# An Evaluation of Natural Language Processing Methodologies

Carol Friedman[1,2], George Hripcsak[2], Irina Shablinsky[2]

[1] Computer Science Department, Queens College CUNY
[2] Department of Medical Informatics, Columbia University

*Medical language processing (MLP) systems that codify information in textual patient reports have been developed to help solve the data entry problem. Some systems have been evaluated in order to assess performance, but there has been little evaluation of the underlying technology. Various methodologies are used by the different MLP systems but a comparison of the methods has not been performed although evaluations of MLP methodologies would be extremely beneficial to the field. This paper describes a study that evaluates different techniques. To accomplish this task an existing MLP system MedLEE was modified and results from a previous study were used. Based on confidence intervals and differences in sensitivity and specificity between each technique and all the others combined, the results showed that the two methods based on obtaining the largest well-formed segment within a sentence had significantly higher sensitivity than the others by 5% and 6%. The method based on recognizing a complete sentence had a significantly worse sensitivity than the others by 7% and a better specificity by .2%. None of the methods had significantly worse specificity.*

## INTRODUCTION

Medical language processing (MLP) systems that extract and codify information in patient reports have been developed[1-5] in order to make clinical information available for a variety of applications. A number of the systems have been evaluated[6-9] in order to assess performance. However these evaluations shed little light on the underlying methodologies that were used. Because language processing is not yet well understood and because systems employ a range of techniques which vary widely in complexity, manageability, and performance, it is important to evaluate and compare the different techniques.

An evaluation that compares methodologies is extremely difficult to undertake for a number of reasons. One impediment is that no common set of reports is publicly available. If such a set were available, comparison of systems using different techniques would be possible provided that a clinical application and gold standard were also made available.

A second impediment is that it is very difficult for different groups of MLP researchers to collaborate in order to evaluate the methodologies they use because of the lack of a common data model and a common controlled vocabulary. A third difficulty is that evaluations are costly and time-consuming. Once an evaluation study is designed and a test set of documents chosen, a gold standard suitable for the test set has to be established. This typically involves recruiting and coordinating clinical experts to encode the test set.

In this paper we describe an evaluation that measures performance of different processing methods. The evaluation was feasible because it was based on modifications to existing work. Variations of techniques were developed by modifying an existing MLP system, called MedLEE[4], which has been operational at Columbia-Presbyterian Medical Center since 1995. In addition, a test set and gold standard from a previous study were used in order to reduce the effort.

## BACKGROUND

**Message Understanding Evaluations.** DARPA recognized the value of natural language processing (NLP) systems for general language extraction tasks and sponsored a series of Message Understanding Conferences (MUC)[10,11] in order to measure performance and portability of the different NLP systems. Each conference was associated with a specified task that consisted of extracting particular information from newspaper articles. In order to evaluate the systems, training and test sets were chosen, and human language analysts were recruited to establish the gold standards for the evaluations.

The results of the evaluations demonstrated that the NLP systems containing simpler pattern matching algorithms using limited linguistic knowledge performed very well compared to those containing

1091—8280/98/$5.00 

more complex linguistic knowledge. The evaluations brought about a better understanding of significant and practical issues associated with the specified language processing tasks.

It is very important to note that the type of texts used and the extraction tasks in the MUC evaluations were quite different from those likely to be needed in the clinical domain, and therefore the results may not be generalizable. The MUC tasks were geared toward extracting very limited information from general narrative text. They also involved extracting new names and locations which were frequently unknown to the language systems.

**Medical Language Systems.** A variety of techniques have been used by medical language processing systems. Some systems, such as the LSP system[1] and Ménélas[12] use comprehensive syntactic and semantic knowledge that involve knowledge about the structure of the complete sentence.

Other systems rely more heavily on semantic and local phrasal information. RECIT[5] uses syntax to recognize the structure of local phrases and interleaves phrase recognition with semantic knowledge in order to assemble semantically relevant groupings and representations from the phrases. MedLEE[4] relies heavily on general semantic patterns interleaved with some syntax, and also includes knowledge of the structure of the entire sentence.

SPRUS[3,13] was initially purely semantically driven. It was based on selecting relevant semantic frames associated with semantic information of the words in the sentence and expectations about findings, locations, and conditions. More recent versions integrated syntax into the processing.

Other MLP systems use methods that are based on pattern matching and keyword search. A more complete description of MLP systems is presented in Spyns[14].

**MedLEE's Techniques.** MedLEE tries to analyze the structure of the entire sentence using a grammar that consists of patterns of semantic and syntactic categories that are well-formed. For example, **finding in bodyloc conj bodyloc** is a well-formed pattern corresponding to sentences such as *pain in arms and legs*. If parsing fails, however, various error recovery modes are utilized in order to achieve robustness. The error recovery techniques use methods such as segmenting the sentence, processing large chunks of the sentence, and processing local segments. Each recovery technique is likely to increase sensitivity but at the expense of decreasing specificity and precision. When MedLEE processes a report, the most specific method is attempted first, and successive less specific methods are used only if needed.

There are five modes of processing as follows:

1. The initial segment is the entire sentence and all words or multi-word phrases in the segment must be defined. This mode requires a well-formed pattern for the complete segment.

2. The sentence is segmented at certain types of words or phrases (i.e. *consistent with*) and an attempt is made to recognize each segment independently. The process of segmenting is repeated until an analysis of each segment is obtained or until segmenting is no longer possible.

3. An attempt is made to identify a well-formed pattern for the largest prefix of the segment. This method is successful when the first part of a sentence contains a well-formed pattern but the end does not.

4. Undefined words are skipped and an analysis is attempted starting with mode 1.

5. The first word or phrase in the segment associated with a primary finding (i.e. *infiltrate*, *mastectomy*, *penicillin*). is identified; an attempt is made to recognize the part of the segment starting with the leftmost modifier of the finding. If no analysis is found, recognition is attempted again starting at the next modifier to the right. This process continues until an analysis is obtained. A modification of this process exists if the leftmost modifier is a negation, because negative terms may have to be distributed over all the segments. After a portion of a segment is successfully analyzed, the remaining portion is processed using the same method.

MedLEE was evaluated previously[8]. The study involved an application that utilized output created by MedLEE. Queries were written to automate the retrieval of reports associated with certain conditions, such as congestive heart failure and neoplasm. It was shown that the automated application was not significantly different from

clinicians in identifying reports associated with the specified conditions. In addition two keyword search techniques were used, but they were shown to perform significantly worse than the clinicians.

To determine a reference standard for the evaluation, twelve clinicians read the test set of reports and checked off zero or more conditions they felt were associated with each report. Two hundred reports were in the test set and each report was read by six clinicians. A condition was considered present in a report if four or more physicians checked the condition.

## METHODS

The two hundred reports that constituted the test set from the prior evaluation study described above was used for the test set in this study because a reference standard was already established for the set. Seven different language processing versions were created by modifying MedLEE in different ways. All the different versions used the same grammar and lexicon to process the test sentences but each version segmented the sentences differently and varied the treatment of words that were unknown to the system:

- V0 consisted of the regular version of MedLEE that tries all recovery modes successively if necessary. This version tests performance if successive methods that are less specific are used when previous methods have failed.
- V1 uses mode 1 only. This version tests performance if the entire sentence is parsed as a unit. This is likely to be the most specific version.
- V2 uses mode 2 only. This version tests performance if the sentence is broken up into smaller units based on predetermined semantic knowledge.
- V3 uses mode 3 only. It is likely be less specific than V1 because although the beginning of the sentence is well-formed, the end of the sentence may be skipped.
- V4 skips unknown words and uses V1.
- V5 skips unknown words and uses V2.
- V6 skips unknown words and uses V3.
- V7 uses mode 5 only and therefore tries to parse the largest segment surrounding a phrase whose semantic category corresponds to a primary finding. This mode is the closest to the technique that is based on processing local phrases, although an attempt is always made to find the largest well-formed segment possible, which could be the entire sentence.

