

# Extracting Noun Phrases for all of MEDLINE

Nuala A. Bennett, Qin He, Kevin Powell, Bruce R. Schatz  
CANIS - Community Architectures for Network Information Systems  
Graduate School of Library and Information Science  
University of Illinois at Urbana-Champaign, Champaign, IL 61820  
Email: {nabennet, hqin, powell, schatz}@canis.uiuc.edu

## ABSTRACT

A natural language parser that could extract noun phrases for all medical texts would be of great utility in analyzing content for information retrieval. We discuss the extraction of noun phrases from MEDLINE, using a general parser not tuned specifically for any medical domain. The noun phrase extractor is made up of three modules: tokenization; part-of-speech tagging; noun phrase identification. Using our program, we extracted noun phrases from the entire MEDLINE collection, encompassing 9.3 million abstracts. Over 270 million noun phrases were generated, of which 45 million were unique. The quality of these phrases was evaluated by examining all phrases from a sample collection of abstracts. The precision and recall of the phrases from our general parser compared favorably with those from three other parsers we had previously evaluated. We are continuing to improve our parser and evaluate our claim that a generic parser can effectively extract all the different phrases across the entire medical literature.

## 1. BACKGROUND

The vast majority of natural language parsers used on medical texts have been fine-tuned in some way to handle texts relating to a specific subject area. Many parsers are tailored for specific sub-domains, such as radiology reports [1], or surgical operative reports [2]. Fine-tuning often involves the SPECIALIST module [3] of the Unified Medical Language System (UMLS), or SNOMED [4]. The SPECIALIST is a prototype system specifically designed for parsing and accessing biomedical text, while SNOMED is a "comprehensive multiaxial nomenclature created for indexing the entire medical record" [5].

In a sample group of disease-related labels [6], over 30% of the labels were not recognized by a UMLS-based parser, as the terms were either absent from UMLS or modifiers were not accepted by the parser. The UMLS lexicon can, of course, be augmented with new terms and thus, have its performance increase significantly. This may be a good local or

short-term solution to parsing medical texts, but in the long term, it will not be as desirable as using one generic parser, which would be able to capture the desired phrases of any medical text, regardless of their healthcare setting.

At the University of Illinois, the Interspace Research Project [7] is developing a prototype environment for semantic indexing of multimedia information in a testbed of real collections. As part of the Interspace Prototype and the Digital Libraries Initiative (DLI) project [8], we have been working specifically on noun phrase extraction. We have set out to develop a generic natural language parser that extracts noun phrases regardless of the domain of knowledge within which a particular document is situated.

In addition to a generic parser, as the Interspace and DLI projects concern large quantities of documents, we also require that the parser can efficiently extract noun phrases from large-scale databases of text. To date, we have tested developing versions of the noun phrase extractor on approximately 630,000 CancerLit abstracts from the National Cancer Institute, 2.6 million Compendex abstracts from Engineering Information, and 3 million INSPEC abstracts from the Institution of Electrical Engineers. Most recently, we have reached an even higher magnitude and have tested the noun phrase extractor on 9.3 million MEDLINE abstracts. This paper considers the algorithms of our noun phrase extractor and demonstrates its performance with an experiment on the entire MEDLINE collection.

## 2. NOUN PHRASE EXTRACTOR

The noun phrase extractor in our system, AZ Phraser [9], was developed in collaboration with our partners in the Artificial Intelligence (AI) Laboratory at the University of Arizona. It is based on a part-of-speech tagger, originally developed by Brill [10], and noun phrase identification rules from NPtool [11], a commercial noun phrase extractor. The AZ Phraser algorithm has three main steps: tokenization, part-of-speech tagging and noun phrase identification.

## 2.1. Tokenization

The goal of the tokenization process is to determine sentence boundaries, and to separate the text into a stream of individual tokens (words) by removing extraneous punctuation. Single spaces are used to delimit tokens, while double spaces delimit sentences. Document texts must be tokenized correctly in order for the noun phrase extractor to parse the text efficiently. The tokenizer has been additionally modified to take specialized nomenclature, such as "1,2-Dimethylhydrazine", which uses embedded punctuation, into account.

## 2.2. Part-of-Speech Tagging

The part-of-speech tagger is based on work from Brill, although the original code has since been modified considerably. Several parts of the tagger have been optimized for better performance, for example, by using data structures such as hash tables. The tagger is divided into two main phases of operation - lexical analysis, and contextual analysis.

