A Broad-Coverage Natural Language Processing System

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

Natural language processing systems (NLP) that extract clinical information from textual reports were shown to be effective for limited domains and for particular applications. Because an NLP system typically requires substantial resources to develop, it is beneficial if it is designed to be easily extendible to multiple domains and applications. This paper describes multiple extensions of an NLP system called MedLEE, which was originally developed for the domain of radiological reports of the chest, but has subsequently been extended to mammography, discharge summaries, all of radiology, electrocardiography, echocardiography, and pathology.

INTRODUCTION

NLP systems have been developed within the clinical domain[1-7] with the goal of enhancing the functionality of the Electronic Medical Record by providing comparable data for computerized applications, such as decision support, clinical research, automated encoding, quality assurance, outcomes analysis, and patient management. Most of the systems have been developed specifically for specialized applications and for limited domains. An NLP system typically requires a long development phase, and, thus, it is beneficial if the system can be extended incrementally, without undue effort, to achieve broad coverage. It is also advantageous if it can be used effectively for multiple applications.

We have developed an NLP system called MedLEE[3] that was originally designed for decision support applications in the domain of radiological reports of the chest (cxr). Subsequent extensions of the system to mammography and discharge summaries were performed that were shown to be effective [6,8]. We have since extended the system substantially to include all of radiology, electrocardiography, echocardiography, and pathology. We are also working on using MedLEE as a tool for a vocabulary development application[9] and for automated encoding of clinical information in text reports into ICD-9[10], SNOMED, or UMLS codes. These applications impose somewhat different requirements on the system than decision support applications.

In this paper we discuss extension of the system to the new domains and new applications. We also discuss some of the associated issues.

BACKGROUND

MedLEE is composed of functionally different modules where each module processes and transforms the text in accordance with a particular aspect of language. Each version of the transformed text is processed by a subsequent module until the final structured output form is obtained. Below is a brief overview of the system. A more detailed description was published previously [11, 12].

Figure 1 illustrates the original design along with subsequent additions, which are shown in gray. In this paper we focus on the first two components of MedLEE, and provide a summary of the others. The first component, the **preprocessor**, performs lexical lookup in order to recognize and categorize words and phrases. The preprocessor also identifies sentences and abbreviations using two sets of rules. As an example, the preprocessor will transform the sentence *possible left ventricular hypertrophy* into a list of words or phrases that are known to the system, as follows:**[possible,[left,ventricular,hypertrophy],.]** The phrase *left ventricular hypertrophy* is bracketed within the sentence list, signifying that it should be treated as an atomic phrase.

The **parser** uses a grammar to identify the structure of the sentence and to generate an intermediate structure that consists of primary findings and different types of modifiers. The grammar is a set of rules based on semantic and syntactic co-occurrence patterns. For the above example, output will be generated where the main finding is a problem **left ventricular hypertrophy** with a **certainty** modifier **possible.**

The **compositional regularizer** is used to compose individual words into phrases when applicable. It uses a table of structural mappings for this purpose. The **encoder** maps words and phrases into codes; it uses a table for the coding. The **recovery component** increases sensitivity by using alternative strategies to structure the text if the initial parsing effort fails.

An extension of MedLEE to the domain of discharge summaries was described previously [13]. The effort consisted of collecting a training corpus of 5,500 discharge summaries. Fifty reports were chosen randomly for manual analysis because the domain was much broader than cxr and mammography. New types of semantic categories, associated with the new types of information, were identified and added to the system. In addition, the representational model also was augmented in order to represent the new types of information and their relationships with other types. Another task consisted of identifying abbreviations and specifying their target output forms. The last task consisted of identifying new words and phrases, and adding them to the lexicon. This required that they be semantically and/or syntactically categorized and their target forms specified.

The system was refined iteratively. After the grammar, lexicon, and representational model were expanded, sample reports were processed and the output that was generated was manually analyzed. Based on problems that were identified, various components of the system were refined and the sample reports processed again until satisfactory results were obtained.