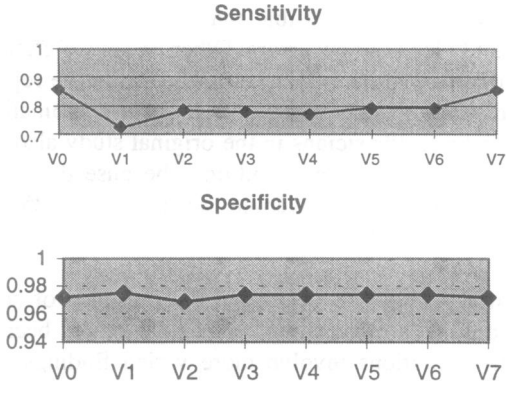
Each version was used to process the test set of reports. The queries from the previous evaluation were used to identify reports associated with the specified conditions. The results obtained were compared against the reference standard determined previously, and sensitivity and specificity measures were computed.

## RESULTS

Figure 1 shows the sensitivity and specificity of the different versions. In addition, confidence intervals and differences between each version and all the other versions combined were also computed. It is not surprising that V1, which is based on the recognition of a complete well-formed sentence, is the least sensitive and the most specific. However, the sensitivity of V1, .72, is significantly lower (by 7%) than the sensitivity of V0 and V7 which are both around .86.

In contrast, the specificity of the different versions varies only slightly. As expected, V1 has the highest specificity although it was significantly better than the other versions by only .2%. Versions V4-V6, which were based on ignoring undefined words, also showed an increase in sensitivity compared to V1 but only a very small decrease in specificity. None of the versions had a significantly worse specificity.

**Figure 1** Sensitivity and specificity of the different language processing versions.




A breakdown of sensitivity and specificity according to individual conditions was also performed. The

results showed that regardless of the version used, performance was better for some conditions than for others. For example, for all versions, the sensitivity for congestive heart failure ranged from .89 to .96 and the specificity was .99. In contrast the sensitivity for pneumonia ranged from .67 to .80 and the specificity from .96 to .97.

## DISCUSSION

It is noteworthy that the specificity of Version 7, which was based on recognizing segments within the sentence, was not significantly lower than that of the sentence based version V1. It appears that recognizing the largest well-formed segment surrounding a primary finding term (i.e. *opacity)* was highly effective, particularly since the sensitivity was increased substantially. Recognizing segments instead of the complete sentence did cause some loss of information. Typical errors in V2 and V7 occurred because modifier information in a portion of the sentence that was segmented was not distributed to the subsequent segments.

An informal analysis of the cause of the errors showed that frequently the errors were not attributable to the output generated by MedLEE. In a few cases, the errors were due to changes in the target vocabulary that were not incorporated into the queries. For example, the term *ill-defined density* was a new complex finding term generated by MedLEE that was not in the original query for neoplasm, and therefore caused a retrieval failure for three of the reports. This type of error is simple to correct, but it is possible that the correction could possibly lead to new errors. Other errors were due in a large part to insufficient information in the reports to distinguish conditions and atypical combinations of findings. Physicians in the original study also had trouble distinguishing conditions because they were shown to have an average sensitivity of .85 and specificity of .98.

Writing queries to determine the presence of some conditions is more complex than for others because some conditions involve more varied findings than others. An automated query can only have a fixed combination of findings and modifiers, and it is therefore impossible to capture all combinations that may actually occur. For example in this study, a false positive occurred because a finding *mass* was noted that made the condition neoplasm seem likely. Subsequently in the same report, another finding *lymph adenopathy* was noted that made the condition neoplasm seem unlikely. If the query were changed to handle this combination, a third finding will eventually occur in another report to make neoplasm seem likely even if *mass* and *lymph adenopathy* are both noted.

False negatives also occurred because of unusual finding-modifier combinations that were not reflected in the query. For example, one false negative occurred because the structured output for a report had a finding *density* with an atypical modifier *uncertain etiology* which was not included in the query.

V2, which always segmented the sentences, had the lowest specificity. This was because modifier information was lost using this method. To compensate for this, procedures would have to be written to recover important information such as body locations and temporal information. That would involve developing recovery algorithms for all the modifiers and therefore would incur additional overhead.

The specificity of versions V4-V6, which skipped unknown words, was not lowered much. This could be due to the fact that the lexicon for MedLEE is very well developed for chest radiological reports, and therefore unknown phrases that are clinically relevant are rare. The results could be considerably different for broader less well trained domains, such as discharge summaries because there are likely to be more words or phrases that are unknown to the system.

One limitation of this study is that the test set for the evaluation was biased because it was already used for a prior evaluation, and therefore the developer had the opportunity to correct processing errors for that set. However, the purpose of the study was not to measure the performance of MedLEE but to compare and analyze the various underlying techniques. A second limitation of the study is that the same grammar was used for parsing in all of the versions although the sentences were initially broken up differently. A third limitation is that in order to reduce the effort the developer, rather than an independent evaluator, analyzed the cause of the retrieval errors because the task required a deep knowledge of the components of the system.

We must be careful not to generalize this study. The results may be substantially different for another domain or for other applications. The structure of the language in radiological reports is much simpler, the

sentences are generally shorter, and the vocabulary more limited than for discharge summaries. These differences could have a substantial effect on the performance associated with the various methods.

## CONCLUSION

We have evaluated various MLP methods using results obtained from previous studies and modifications to an existing MLP system. The study was undertaken in an attempt to better understand language processing methodology and related issues. The results of our study demonstrated that the methods based on analysis of sentence segments rather than complete sentences showed substantial increases in sensitivity while incurring only a small loss in specificity. However, it will be important to perform other studies to see if the results are generalizable.

Informal analysis of the errors showed that most of the time the language processor encoded the information correctly. Errors occurred because the queries functioned with incomplete information and inconclusive findings. However, the medical experts also incurred errors, since their sensitivity and specificity measures (.85 and .98 respectively) were not significantly different.

## ACKNOWLEDGEMENTS

This publication was supported in part by grants LM06274 and LM05627 from the National Library of Medicine and by the Columbia CAT supported by the NYS Science and Technology Foundation.

## References

1. Sager N, Lyman M, Buchnall C, Nhan N, Tick L. Natural language processing and the representation of clinical data. JAMIA 1994;1(2) :42-160.

2. Zweigenbaum P, et al. An access system for medical records using natural language. Comput Meth Prog Bio 1994;45:117-120.

3. Haug P, Koehler S, Lau M, Wang P, and Rocha R. Experience with a mixed semantic/syntactic parser.In Gardner R.ed., *SCAMC 95.* Philadelphia. Hanley & Belfus. 1995; 284-288.

4. Friedman C, Alderson P, Austin J, Cimino J, Johnson S. A general natural language text processor for clinical radiology. JAMIA 1994;1(2):161-174.

5. Baud R, Rassinoux A, Scherrer J. Natural language processing and semantical representation of medical texts. Meth of Inf in Med 1992;31(2):117-125.

6. Lyman M, Sager N, TIck L, Nhan N, Borst F, Scherrer J. The application of natural-language processing to healthcare quality assessment. Med Decis Making 1991;11(suppl):S65-S68.

7. Gundersen M, Haug P, Pryor T, van Bree R, Koehler S, Bauer K, Clemons B. Development and evaluation of a computerized admission diagnoses encoding system. Comp and Biom Res 1996;29:351-372.

8. Hripcsak G, Friedman C, Alderson P, DuMouchel W, Johnson S, Clayton P. Unlocking clinical data from narrative reports. Ann of Int Med 1995;122(9):681-688.

9. Hripcsak G, Kuperman G, Friedman C. Extracting findings from narrative reports: software transferability and sources of physician disagreement. Meth Inform Med 1998;37:1-7.

10. Sundheim B. *Proceedings of the Fifth Message Understanding Conference (MUC-5).* San Mateo,CA. Morgan Kaufmann. 1994.

11. Sundheim B. *Proceedings of the Sixth Message Understanding Conference (MUC-6).* San Mateo, CA. Morgan Kaufmann. 1996.

12. Zweigenbaum P, Bachimont B, Bouaud J, Charlet J, and Boisvieux J. A multi-lingual architecture for building a normalized conceptual representation from medical language.In Gardner R.ed., *Proceedings of SCAMC 1995.* Phil. Hanley & Belfus.1995; 357-361.

13. Haug P, Ranum D, Frederick P. Computerized extraction of coded findings from free-text radiologic reports. Radiol 1990;174:543-548.

14. Spyns P. Natural language processing in medicine: An overview. Meth Inform in Med 1996;35:285-301.