Recognizing Noun Phrases in Medical Discharge Summaries: An Evaluation of Two Natural Language Parsers

Kent A. Spackman, M.D., Ph.D., William R. Hersh, M.D.
Biomedical Information Communication Center
Oregon Health Sciences University
Portland, OR

We evaluated the ability of two natural language parsers, CLARIT and the Xerox Tagger, to identify simple noun phrases in medical discharge summaries. In twenty randomly selected discharge summaries, there were 1909 unique simple noun phrases. CLARIT and the Xerox Tagger exactly identified 77.0% and 68.7% of the phrases, respectively, and partially identified 85.7% and 80.8% of the phrases. Neither system had been specially modified or tuned to the medical domain. These results suggest that it is possible to apply existing natural language processing (NLP) techniques to large bodies of medical text, in order to empirically identify the terminology used in medicine. Virtually all the noun phrases could be regarded as having special medical connotation and would be candidates for entry into a controlled medical vocabulary.

INTRODUCTION

There are many potential uses for automated systems that extract clinical findings from dictated medical narratives. Gathering data for outcomes research could become much less difficult and time-consuming. Automated quality assurance systems could perform concurrent evaluation of a patient's clinical course. In addition, automated extraction of new vocabulary could help to enhance and maintain standard clinical vocabulary systems such as SNOMED.

Computer understanding of natural language has long been a topic of active research.[1] Several researchers have investigated natural language processing (NLP) of medical narratives, and these studies have generally focused on building systems that take advantage of the relatively limited (when compared with all of the English language) vocabulary and structure of medical narratives.[2] Reports of the accuracy of such systems indicate that no-one has yet reached the goal of complete recognition of medical concepts using natural language processing.[3,4] Though imperfect, NLP systems have been shown to be of use for quality assurance systems and for answering questions from medical databases.[5]

Another important application of NLP methods may be the empirical extraction of phrases and concepts from large bodies of medical narrative. Collections of such concepts and phrases could be used to create and maintain large medical vocabularies. Even the most complete vocabulary, SNOMED[6], must continually be updated both to increase its coverage of medical concepts and to remain current.

Our goals in this study were to begin to evaluate the extent to which existing NLP programs could identify concepts in medical narrative. Since many of the important medical concepts in such narrative are expressed as noun phrases, we examined the proportion of noun phrases that were recognized by NLP analysis of discharge summaries.

METHODS

We randomly selected 20 discharge summaries of length greater than 200 words, from a corpus of 15,000 discharge summaries at Oregon Health Sciences University. We then manually marked each simple noun phrase in each discharge summary. A simple noun phrase is defined as a noun head with its modifiers. We excluded phrases that involved conjunctions or prepositions. Thus the phrase "history of deep venous thrombosis" would be split into two simple phrases, "history" and "deep venous thrombosis." We excluded from the analysis the names of patients, physicians, and proper place names. We also excluded isolated numbers and dates (e.g. "3/22/96"). For example, phrases such as "sodium 140" were marked as "sodium."

We evaluated two NLP systems: CLARIT[7] and the Xerox Part-of-Speech Tagger.[8] The CLARIT system is designed to quickly identify noun phrases in large

0195—4210/96/$5.00

bodies of text. It can identify simplex noun phrases consisting of the noun phrase head and its modifiers, or it can identify complex noun phrases. For the purposes of this experiment, we used only the simplex noun phrase option. The CLARIT parser was provided to us under a research agreement with Claritech Corporation, Pittsburgh PA. We used version 3.2, with its standard English lexicon dated 11/12/1995, running on a SUN Sparc workstation under SunOS v. 5.4.

The Xerox Part-of-Speech Tagger was obtained by FTP from parcftp.xerox.com in the directory pub/tagger. We used version 1.2 (tagger-1-2.tar.Z) running on a SUN Sparc workstation under SunOS v. 5.3. The Tagger requires a Common Lisp environment; we used CMU Common Lisp version 17f.

The Tagger can be modified in many ways. A finite-state automaton can be reprogrammed to break up text into tokens or words. The lexicon can be modified to include additional terms. In addition, the system uses probabilities in a hidden markov model that can be trained on a correctly-tagged corpus of text. We intend to modify the system for medical narrative in future studies; however, for this baseline analysis we used the "tag-english" system as distributed.

The Tagger default output simply marks the most likely part of speech for each word, using a lexicon and part-of-speech tags derived from the Brown corpus.[9] Within this tagging system, nouns are identified with the tags NN, NNS, NP, NPS, NR, and NRS, and adjectives are identified with the tags JJ, JJR, JJS, and JJT. In order to identify noun phrases, we added a 20-line Lisp routine that identifies continuous sequences of words that are tagged as nouns or adjectives and that end with a noun. An example of such a simple noun phrase is "aortic root aneurysmal dilatation."

