement")
=> nullification, override --
(the act of nullifying; making null and void; counteracting or overriding the effect or force of something)
=> reversal --
(a change from one state to the opposite state; "there was a reversal of autonomic function")
=> beginning, start, commencement --
(the act of starting something; "he was responsible for the beginning of negotiations")
=> arousal, rousing -- (the act of arousing; "the purpose of art is the arousal of emotions")
=> cooking, cookery, preparation --
(the act of preparing something (as food) by the application of heat; "cooking can be a great art"; "people a
re needed who have experience in cookery"; "he left the preparation of meals to his wife")
=> seasoning -- (the act of adding a seasoning to food)
=> infusion --
(the act of infusing or introducing a certain modifying element or quality; "the team's continued success is attributable to a steady infusion of new talent")
=> improvement --
(the act of improving something; "their improvements increased the value of the property")
=> beautification -- (the act of making something more beautiful)
=> decoration -- (the act of decorating something (in the hope of making it more attractive))
=> worsening -- (changing something with the result that it becomes worse)
=> degradation, debasement -- (changing to a lower state (a less respected state))
=> change of color -- (an act that change the light that something reflects)
=> soiling, dirtying -- (the act of soiling something)
=> wetting -- (the act of making something wet)
=> chew, chewing, mastication --
(biting and grinding food in your mouth so it becomes soft enough to swallow)
=> defoliation -- (causing the leaves of trees and other plants to fall off (as by the use of chemicals))
=> specialization, specialization --
(the act of specializing; making something suitable for a special purpose)
=> spiritualization, spiritualization --
(the act of making something spiritual; infusing it with spiritual content)

5.3.2.5.2 Adjective Classification

WordNet organizes adjectives differently from nouns since adjectives seem to fall into synsets centered around two direct antonyms. For example, *short* and *tall* are direct antonyms of each other. Adjectives in the *short* synset include *chunky* and *pint-sized* while adjectives in the *tall* synset include *gangly* and *leggy*. Since these adjectives are semantically similar to their respective adjective, the synset adjectives also serve as indirect antonyms to each other: *chunky* versus *leggy*. Relational adjectives which are derived from nouns, i.e. abdominal from abdomen, are not organized into the direct antonym clusters; instead they point to the base noun form. Adverbs derive lexically from adjectives, i.e. add –ly, so they follow the adjective organization where possible.

Rather than following the WordNet's scheme of clustering adjectives into direct antonym pairs, I used the adjective classification scheme used in GermaNet[18], a derivative of WordNet for the German language. GermaNet uses a more hierarchical approach.

Class	Subclass
Perceptional	Lightness, Color, Sound, Taste, Smell, Temperature
Spatial	Dimensional, Directional, Localisational, Origin, Distribution, Form
Temporal-related	Temporal, Age, Habitual, Velocity
Motion	
Material-related	Material, Consistence, Ripe, Dampness, Purity, Gravity, Physical, Composition, State, Stability
Weather	
Body-related	Life, Constitution, Affliction, Desire, Appearance, State
Mood-related	Mood, Stimulus
Spirit-related	Intelligence, Knowledge, Language Characterizing
Behavior-related	Character, Behavior, Discipline, Skill, Relation, Inclination
Quantity-related	Number, Quantity, Costs, Return
Social-related	Social, Institutional, Religion, Region, Race
Relational	Certainty, Effectiveness, Comparison, Correlation, Completeness
Pertainyms	"derived from"

Table 1: GermaNet Adjective Classification Scheme

The GermaNet classification is similar to the *property* hierarchy and thus made it amenable to adding adjectives to the MeSH hierarchy.

5.3.2.5.3 Verb Classification

WordNet organizes verbs similarly to nouns in that they also have a hierarchy although it is flatter (up to 4 levels). The primary semantic relationship used is tropnymy, which relates two verbs in how one verb dictates the manner in which the other verb is acted out. For example, *jog*, *sprint*, *rush*, and *lope* are all troponyms of *run* since they characterize how fast the person is running. There is the hyponymy relationship. There is also opposition in which motion verbs like *rise* and *fall* oppose each other in direction, and entailment in which completing one action actually incorporates another action. For example, *snoring* entails *sleeping*, but entailment is only one-directional, meaning *sleeping* does not entail *snoring*.