Lexical analysis involves looking up each word in a lexicon. To ensure generality and domain independence of the noun phrase extraction, the lexicon mostly comprises the Wall Street Journal corpus and the Brown corpus, neither of which is specifically related to medicine. The lexicon contains all the possible parts of speech, such as noun, verb, or adjective, appropriate to each word contained therein. Each word (token) from the text document is first marked with all the parts of speech listed for that particular word in the lexicon. If a word does not appear in the lexicon, the tagger will default to mark it as an unknown noun. The tagger originally included several learning rules, but to ensure generality of our algorithm, they were not used in this instance. Using the lexical rules of the tagger, one part-of-speech is chosen, leaving each word marked with its "best guess" part-of-speech tag.

Using several contextual rules, the contextual analysis phase processes the text further to ensure that the part-of-speech tags are disambiguated. With this information, the tagger is able to determine the final part-of-speech tag for each word. Following the completion of the tagging process, the noun phrases will be identified.

## 2.3. Noun Phrase Identification

Noun phrases are extracted using a finite set of rules, composed of different sequences of part-of-speech tags. The noun phrase rules used are based on the

rules used by NPtool. For this particular experiment with MEDLINE, the limit for the longest recognizable noun phrase pattern was set to seven words in length, with the shortest pattern being obviously a noun phrase of length one, the single noun. The seven-word limit can lead to some error, as the tagger is likely to misidentify noun phrases longer than seven words as two completely separate noun phrases, which themselves may or may not be valid terms. An example is the sentence beginning, "We interpreted this finding as evidence of redistribution of blood flow in the lung...", which gives the noun phrase "finding as evidence of redistribution of blood", but consequently will lead to the phrase "blood flow" being lost.

The rules were applied to the tagged words from the text, using a sliding "window" of seven words. As the window slides over the words of the text, the noun phrase patterns are applied to the window contents. When encountered, the sentence delimiters will truncate the window. Since some of the rules are subsets of other rules, the longest matching rule is used to determine the "best" noun phrase. Once a noun phrase is located, the window will slide to the next word following the phrase and commence reading the contents of a new seven-word window.

## 2.4. Parser Evaluation

We undertook some preliminary evaluations of the quality of terms extracted using the AZ Phraser. In testing the noun phrase quality, we found that the AZ Phraser compared favorably with three other tested parsers [12]: FastNPE [13], which was the original parser used on our text systems, and which relies on concatenation of adjacent tokens to identify phrases; Chopper, developed by Haase at MIT, which will parse a text by breaking it down into constituent sentences or phrases; and NPtool, the commercial tool mentioned previously.

The preliminary tests were carried out on forty document abstracts. The noun phrases extracted by each parser were compared against manually pre-detected phrases in each document. Using well-known terms taken from information retrieval literature, the parsers were evaluated using recall and precision measures. Recall was defined to be the number of noun phrases correctly identified, divided by the total number of actual noun phrases manually identified by a human expert in the texts. Precision was taken to be the number of phrases correctly identified by the parser, divided by the total number of nouns identified by the parser.

	FastNPE	NPtool	Chopper	AZ Phraser
<b>Recall</b>	50%	95%	97%	92%
<b>Precision</b>	80%	96%	90%	86%

**Figure 1: Recall and precision results**

As shown in Figure 1, NPtool was found to perform the best overall with the test data. However, NPtool is a commercial system, and its source code is not available, although a binary version is sold (on a year by year basis) for research purposes. Similarly, the Chopper source code was not available for use at the time. The AZ Phraser showed great potential for use and development in generating good noun phrases with minimal noise in the output phrases. According to a separate study conducted by the Arizona AI Lab [14], a version of the AZ Phraser, when enhanced by the SPECIALIST lexicon, did perform slightly better than the generic version. This study was on a collection of 630K CancerLit abstracts, but the difference was not found to be statistically significant. We therefore decided to proceed with using the AZ Phraser on MEDLINE.

### 3. PROCESSING MEDLINE

MEDLINE is the premier bibliographic database of the National Library of Medicine (NLM). It covers the fields of medicine, nursing, dentistry, veterinary medicine, the health care system, and the pre-clinical sciences. Records are made up of bibliographic citations and author abstracts from over 3,900 biomedical journals published in 70 different countries. In mid-1998, we ran the noun phrase extractor on the entire MEDLINE collection, 9.3M records from 1966 to December 1997.

#### 3.1. Computing Environment

This experiment was carried out on the SGI/CRAY Origin 2000 supercomputer at the National Center for Supercomputing Applications (NCSA) at the University of Illinois. The Origin 2000 is a scalable shared memory multiprocessor, which is designed to provide the benefits of both shared memory multiprocessor and distributed memory message-passing multiprocessor approaches. The data from the NLM was received on 125 tapes. In our experiment, we used 125 processors in a dedicated model and ran one noun phrase extractor with one processor on each of the 125 tapes. This resulted in the 125 tapes being processed in parallel to one another, with the noun phrase extractor running simultaneously on 125 processors.