We catalogued the list of unique noun phrases identified by manual markup of the discharge summaries. These were the "gold standard" against which the other two systems were compared. We then created a list of unique noun phrases identified by each NLP system.

Phrases identified manually were compared with those identified by each system using a strict lexical match after converting all phrases to lower case and eliminating punctuation. We also performed a manual match to determine "partial" identification of noun phrases; we counted the number of "gold standard" noun phrases partially matched by a NLP noun phrase. A partially matching phrase was defined as one with the same noun head but lacking some of the modifiers in the gold standard phrase. Extraneous words resulted in a failed match. For example, if the gold standard phrase was "pre-transplant anginal symptoms", "anginal symptoms" was counted as a partial match, but the phrases "anginal symptoms he" and "had pre-transplant anginal symptoms" were not counted as a match, since they misidentified the right and left boundaries, respectively.

We also examined the gold standard phrases that were not found by NLP methods and looked for systematic problems.

RESULTS

In twenty discharge summaries there were 3,111 lines of text and 15,908 words. We manually identified 1,909 unique simple noun phrases. The noun phrases identified by CLARIT included exact matches of 1,469 (77.0%) and partial matches of an additional 167 (8.7%) for a total of 1636 (85.7%). The noun phrases identified by the Xerox Tagger included exact matches of 1311 (68.7%) and partial matches of an additional 231 (12.1%) for a total of 1542 (80.8%). Combining output of the two NLP programs, 1638 (85.8%) of the noun phrases were matched exactly. Table 1 shows the number of phrases that were identified by each system, both systems, or neither system.

	#	%
Identified by both	1142	59.8
Xerox Tagger only	169	8.9
CLARIT only	327	17.1
Missed by both	271	14.2
Total Simple Noun Phrases	1909	100.0

Table 1. Number of Unique Noun Phrases Exactly Identified by Two NLP Programs

The proportions given in Table 1 correspond to the "Information Recall" (I-R) measure as defined by Sager.[2] "Information Precision," (I-P) on the other hand, is a way to express the proportion of total candidate noun phrases identified by each system that actually matched the gold standard. CLARIT

identified 2198 candidate noun phrases, for an I-P of 1469/2198=66.8%. The Xerox Tagger identified 1691 candidate noun phrases, giving an I-P of 1311/1691=77.5%.

In examining the phrases that were missed by the NLP methods, several repeated patterns emerged that may represent systematic problems. Some of these problems could be readily eliminated by simply "tuning" the systems to the domain of medical narratives.

Phrases that were missed were more likely to contain:

- misspelled words,
- hyphenated words used as adjectives (e.g. well-healed, side-to-side, wet-to-dry, crusted-over, follow-up),
- medical abbreviations, especially those associated with medication dosage and frequency (e.g. QD),
- adverbs modifying adjectives (e.g. totally normal coronary pattern),
- modifiers that may act as either a verb or an adjective (e.g. increased, increasing, improved, improving, etc.),
- references to dates, numbers, or quantities.

An open question is whether the noun phrases all have medical import, or are some of them simply verbal "chaff." To begin to answer this question, we used the SAPHIRE[10] engine to search for automatic matches for each of the noun phrases with a term or terms from SNOMED. SAPHIRE was able to find a single SNOMED term that matched 546 (28.6%) of the simple noun phrases; composite SNOMED terms were found that matched all or part of an additional 1079 (56.5%) of the phrases. Of the remaining 284 phrases, virtually all had medical import, and SNOMED terms could be found for most of these by manual lookup.

Based on these data, we believe it is unusual in medical narrative for even simple noun phrases to be "non-medical," that is, having no special meaning in the medical context.

Another open question is what proportion of all unique simple noun phrases would these NLP methods be able to identify in a very large corpus of text, rather than the very small corpus we analyzed.

The number of phrases in SNOMED is over 130,000, and many of these terms can be combined to form noun phrases. Thus the total number of simple noun phrases that can occur in dictated medical narratives is very large. Until the number of phrases extracted from a corpus approached that very large number, one would expect a roughly linear increase in new noun phrases identified per discharge summary (on average), and also a nearly linear increase in noun phrases that are missed by current NLP methods. Figure 1 presents our data that shows that this near-linear increase does occur, at least for a relatively small data set.