I only used the hyponymy relationship to classify the verbs within the MeSH hierarchy and did not use the tropnymy, opposition, nor entailment relationships. In most cases, the verbs were added within the *activity* branch by being listed under their noun counterpart, i.e. *terminated* under *termination*.

5.4. Labeling Compounds

The next step in the data processing was labeling each compound with its appropriate semantic relationship.

5.4.1 Noun-Noun Relationships

Rosario started off with Levi's [19] noun compound relationships and Warren's [20] taxonomy and adapted the list so that relationships were better suited for medical data. So in addition to general classes like Cause, Purpose, and Location, Rosario also added specifically medical relationships like Defect and Person/center who treats. I used the same set of relationships to classify the Noun-Noun set. The initial set had 1299 compounds and 24 relationships and I narrowed it down to 1132 compounds and 18 relationships by removing relationships with a sample count of 20 or less.

Name	**Number**	**Rosario Examples**	**CHED Examples**
Activity/Physical Process	118	Bile delivery, virus reproduction	Seizure activity, scalp swelling
Ending	4	Migraine relief	Surgery closure
Beginning	8	Headache onset	Eczema flare
Change	28	Disease development	Disease progression
Cause(1-2)	99	Asthma hospitalization	Strep pneumonia
Characteristic	52	Cell immunity	Extremity strength
Physical Property	56	Blood pressure	Fracture line
Defect	32	Hormone deficiency	Joint pain, vascular anomaly
Physical Make Up	7	Blood plasma	Blood glucose
Subtype	16	Migraine headache	Hbv carrier
Person/center who treats	21	Children hospital	Dermatology clinic
Attribute of clinical Study	7	Biology analyses	Laboratory findings
Procedure	94	Brain biopsy	Head MRI
Frequency/time of	41	Headache interval	Morning stiffness
Measure of	42	Relief rate	Cell count
Instrument	104	Laser irradiation	Albuterol nebulizer
Object	97	Kidney transplant	Family history
Misuse	4	Drug abuse	Tylenol overdose
Purpose	75	Influenza treatment	Erythomycin ointment
Topic	57	Health education	Asthma instructions
Location	84	Brain artery	Ear canals
Modal	14	Emergency surgery	Home visits
Material	54	Aloe gel	Iron supplements
Defect in Location	185	Lung abscess	Shoulder bursitis

Table 2: Relationships for Noun-Noun Compounds

5.4.2 Adjective Relationships

Since the relationships between noun and adjective are similar to those between noun and noun[21], I started off with the same set of noun-noun relationships. However, I soon noticed that I was labeling the majority of compounds as either “Characteristic” or “Measure of”. Another problem was that there was no counterpart to “Defect” for compounds like “strong pulse” so they were then put under “Characteristic”. I looked back to Warren's set of adjective-noun relationships and found the relationship, Norm-Adherent, which seemed to work as the opposite to "Defect". I divided "Measure of" into Degree, Quantity, and Duration. To break up the "Characteristic" category, I created the "Result of examination" category for test results and "Strength of certainty". The set of 23 relationships of 1422 compounds was narrowed down to 13 relationships with 1297 compounds.

Name	**Number**	**CHED Examples**
Activity	15	Clonic movements, respiratory effort
Cause(1-2)	11	Croupy cough, erythematous papule
Change	15	Dietary changes
Characteristic	190	Anxious look, clear fluid, cold air
Defect	59	Poor growth, bad breath
Defect in location	105	Abdominal pain, cardiac problem
Frequency/Time of	177	Chronic aspiration, previous symptoms
Location	204	Distal femur, interphalangeal joints
Material	15	Chromic sutures
Measure of – Quantity/Amount	226	Few episodes, minimal distress
Measure of – Degree	68	Significant effusion, slight limp
Measure of – Duration	34	Persistent cough, new murmur
Norm-Adherent	90	Good hemostatis, normal tone
Object	9	Apneic episode
Person/center who treats	9	Pulmonary clinic
Physical Make up	13	Bloody vomiting, purulent fluid
Physical property	26	Occipital region
Procedure	33	Expiratory films, nonfocal exam
Purpose	13	Corrective surgery
Result of examination	31	Negative monospot, known fever
Strength of certainty	39	Possible ulcer, probable reflux
Subtype	9	Medical team
Topic	16	Clinical findings, skeletal survey

Table 3: Relationships for Adjective-Noun Compounds

The Noun-Adjective compound set was much smaller with only 162 compounds and 9 relationships. Because it was a small set, I kept all the compounds.