#### 3.2. The Extraction Process

Using a separate pre-processing program, the MEDLINE data was first converted into the XML format, which is currently the most acceptable format for the noun phrase extractor. Once the pre-processing was complete, the noun phrase extractor was called to extract the noun phrases from the 125 XML files corresponding to the 125 tapes. A segment of the final output from the noun phrase extractor is shown in Figure 2. In each line of the Figure, the first number is the document ID. The second number is the frequency of the noun phrase within the current document, and the third is the number of words in the current noun phrase. The fourth number designates the field from which the noun phrase was extracted. In this experiment, we chose to only extract the noun phrases from four fields of each document - title, author, abstract and MeSH thesaurus terms. Finally, the words in each line make up the actual extracted noun phrase.

96037199 1 2 5 Aardema MJ
96037199 1 2 2 adult rat
96037199 1 3 9 alveolar epithelial cell
96037199 1 3 2 alveolar type II
96037199 2 4 9 alveolar type II cell
96037199 1 4 9 analysis of genomic DNA
96037199 1 1 9 antigen
96037199 1 2 9 antigen gene
96037199 1 2 6 Base Sequence
96037199 1 2 5 Burns JL
96037199 1 2 5 Carter JM

**Figure 2. Noun phrase extractor sample output**

### 4. RESULTS

There were 9,315,615 documents, or records, extracted from the 125 MEDLINE tapes. Of these, 142 records were so short that no noun phrases were found in the four specified fields. Consequently, 270,729,881 noun phrases were extracted from 9,315,473 records. The total number of unique noun phrases generated was 45,449,799, of which 18,486 were MeSH thesaurus terms.

The next stage after the noun phrase extraction was to create concept spaces and category maps to serve as a large testbed for users. These semantic indexes have been incorporated into the Interspace Prototype [15]. Our physician collaborators are currently evaluating the usefulness of the phrases extracted from MEDLINE, using a Web interface to the research prototype of the Interspace. Thus far, the reaction of the participants has been highly positive [16]. Our immediate goal is to do more in-depth evaluation of the noun phrase quality on the much larger dataset generated by working with MEDLINE, and working with a wider variety of users, including librarians, physicians, and medical students. More thorough validation tests, using coarse-grain instrument and fine-grain interview techniques from our DLI project [17] to validate the quality of the extracted phrases, will be used.

#### 4.1. Time and Memory

The time usage (h:m:s) of the noun phrase extractor for the MEDLINE tapes is as follows:

Maximum:	2:34:44.50
Minimum:	0:00:59.27
Average:	1:44:22:94

These differences are due to the 125 tapes varying in size from 1,172,012 to 145,111,036 bytes. The number of records per tape ranged from 710 to 125,000. The raw data in each record also varied from 256 to 6020 bytes. Only 4,489,262 records (48.19%) were found to have an abstract, differing in size from 54 to 4,299 words per abstract. 99.97% of the records were found to have MeSH terms, ranging from 1 to 60 terms per record.

At the beginning of the process, all lexicon rules are read into and retained in the memory. A memory space large enough for the longest record is also allocated at the same time. The records are then pre-allocated space and processed one by one, thus making the total memory space used by the program constant, being roughly equal to the sum of the lexicon size, rules size, and maximum record size.

#### 5. FUTURE WORK

There are several limitations in the current implementation of the noun phrase extractor. One major drawback is the current seven-word limitation on the length of noun phrase. A possible way to remove this problem is to generate a dynamic noun phrase length so that there are no limits imposed. The current method of noun phrase identification is

computationally expensive when compared with other more ad hoc techniques that rely solely on tokenization and concatenation of adjacent tokens to identify phrases. In cases where the quality of noun phrases is not as important, a simple method, such as FastNPE, might be used.

We are also considering generation of noun phrase variants [18]. Jacquemin described similar work with French terms [19] where successful expansion and thus conflation of terms, can increase indexing coverage up to 30% with precision of 90% for correct identification of related terms. We are planning to start working with Jacquemin's system in the immediate future. Another consideration, since names are a major practical use of concept spaces, is to use a name finding system, such as Nymble [20], to enhance noun phrase parsing and aid in finding specific terms for effective search.

This MEDLINE experiment was carried out on a supercomputer at NCSA. As this approach is not feasible in every case, we have started investigating the feasibility of conducting a large-scale experiment on a group of PCs or workstations [21]. Our goal will be to prove that the large datasets can be processed efficiently in a smaller laboratory or community situation.

#### Acknowledgements

This work is supported by the DARPA Information Management Program Contract N66001-97-C-8535, and NSF/DARPA/NASA Digital Libraries Initiative Cooperative Agreement No. IRI 94-11318. We would like to particularly thank our colleagues at the University of Arizona, Kris Tolle and Dr Hsinchun Chen, without whom this work would not have been possible. We also wish to thank all those members of CANIS, who worked on this project.

#### References

1. Friedman, C., P.O. Alderson, J.H. Austin, J.J. Cimino, S.B. Johnson (1994), A General Natural-Language Text Processor for Clinical Radiology. *JAMIA*, 1(2):161-174.
2. Lamiell, J.M., Z.M. Wojcik, J. Isaacks (1993), Computer Auditing of Surgical Operative Reports Written in English. *Proc. Ann. Symp. Comp. Apps. Medical Care*, 269-273.
3. McCray, A.T. (1991), Extending a Natural Language Parser with UMLS Knowledge. *Proc. Ann Symp Comp. Apps. Medical Care*, 194-198.

4. Do Amaral Marcio B., Y. Satomura (1995), Associating Semantic Grammars with the SNOMED: Processing Medical Language and Representing Clinical Facts into a Language-Independent Frame, *MEDINFO*, 8(Pt 1):18-22.
5. SNOMED, <http://www.snomed.org>
6. Goldberg, H.S., C. Hsu, V. Law, C. Safran (1998), Validation of Clinical Problems Using a UMLS-Based Semantic Parser. *Proc AMIA Symposium*, pp. 805-809.
7. Schatz, B.R. (1997), Information Retrieval in Digital Libraries: Bringing Search to the Net. *Science*, Jan. 17, 275(5298):327-334.
8. Schatz, B.R., W.H. Mischo, T.W. Cole, et. al. (1999), Federated Search of Scientific Literature, *IEEE Computer*, 32(2), February, pp. 51-59.
9. Tolle, K.M., H. Chen, T. Ng (1999), Improving Concept Extraction from Text Using Noun Phrasing Tools: An Experiment in Medical Information Retrieval. Submitted to 4<sup>th</sup> Int. ACM Conf. on Digital Libraries (DL '99).
10. Brill, E. (1993), A Corpus-Based Approach to Language Learning. Ph.D. Thesis, Department of Computer and Information Science, University of Pennsylvania, Philadelphia, PA.
11. Voutilainen, A. (1993), NPtool: A Detector of English Noun Phrases. *Proc. Workshop on Very Large Corpora*, Columbus, OH, June 22.
12. Bennett, N.A., Q. He, C. Chang, B.R. Schatz (1998), Concept Extraction in the Interspace Prototype, *CANIS Technical Report*, <http://www.canis.uiuc.edu/interspace>.
13. Chen, H., T.D. Ng, J. Martinez, B.R. Schatz (1997), A Concept Space Approach to Addressing the Vocabulary Switching Problem in Scientific Information Retrieval: An Experiment on the Worm Community System. *JASIS*, 48(1), pp. 17-31.
14. Tolle, K.M., H. Chen (1999), Bridging the Medical Information Gap through the Use of UMLS-enhanced Medical Noun Phrasing. Submitted to *JASIS*.
15. Chung, Y., Q. He, K. Powell, B. Schatz (1999), Semantic Indexing for a Complete Subject Discipline, Submitted to 4<sup>th</sup> Int. ACM Conf. on Digital Libraries (DL '99).
16. Alper J. (1998) Taming MEDLINE with Concept Spaces, *Science*, 281(5384):1785, Sept 18.
17. Bishop, A.P. (1998), Digital Libraries and Knowledge Disaggregation: The Use of Journal Article Components, *Proc. 3<sup>rd</sup> Int. ACM Conf. on Digital Libraries*, pp. 29-39, Pittsburgh, PA.
18. Tzoukermann, E., J.L. Klavans, C. Jacquemin (1997), Effective Use of Natural Language Processing Techniques for Automatic Conflation of Multi-Word Terms: The Role of Derivational Morphology, Part of Speech Tagging, and Shallow Parsing. *ACM SIGIR Special Issue*, 148-155.
19. Jacquemin, C., J. Royaute (1994), Retrieving Terms and their Variants in a Lexicalized Unification-Based Framework. *Proc. 17<sup>th</sup> Ann. Int. ACM SIGIR Conf. on Research and Dev. in IR*, pp. 132-141.
20. Bikel, D., S. Miller, R. Schwartz, R. Weischedel (1997), Nymble: a High-Performance Learning Name-finder, *Proc. 5<sup>th</sup> Conf. Applied NLP*, pp. 194-201.
21. Chang, C., B.R. Schatz (1999), Concept Space Computation in a Distributed Analysis Environment. Submitted to 22<sup>nd</sup> Ann. Int. ACM SIGIR Conf. on Research and Dev. in IR.