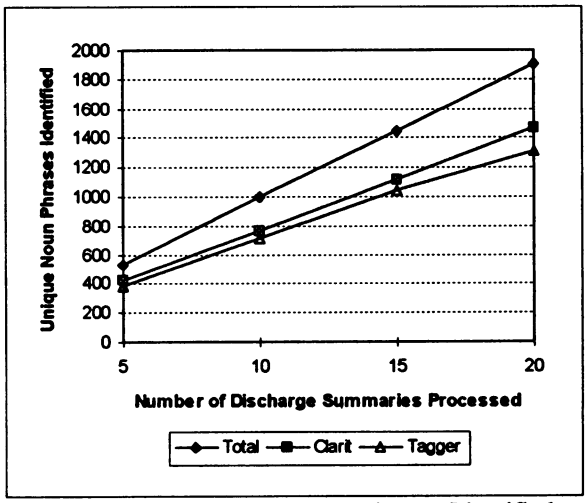

Figure 1: Total Unique Noun Phrases Identified After Processing N Discharge Summaries

DISCUSSION

Off-the-shelf systems designed to parse English sentences are quite readily available to the research community, and provide a convenient starting point for NLP research efforts. This study provides data that can serve as a baseline against which NLP systems can be compared. Tuning the off-the-shelf systems for special purposes in the medical domain may be more cost-effective than *de novo* creation of new parsers from scratch.

We expect that a modest amount of modification will enable a significant improvement in simple noun phrase identification. The failure analysis presented above suggests several relatively simple fixes that we estimate will increase the noun phrase identification rate from around 85% (partial matches) to above 95%.

The two systems that we tested illustrate the expected trade-off between the I-R and I-P measures; that is, CLARIT generated a larger number of candidate phrases and thus had higher recall and lower precision, while the Tagger generated fewer candidate phrases and had lower recall but higher precision.

Our results show that simple noun phrases are quite readily identified. Most parsers assign the "noun" part of speech to words that they cannot find in the lexicon, and this makes it more likely that such systems will correctly identify new terms such as drug names. On the other hand, many important parts of a complete concept lie outside the simple noun phrase. For example, the phrase "no evidence of sepsis" is not a simple noun phrase (by our definition) because of the preposition. It also involves negation. It is clear from this example that NLP systems need to recognize complex noun phrases, and to recognize the multiple ways that negation can be expressed.

Our study demonstrates that discharge summaries express many concepts as simple noun phrases. A question that remains to be answered is what proportion of medical concepts are expressed as noun phrases, and what proportion are expressed as verb phrases, adjective phrases, etc. NLP programs can be used to identify these other parts of speech as well.

CONCLUSION

We have evaluated the ability of existing natural language processing systems to identify simple noun phrases in medical narrative. Without special modification for the medical domain, two such systems were able to exactly identify a majority of the simple noun phrases in twenty discharge summaries. These findings indicate that readily-available NLP systems can be used to extract noun phrases from discharge summaries. Such automatic identification of noun phrases may be useful not just for building large vocabularies but also for deriving patient-specific information from medical narrative.

Acknowledgments

This research was supported in part by grant U01-LM05879 from the National Library of Medicine and grant FG06-994ER61918 from the U.S. Dept. of Energy. The authors thank Emily Campbell and Larry Donohoe for their assistance, and Alan Ertle for interesting discussions.

References

1. Joshi AK. Natural language processing. *Science*, 1991; 253:1242-1249.

2. Sager N, Lyman M, Bucknall C, Nhan N, Tick L. Natural language processing and the representation of clinical data. *JAMIA*, 1994; 1:142-160.

3. Haug PJ, Koehler S, Lau LM, et al. Experience with a mixed semantic/syntactic parser. In: *19th Annual Symposium on Computer Applications in Medical Care*, 1995, pp. 284-288.

4. Friedman C, Alderson PO, Austin JHM, et al. A general natural-language text processor for clinical radiology. *JAMIA*, 1994; 1:161-174.

5. Grishman R, Hirschman L. Question anwering from natural language medical databases. *Artificial Intelligence* 1978; 11:25-43.

6. Cote RA, Rothwell DJ, Beckett RS, Palotay JL. *SNOMED International*. College of American Pathologists, 1993.

7. Evans DA, Lefferts RG, Greffenstette G, et al. CLARIT TREC design, experiments, and results. Proceedings of the First Text REtrieveal Conference (TREC), Gaithersburg, MD, 251-286, 1992.

8. Cutting D, Kupiec J, Pedersen J and Sibun P. A practical part-of-speech tagger. In Proceedings of the Third Conference on Applied Natural Language Processing, Trento, Italy, April 1992. ACL. Also available as Xerox PARC technical report SSL-92-01.

9. Francis WN and Kucera F. *Frequency Analysis of English Usage*. Houghton Mifflin, 1982.

10. Hersh WR, Hickam DH. A performance and failure analysis of SAPHIRE with a MEDLINE test collection. *JAMIA* 1994; 1:51-60.