Name	**Number**	**CHED Examples**
Procedure	8	Culture done

Physical Property	14	Diapers wet
Result of examination	41	Ketones negative
Defect	20	TM abnormal
Frequency/time of	13	Motrin earlier
Location	11	Knees bilat (bilateral)
Norm-Adherent	46	Pulse regular
Material	5	Clotrimazole topical
Strength of certainty	4	Parents unsure

Table 4: Relationships for Noun-Adjective Compounds

5.4.3 Verb Relationships

Since I was unable to find any literature about verb-noun relationships, I decided to use Rosario's set of noun-noun relationships. However, as with the adjective classification, I found that most of the relationships were being classified under Activity/Physical Process. I looked to Longacre's verb semantic classes to expand the number of relationships and included Cognition, Sensation/Perception, Communication, and Possession. There were 820 Verb-Noun compounds and 501 Noun-Verb compounds with the same 11 relationships.

Name	**Number: V-N**	**Number: N-V**	**CHED Examples: V-N**	**CHED Examples: N-V**
Defect	84	20	Had heartburn	Patient vomited
Procedure	78	25	Sutured wound	Surgery performed
Possession	156	82	Took bottle	Toradol given
Cognition	39	54	Reviewed films	Benadryl detected
Sensation	124	69	Saw physician	Blood seen
Beginning	45	34	Started medications	Exacerbation started
End	31	24	Finished amoxicillin	Nosebleed resolved
Change	87	28	Decreased appetite	Secretions increased
Activity/Physical Process	93	60	Struck leg	Patient carried
Communication	59	81	Called ambulance	Parents asked
Frequency/time of	24	24	Continued pain	Headaches persisted

Table 5: Verb Relationships

The verb classification included an additional parameter – whether the noun acted on was the agent or object of the action. An example in the Noun-Verb set is *Physician noted*; *Physician* is the agent of the phrase. In contrast, *Penicillin* in *Penicillin prescribed* is the object of the action verb, *prescribed*. For the Verb-Noun set, all of the compounds had the noun acting as a object and unstated agent; for example, for the compound *Recommended antibiotics*, *antibiotics* is the object of the verb *recommended* and the implied agent, i.e. the doctor, is not stated. So the additional parameter was not used for the Verb-Noun set.

5.4.4 UMLS Classification

I also tried an alternative in which I used the UMLS Metathesaurus to find the corresponding concept and semantic type for each word and then looked up the semantic relationships between the semantic types as found in the Semantic Network. However, I ran into similar problems where common, non-medical terms like intake and maneuver are not in the Metathesaurus. Even quasi-medical terms like dorsiflexion and extensor were not in it. For terms that did have a corresponding UMLS concept, some of them were not designated to be in a semantic relationship although there was clearly one. For example, for the compound *allergy clinic, allergy* has the following semantic types, Finding and Pathologic Function, and *clinic* has the following semantic types, Manufactured Object and Health Care Related Organization. There is no semantic relationship between any of these types in the Semantic Network. However, according to Rosario's classification, the compound would be labeled as Person/center who treats. In another example, *cell* is assigned the Cell semantic type while *culture* is given the Idea or Concept, Qualitative Concept, and Laboratory Procedure semantic types. For the latter, UMLS also interprets *culture* in terms of anthropology or society. Nonetheless, again there is no relationship between the semantic types Cell and Laboratory Procedure. Rosario's classification would label it as Procedure. Although the UMLS is a helpful tool, in this particular aspect, it was not useful. However, rather than relying on the designated semantic relationships, an interesting future project might be to look at the designated semantic types of each word and determine if any information or patterns could be extracted from semantic type compounds.

5.5 Models for Neural Networks

Once the compounds were classified with their semantic relationships and each term could be found, directly or indirectly, in MeSH, the next step was to create the different models that would be used in the Neural Network classification. Each model represents a different level of the MeSH hierarchy so the models for the top 5 levels of MeSH were generated. A compound was represented by concatenating its terms' MeSH codes. For example, the MeSH code for *Measles* is C02.782.580.600.500.500 and the code for *Immunization* is G03.850.780.200.425. Each term actually has more than one unique MeSH code, but for this example, only one code is shown. The two MeSH codes are concatenated and the periods are removed to form a single input.

Model 2	C 02 G 03
Model 3	C 02 782 G 03 850
Model 4	C 02 782 580 G 03 850 780
Model 5	C 02 782 580 600 G 03 850 780 200
Model 6	C 02 782 580 600 500 G 03 850 780 200 425

Table 6: Model Generation

If the codes for a compound only extend to level 3 (i.e. G01.400.500), then the compound will only be included in Models 2, 3, and 4.

This single input for each model that corresponds to a compound then serves as an input node into the neural network for that model. The input layer consists of all the possible MeSH code concatentations for that level. For example, the input layer for Model 2 would contain A 01 B 01, G 03 G 04, and so on. The input nodes within each model are unique. The input vector of the compound is a sequence of values with 1 for the its corresponding input node and 0 for all the other input nodes. For example, if the input layer for Model 2 consisted of the following input nodes: A 01 B 01, G 03 G 04, and C 02 G 03, then the input vector for *measles immunization* is [0 0 1]. Each compound also has a target vector which is a sequence of values with 1 for its corresponding relationship and 0 for the others. So if there were 3 output nodes for the neural network, representing the set of 3 relationships, and *measles immunization* was labeled with relationships #1, then its target vector is [1 0 0]. The set of input nodes, the input vector for each compound, and the target vector for each compound is fed into the neural network for each compound.

6 Classification Results

Two methods for classifying the processed data were used: Neural Networks and Classification Trees. In order to provide a baseline for measurement, the accuracy from random guessing was calculated for each syntactic category:

	N-N	A-N	N-A	V-N	N-V
Guessing	0.16	0.17	0.28	0.19	0.16

Table 7: Guessing accuracy for Syntactic Categories

6.1 Neural Networks

Rosario used a feed-forward neural network with conjugate gradient descent to classify each model. The input layer for each model consisted of the set of input nodes. The hidden layer used the hyperbolic tangent function. The output layer represented the set of semantic relationships and it used a logistic sigmoid function to map the outputs in the range [0,1]. I used the Neural Network package in Matlab. Because Rosario does not report the number of nodes in the hidden layer, my neural networks were run several times with different numbers of nodes in the hidden layer. A range of 25-35 nodes with full convergence yielded the best results. The data set for each model was split into 60% training and 40% testing for each relation. Table 8 shows the accuracy for each model on the test set in each syntactic category.

	Rosario Results	N-N	A-N	N-A	V-N	N-V
Model 2	0.52	0.10	0.08	0.24	0.12	0.11
Model 3	0.58	0.39	0.40	0.55	0.43	0.48
Model 4	0.60	0.44	0.46	0.58	0.48	0.49
Model 5	0.60	0.34	0.23	0.21	0.26	0.23
Model 6	0.61	0.22	0.10	0.09	0.10	0.11

Table 8: Neural Networks Results

Although the neural networks did not achieve the same level of accuracy as Rosario, they did perform better than random guessing for each model and performed consistently among the different syntactic categories except for the Noun-Adjective compound set. The Noun-Adjective set performed quite well and is probably due to its small size (and thus small testing set) and its two distinct and dominant relationships: Norm-Adherent and Result of Examination. The Noun-Noun compound set performed worse than the Adjective and Verb compound sets. This is probably due to the fact that the Noun-Noun set had 1241 unique nouns as mentioned earlier. In contrast, the Adjective-Noun compound set (before narrowing the set) had 932 unique words and the Verb-Noun set had 756 words, of which approximately 120 were verbs. So the number of input nodes for the Noun-Noun compounds exceeded those of the Adjective and Verb compounds sets. Another possible factor is that within the Adjective and Verb compound sets, certain words had only one associated relationship. For example, the adjective "positive" was almost always designated with a Result of Examination relationship and thus had a better chance of being correctly classified. In contrast, the noun "heart" could have multiple relationships: Measure Of for "heart rate", Defect in Location for "heart arrhythmia", and Activity for "heart motion". Although there were nouns with this characteristic in the Noun-Noun set, there were a greater number of such words within the Adjective and Verb classes.

In general, Models 3 and 4 had the most number of input nodes so they seemed to perform the best with a range of 35%-46% accuracy. Models 5 and 6 have lower numbers because there were fewer compounds which had MeSH codes that extended beyond 4 or 5 levels. As mentioned above, almost 75% (895 out of 1241 words) had to be added to the MeSH hierarchy. I assigned most words a MeSH depth of only level 3 or less; this is especially true for words within the *property* and *activity* branches.

6.2 Classification Tree

A classification tree uses a property-testing algorithm to classify data [23]. The data space has a set of known properties and the tree is generated by testing each property to see if it separates the data space "the best", i.e. divides the data into subsets which are optimally homogenized, i.e. all samples have the same relationship. The heuristic to determine this is the Average Disorder Formula[23]:

$$\text{Average Disorder} = \sum_{b}\left(\frac{n_b}{n_t}\right)\times\left(\sum_{c}\left(-\frac{n_{bc}}{n_b}\right)\log_2\left(\frac{n_{bc}}{n_b}\right)\right)$$

where

n_b = number of samples in branch b
n_t = total number of samples in all branches
n_{bc} = total number of samples in branch b of class c

The equation gives a high number if the resulting subsets are highly inhomogeneous and gives a low number if the resulting subsets are highly homogenous. The property with the lowest average disorder is chosen then as the split and the algorithm continues by applying the leftover properties to the resulting subset. The process continues until all the resulting leaf nodes are homogenous.

Generating a classification tree for the MeSH-coded compounds required a different way of splitting up the set of properties since each compound has a right and left word and the following split follows the MeSH hierarchy. The first split is made between the Right and Left words to determine if classifying among the right words first gives more homogenous sets than classifying the left words. Let's say that the Left words are chosen. All of the top-level MeSH codes (i.e. A01, A02, A03, etc.) for the left words in all the compounds are then found and the average disorder of each code is calculated. Let's say that A01 is chosen. At the next step, the neighbors of A01 (the other top level codes like A02 and A03 which are on the left side) are placed on a held off list and the children of A01 are now investigated. The Right words set is still being tested as well. However, the subset of the Right words is narrowed to those who also have A01 as either its corresponding left word or parent of its corresponding left word. The average disorder is again calculated and will determine whether first looking at left A01 and then the right word of a compound or whether continuing downwards to a child of A01 will facilitate the compound's classification. Once a homogenous leaf node is found, its path is recorded and the program goes up one level to retrieve the most recent held off list and then looks at the rest of the codes again. The process continues until each leaf node is homogenous.

However, a problem emerges in that the tree is now over fitted to the data and might not do well on a set of new words. To make the tree more generalized, a modified version of Fisher's exact test [23] was implemented. Basically, the Fisher's test looks at the distribution of classes in an inhomogeneous leaf node to see if the observed value is significant. If a particular class constitutes 75% of the node, it is chosen as the significant class since guessing that class will be 75% accurate. Also, nodes with fewer than four samples were also left alone rather than being classified further. The generalized classification tree includes more noise in that there are some inhomogeneneous nodes, but it has the benefit of being able to be used on other sets of data.

	N-N	A-N	N-A	V-N	N-V
Original	0.58	0.59	0.65	0.61	0.63
Generalized	0.50	0.55	0.63	0.58	0.57

Table 9: Classification Tree Accuracy

The classification tree outperformed the neural networks by a difference of 15-20%. As with the neural networks, the Adjective and Verb sets performed better than the Noun-Noun set. The generalized classification tree performed slightly worse than the original classification tree since all words in the testing set were present in the training set and it allowed some noise into the tree to avoid overfitting. The difference between the generalization and original accuracy for the Noun-Noun was larger than that of the other categories. One possible reason for this is that the number of samples within each rule was smaller in the Noun-Noun compounds due to the larger number of unique nouns; thus the modified Fisher's exact test was used more often in the Noun-Noun set than in the Adjective and Verb sets.

6.3 Sources of Error

There are a number of possible sources of error. The MeSH hierarchy dictates how the words are processed by the classification algorithms since the algorithm sees each word as a MeSH code. The structure of the additional branches of the MeSH hierarchy – *property* and *activity* – might be a factor in the classification performance. An experienced taxonomist would be useful for that task. Another source of error was the lack of an iterative process in both creating new semantic relationships for the Adjective and Verb sets and also labeling each compound with its correct relationship. The process is a subjective one, and an iterative process, similar to what Rosario used, in which several people classify and check each other's labeling might provide better objectivity and consistency in labeling. An experienced linguist might also be helpful. The neural network configuration used was taken from Rosario's research, and it is possible that another configuration with more layers or a different training function might be better suited for this particular set of data.