METHODS

The same basic method that was used to extend MedLEE to discharge summaries was also used for the incremental extensions to the domains reported in this paper. However, in order to minimize the manual effort, we initially extended the lexicon only. This approach seemed feasible because discharge summaries are so broad that we hypothesized that most of the work concerned with identifying new semantic categories, co-occurrence patterns, and components of the representational model would have been performed during the extension to discharge summaries, and therefore very few grammar changes would be needed for the subsequent extensions.

Training corpora of sample reports ranging from 5 - 6 megabytes were collected for each new domain. Since the domain of radiology is quite large, it was divided into sub-domains (i.e. abdomen, musculoskeletal, neurological, interventional, nuclear, computerized axial tomography, and magnetic resonance imaging). Dividing it up made the task more manageable and enabled us to quantify the amount of work required for each smaller sub-domain. A similar procedure was followed for echocardiography, electrocardiography, and pathology reports, but these domains were not subdivided.

The system was extended for one domain at a time before work began on the next domain. Working in parallel domains may have sped up the overall effort, but since there is a considerable amount of terminological overlap between domains a parallel effort would also have created extra work because then many of the same words and phrases would be identified and categorized in parallel.

A knowledge engineer, who was familiar with medical terminology, was trained to add terms to the lexicon. Tools that were developed previously were used to assist in lexical development. A statistical tool was used to suggest candidate multi-word terms. Appropriate terms were added to the lexicon based on manual review of the candidate terms. Single words that were unknown to the system were automatically identified and their frequencies calculated. Unknown words that occurred more than four times were listed along with sample sentences in order to provide contextual information. The knowledge engineer scanned the list of undefined words and contexts, and specified new entries for the lexicon.

Infrequently, a lexical entry could not be completed because it did not correspond to any of the semantic categories. When this occurred, a manual analysis of sample reports was performed, a new category was added to the grammar, and appropriate changes were made to the grammar and representational model.

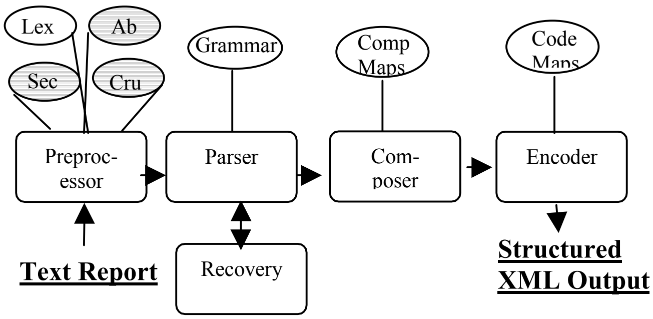

Figure 1- An overview of components of MedLEE. The oval components are knowledge bases; the other components are the programming engines. The preprocessor uses a lexicon (Lex), a list of abbreviations (Ab), a list of section names (Sec), and context rules for disambiguation. The grey ovals represent additions to the system for the new extensions.

RESULTS

The total number of entries in the lexicon increased from an initial amount of 4,500 entries covering chest radiological reports to a total of 15,307 entries covering findings in discharge summaries, radiology, electrocardiography, echocardiography, and pathology. The lexical entries were partitioned into two groups: domain specific (e.g. *infiltrate*, *lung*) and general English-like (e.g. *was*, *possible*, *moderate, due to*) in order to see if there were differences in the rate of growth between the two categories. The total number of domain specific entries and general English-like entries for the cxr domain were 2,300 and 2,200 respectively; after expansion to the new domains the number of domain specific entries increased substantially to 11,786 but the number of general English-like terms increased to only 3,521 entries. The biggest increase of entries for general English-like terms occurred when extending coverage to the discharge summary domain (i.e. 1,219 entries were added); the current extensions contributed only a total of 102 new entries of English-like terms.

Figure 2 shows the total number of domain specific entries after the system was extended for each new domain. