# Towards a Comprehensive Medical Language Processing System: Methods and Issues

Carol Friedman, Ph.D.
Department of Computer Science, Queens College CUNY
Department of Medical Informatics, Columbia University

*Natural language processing (NLP) systems can help solve the data entry problem by providing coded data from textual reports for clinical applications. A number of NLP systems have shown promise, but have not yet achieved wide-spread use for practical applications. In order to achieve such use, a system must have broad coverage of the clinical domain and not be restricted to limited applications. In addition, an NLP system must perform satisfactorily for real-world applications. This paper describes methods and issues associated with an ongoing extension of MedLEE, an operational NLP system, from a limited domain to a domain that encompasses comprehensive clinical information.*

## INTRODUCTION

Natural language processing (NLP) systems have the potential to facilitate access to coded data by providing a method whereby clinical information in textual patient reports are automatically extracted, structured, and encoded. The information is then in a form that can be accessed reliably for applications such as decision support, literature search, ICD9 encoding, quality assurance, and outcomes analysis. Although a number of NLP systems have been developed within the medical domain[1-10], they have not yet achieved broad use for real-world applications. In order to achieve general use, an NLP system must be capable of processing patient reports from numerous domains, and also be robust, sensitive, and accurate enough for real-world applications. In addition, the target form must be relatively easy to access for different types of applications.

We have developed an NLP system, called MedLEE[5], that is operational at Columbia-Presbyterian Medical Center (CPMC), and is being used for decision support. Although MedLEE was developed as a general purpose medical language processor, it was initially applied to radiological reports of the chest. An independent evaluation of MedLEE was performed and the results demonstrated that it behaved similarly to physicians in interpreting x-ray reports to identify specified conditions[11,12].

A radiological report of the chest has a limited vocabulary, the language structures are primarily simple, and the formal representational model corresponds a small number of informational types and relations. Our aim was to first apply the methodology to a well-defined and restricted domain to see if performance was effective enough for a realistic application, and if it was, to incrementally extend the system to other domains until broad coverage of the clinical domain was achieved.

The first extension of MedLEE was to mammography reports. The task was relatively simple because of the language similarity between the two sub-specialties, and also because of the small vocabulary used in mammography reports. A second extension is currently being implemented to cover discharge summaries, which incorporate comprehensive types of clinical information. This task is considerably more complex and challenging. Discharge summaries are much more diverse than radiology reports: the vocabulary is much larger, the language structures are more varied, and there are many more types of information that have to be modeled.

In this paper we describe the methodology used to extend MedLEE, and discuss problems and issues related to the extension.

## BACKGROUND

MedLEE is composed of functionally different modular components where each component processes the text in some way and generates output used by subsequent components. Modularization simplifies management of the overall process. Ideally each module results in a further regularization of the text without significant loss of information. MedLEE is written in Quintus Prolog, and can support AIX, UNIX, and Windows platforms. Figure 1 is a diagram showing the different components, and a summarization is described below. A more detailed description can be found in Friedman and coauthors [5].

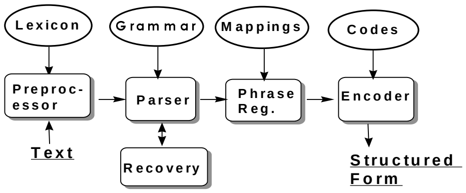

**Figure 1 - Overview of MedLEE**

The first component is the preprocessor, which reads the original text. It utilizes tokenization rules to determine word and sentence boundaries, resolve abbreviations, perform lexical lookup, and then generate output which consists of lists of sentences and corresponding lexical definitions. The lexical lookup phase finds definitions for words and phrases in the sentences, and is required for the parsing stage. A lexical entry specifies semantic or syntactic categories and canonical target forms. For example, the word *abdominal* is a body location category and the target form is **abdomen.**

The second component is the parser, which utilizes the lexical definitions and grammar to determine the structure of a sentence, interpret the relationships among the sentence elements, and generate an intermediate output form. The grammar specifies semantic and syntactic structures, which are interspersed, and also corresponding target forms. A parse is successful if the words and phrases of a sentence or fragment satisfy specified structures of the grammar. If there is no match, the recovery component is successively called to segment the sentence until partial parses are obtained. A parse is always obtained if the sentence contains semantically relevant information, but segmenting could result in lower accuracy.

The formal representational model for chest x-rays[13] was previously in the form of conceptual graphs (CGs)[14]. We currently use a frame-based representation that is consistent with the CG model. Each frame specifies the informational type, the value, and the modifier slots which are also frames. Thus, the output form for *severe pain in chest*, as shown below, is a frame denoting a clinical **problem**, which has the value **pain;** in addition, there are **degree** and **body location** modifiers with the values **severe** and **chest** respectively:

[**problem,pain,[degree,severe],[bodyloc,chest]]**.

The two components following the parsing stage are the phrase regularization component, which regularizes the intermediate target form further by composing multi-word terms which have been separated in the output, and the encoding component, which maps regularized target terms to controlled vocabulary concepts. This component utilizes a coding knowledge base to associate target terms to controlled vocabulary concepts[5,15]. The format has recently been modified so that the encoding can be fine-tuned to the domain. At CMPC, the controlled vocabulary is maintained by the Medical Entities Dictionary[16], but another knowledge base could be used to map to another controlled vocabulary. An example of the knowledge base is shown in Figure 2.

```
synonym(‘Paget’’s disease’,finding,’Paget’’s disease of
            breast’,finding,mammography).
synonym(‘Paget’’s disease’,finding,’osteitis deformens’
            ,cxr).
synonym(‘left lower lobe’,bodyloc,’left lower lobe of
            lung’,cxr).
```

**Figure 2. Encoding Knowledge Base**

According to Figure 2, the controlled vocabulary concept for the regularized term **Paget’s disease** is **Paget’s disease of breast** if the domain is mammography and **osteitis deformens** if the domain is chest x-ray (cxr) . Similarly, if the domain is cxr, **left lower lobe** will be encoded as **left lower lobe of lung**

After the encoding stage the output is in a form suitable for further processing. In the operational system at CPMC, this form is translated into an HL7 format and is uploaded to the CIS patient database[17]. In another application, the output may be uploaded to a relational research database.

## METHODS

The effort associated with extending MedLEE primarily involves the lexicon, grammar, representational model, and encoding knowledge base. The knowledge engineering task needed to establish the mappings from target terms to controlled concepts is substantial and requires knowledge of clinical terminology and ontology. There are important issues associated with the mapping to a controlled vocabulary, but this paper focuses on the three other components of the effort.

The task of extending a natural language processor may be very difficult or easy, depending on the new domain. An extension to a similar domain is generally simple (i.e. to another sub-domain within radiology such as abdomen, skull, or kidney). In that case, the effort mainly consists of adding new entries to the lexicon, and possibly adding some patterns to the grammar. Extension to a completely different domain is likely to require additional types of modifications, such as adding new semantic categories to the grammar and lexicon in order to classify and structure new types of information, adding new elements to the representational schema in order to model the appropriate output form, and adding new rules for recognizing sentence boundaries.

The first step in an extension involves collecting a training corpus for the new domain. For the domain of discharge summaries, we collected 5,500 reports of patients discharged at CPMC during a specified time period. Fifty sample reports were chosen for manual analysis to determine the adequacy of the semantic and syntactic categories; new categories were added when applicable. For example, some of the new types of information in discharge summaries were medications (c*oumadin*), laboratory procedures (c*hem7*), body measurements (*respiratory rate*), and behavioral information (*smokes cigarettes*).

When a new informational type is found, it generally means that the new informational type and associated modifiers have to be incorporated into the representational model. For example, medication information frequently occurs in discharge summaries, and therefore a medication frame containing new types of qualifiers, such as dose, duration, and frequency, must be modeled.

Another task associated with an extension consists of adjusting the algorithm for recognizing sentence boundaries because the new domain may contain abbreviations that are unknown to the system. This is accomplished by looking at occurrences of period ('.') to check for situations where a period does not signal the end of a sentence. For example, *coumadin 6 mg P.O. q.h.s. was started* corresponds to one sentence in which *P.O.* and *q.h.s.* are common abbreviations.

An extension also involves adding new entries to the lexicon. Single and multi-word phrases occurring in the new domain must be semantically categorized and their target forms specified. Identifying multi-word phrases is critical to accuracy and is facilitated by using a statistical tool which identifies candidate phrases automatically[18]. After multi-word phrases are added to the lexicon, single words which are not yet in the lexicon are automatically identified by scanning the corpus. When new entries are added, generally the most frequent words are added first. Specifying a new lexical entry is a straightforward task but requires domain knowledge and knowledge of the semantic categories.

A more complicated task consists of adding new semantic and syntactic rules to the grammar and specifying their target forms. This requires natural language processing expertise and is presently accomplished manually. Statistical methods are also used to simplify this task. The words and phrases in the training corpus are replaced by their semantic categories, and frequent semantic patterns are identified.

Subsequent steps in the extension consist of successive cycles of refinement where sample reports are processed and the system is adjusted. During each cycle, the output is analyzed, problems are identified, and the appropriate corrections are made. This process continues until satisfactory performance is achieved.

## RESULTS

When MedLEE was initially trained for radiological reports of the chest, the lexicon and grammar encompassed 30 semantic categories, 4 syntactic categories, and contained about 4,500 single and multi-word phrases. Remarkably, about one half of the lexical entries consisted of modifiers that were general across domains, such as *severe, possible*, and *consistent with*. The grammar contained about 450 rules. Extension of MedLEE to mammography involved adding about 250 new entries (i.e. *microcalcification*, *architectural distortion*) to the lexicon, and a few new patterns to the grammar. For example one new pattern was added to delineate the clock position of a finding (*lesion at 1:00 o'clock).* No new semantic categories were added and no extension of the representational model was necessary.

The extension to discharge summaries is still in progress. So far, we have added 23 new semantic categories (i.e. **laboratory procedure**, **medication**, **body measurement**) to the grammar and lexicon, 300 new rules to the grammar, and 6,000 new entries to the lexicon. These entries correspond to the most frequent lexical elements in the training corpus of discharge summaries. The lexicon will continue to increase substantially until it reaches a critical mass ranging from roughly 50,000 - 70,000 entries, but we have noted that the increase in the size of the grammar is significantly tapering off.

The representational model was extended to accommodate new informational types and their modifiers. Only a few new frames were developed for primary information such as medications, behavior, demographic information, body measurements, and their appropriate qualifiers. In addition, 10 new modifier types were designed for information such as frequency, duration, dose, age, ethnic background, and temporal information. Temporal information is much more complex and relevant in discharge summaries than in radiological reports. Some examples of temporal phrases that are represented are *on the morning of 3/6/1996, 2 days before last admission, 3 hours after fever*, and *between March 3rd and March 10th.*

Figure 3 below shows two sample output forms for information extracted from discharge summaries. The output has been simplified for demonstration purposes. The actual output form also includes contextual information, such as the parser recovery method that was used and the section of the report, because context is important for certain applications. The recovery method is associated with parsing accuracy (additional segmenting results in lower accuracy), and the section of the report occasionally affects the underlying meaning of the information. For example, *possible pneumonia* has a different meaning when it occurs in a clinical information section than when it occurs in an impression section.

In Figure 3, the first output frame corresponds to the output form for p*atient was discharged on coumadin 6mg P.O. q.h.s.* The frame corresponds to **med**ication information which has the value **coumadin**. In addition, there are modifiers **status**, **dose**, **manner**, and **frequency**, which further qualify **coumadin**. The second frame represents output for *patient experienced pain in chest 2 days before admission*. In this case the frame is a **problem** type with the value **pain** which is qualified by **certainty** with the value **high**, and **temporal** information consisting of a time point **admission**. The time point also has a relative time qualifier **reltime** specifying that the event occurred **before** the **admission** by **2 days**.

```
[med, coumadin,[status,discharge],  [dose, [unitval,
       [6,mg]]],[manner, po],[frequency, qhs]].
[problem,pain,[bodyloc,chest],[certainty,high],
       [timept,admission,[reltime, before, [timeunit,
       [2,day]].
```

**Figure 3. Output for New Types of Information**

## DISCUSSION

Since discharge summaries contain comprehensive information, it is practical to start with a meaningful subset, and then to extend the system incrementally. Because MedLEE was trained to capture the most frequent clinical information in discharge summaries, it presently can structure clinical information that is generally associated with the most prevalent health conditions, such as heart diseases, cancer, HIV infection, cerebrovascular diseases, diabetes mellitus, substance abuse, etc. It therefore provides access to crucial clinical data. In the near future, we will evaluate the system by applying a realistic application. If performance is satisfactory it will demonstrate that it is possible to extract limited but relevant information from a broad domain for practical use.

Generally, as the sensitivity of a system increases, accuracy decreases. The size of the lexicon should not adversely affect performance or manageability because each lexical entry is independent of the other entries. However, because most grammar rules are interdependent, the size of the grammar may have an adverse effect on performance. We plan on re-evaluating the performance of the extended system by applying it to chest x-ray reports in order to measure changes in performance. If the results demonstrate a considerable degradation in performance, it may imply that the only way to achieve adequate performance is to customize the grammar for a specialized domain. This would incur a considerable management overhead because different versions of grammars would have to be developed and maintained.

Manageability of one grammar is also a concern when the grammar is large because the rules are interdependent. Therefore a change in one rule may require changes to other rules. So far, the grammar has grown from 450 rules to 730 rules, and is still quite manageable. We have noted a substantial leveling off in the number of new rules. As long as the size of the grammar does not substantially increase, it should continue to be manageable.

Another area of concern is related to accessibility of the structured output. An analysis of a previous evaluation[11] demonstrated that a majority of errors in the application were attributable to the queries which accessed the output form and were not attributable to MedLEE. Although a simplified output form is generated in order to facilitate access, the queries still proved complicated to write. The errors due to the queries were basically caused by omission of three types of information: primary findings, unusual or unforeseen finding-modifier combinations, unusual or unforeseen combinations of findings.

It is reasonable to assume that the queries associated with discharge summaries will be more complex and therefore more error-prone because there are more informational types, more modifier types, and more values to consider. In particular, temporal information is more prevalent and more complex to use for an application. For example, a temporal reference may correspond to an incomplete date (i.e. *on the 1st*), an undefined time point (i.e. *after the flu*), or a time point modified by another unit of time (i.e. *2 days after last admission*).

## CONCLUSIONS

For a natural language processor to achieve broad use it must cover comprehensive clinical information and demonstrate effectiveness for practical clinical applications. We have described the method that was used to extend MedLEE from the domain of chest x-ray reports to the domain of discharge summaries. It will be important to evaluate the extended system by developing a realistic clinical application, and by measuring performance of both MedLEE and the overall application. If performance is satisfactory, it will demonstrate that natural language methodology can be used effectively for practical clinical applications.

**Acknowledgements**

This publication was supported in part by grants LM05397 from the National Library of Medicine and by the Columbia CAT in High Performance Computing and Communications in Healthcare supported by the New York State Science and Technology Foundation.

## References

1. Sager N, Lyman M, Nhan N, Tick L. Medical language processing: applications to patient data representation and automatic encoding. Meth Inform Med 1995;34:140-6.

2. Zweigenbaum P, et al. An access system for medical records using natural language. Comput Meth Prog Bio 1994;45:117-20.

3. Rassinoux AM, Wagner JC, Lovis C, et al. Analysis of medical texts based on a sound medical model. Proceedings of SCAMC 95, 1995, pp 27-31.

4. Baud R, Rassinoux A, Wagner J, Lovis C. Representing clinical narratives using conceptual graphs. Meth Inform Med 1995;1/2:176-86.

5. Friedman C, Hripcsak G, DuMouchel W, et al. Natural language processing in an operational clinical information system. J of Nat Lang Eng 1995;1(1):83-108.

6. Haug P, Ranum D, Frederick P. Computerized extraction of coded findings from free-text radiologic reports. Radiology 1990;174:543-8.

7. Pietrzyck P. A medical text analysis system for german syntax analysis. Meth Inform Med 1991;30:275-83.

8. Gabrieli E. Computer-assisted assessment of patient care in the hospital. Meth of Inform Med 1988;12(3):135-46.

9. Gundersen M, Haug P, Pryor T, et al. Development and evaluation of a computerized admission diagnoses encoding System. Computers and Biomedical Research 1996;29:351-72.

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

11. Hripcsak G, Friedman C, Alderson P, et al. Unlocking clinical data from narrative reports. Ann of Int Med 1995;122(9):681-8.

12. Jain NL, Knirsch CA, Friedman C, et al. Identification of suspected tuberculosis patients based on natural language processing of chest radiograph reports. Proceedings of the 1996 AMIA Annual Fall Symposium, 1996, pp 542-546.

13. Friedman C, Cimino JJ, Johnson SB. A schema for representing medical language applied to clinical radiology. JAMIA 1994;1(3):233-48.

14. Sowa J. Conceptual Structures: Information processing in mind and machine. Reading: Addison-Wesley; 1984.

15. Friedman C, Alderson P, Austin J, et al. A general natural language text processor for clinical radiology. JAMIA 1994;1(2):161-74.

16. Cimino J, Clayton P, Hripcsak G, Johnson S. Knowledge based approaches to the maintenance of a large controlled medical terminology. J Am Med Inf Assoc 1994;1:35-40.

17. Johnson SB, Friedman C, Cimino JJ, et al. Conceptual data model for a central patient database. In SCAMC 1992, p 381-385.

18. DuMouchel W, Friedman C, Hripcsak G, et al. Fisher D, Lenz H, editors.AI and Statistics. NY: Springer-Verlag; 1996;Two applications of statistical modeling to natural language processing. p. 413-21.