7 Conclusion

This paper has shown the viability of using the MeSH hierarchy to help classify medical text. The MeSH hierarchy is a useful choice for classifying words because it inherently incorporates semantics into its hierarchy. The more top-level MeSH codes function as the semantic type while its children were the specific instances of that type; for example, *head* (A01.456), *back* (A01.176), and *pelvis* (A01.598) were all placed under *Body Regions* (A01). It has also shown that this process can be applied to different syntactic structures.

Once the medical data has been semantically labeled, different classification algorithms can be used to organize the data. The Neural Networks and the Classification Tree methods were both successful in using intelligence to classify the data to its specified semantic relationship. Both performed better than random guessing but the Classification Tree performed better than the Neural Networks. One possible reason why the Classification Tree performed better than the Neural Networks is due to its own inherent hierarchical nature which matched the hierarchical structure of MeSH. Also, since the Neural Networks only used one layer of hidden nodes, it was not able to distinguish sharper boundaries between different compounds and relationships.

In the future, these classification algorithms might be trained for application purposes within clinical information systems or electronic patient records systems. Another extension is to apply this process to compounds with more than two words and more complex syntactic structures.

A.1 References

1. Rosario and Hearst. Classifying the Semantic Relations in Noun Compounds via a Domain-Specific Lexical Hierarchy.
2. MEDPARSE homepage http://www.medparse.com
3. Baud, Lovis, Rassinoux, and Scherrer. Morpho-Semantic Parsing of Medical Expressions. *AMIA Annual Symposium* 1998.
4. Sager, Lyman, Bucknall, Nhan, and Tick. Natural Language Processing and the Representation of Clinical Data. *Journal of the American Medical Informatics Association.*
5. Friedman, C. Towards a Comprehensive Medical Language Processing System: Methods and Issues.
6. UMLS Documentation homepage http://www.nlm.nih.gov/research/umls/umlsdoc.html
7. Nelson, Stuart J.; Schopen, Michael; Savage, Allan G.; Schulman, Jacque-Lynne; Arluk, Natalie. The MeSH Translation Maintenance System: Structure, Interface Design, and Implementation. In: Fieschi, M. et al., editors. Proceedings of the 11th World Congress on Medical Informatics; 2004 Sep 7-11; San Francisco, CA. Amsterdam: IOS Press; 2004. pp. 67-69.
8. Aronson, Alan R. Effective Mapping of Biomedical Text to the UMLS Metathesaurus: The MetaMap Program. AMIA 2001 proceedings.
9. Link Grammar Parser homepage http:// www.link.cs.cmu.edu/link/
10. Stanford parser homepage http://www-nlp.stanford.edu/software/lex-parser.shtml
11. Penn Treebank homepage http://www.cis.upenn.edu/~treebank/
12. The Brown Corpus homepage http://clwww.essex.ac.uk/w3c/corpus_ling/content/corpora/list/private/brown/brown.html
13. Brill, E. Transformation-Based Error-Driven Learning and Natural Language Processing: A Case Study in Part of Speech Tagging. 1995 Assoc. for Computational Linguistics.
14. Shu, J. and Douglass, M. Information Extraction from Medical Patient Notes.
15. LT CHUNK homepage http://www.ltg.ed.ac.uk/software/chunk/
16. LT POS homepage http://www.ltg.ed.ac.uk/software/pos/
17. Fellbaum, C., (1998). A lexical database of English: The mother of all WordNets. Special issue of *Computers and the Humanities*, P. Vossen (ed.), 209-220, 1998.
18. Hundsnurscher, F. & J. Splett. Semantik der Adjektive im Deutschen:Analyse der semantischen Relationen. Westdeutscher Verlag, 1982.
19. Levi, J. The Syntax and Semantics of Complex Nominals. New York: Academic Press, 1978.
20. Warren, Beatrice. Semantic Patterns of Noun-Noun Compounds. Acta Universitatis Gothoburgensis.
21. Barker, Ken. Semi-Automaic Recognition of Semantic Relationships in english Technical Texts. PhD Thesis.
22. Longacre, R. E., *An anatomy of speech notions*, Peter de Ridder Press, 1976.
23. Winston, P. Artificial Intelligence. 3rd ed. Boston: Addison-Wesley, 1992.

A.2 Acknowledgements

I would like to thank my thesis advisor, Prof. Peter Szolovits, for having me as a graduate student, encouraging my interest in medical informatics, and giving cool history lessons. I also would like thank Fern DeOliveira and Stephanie Hope for their support and Jenny Shut for her help with the tagger.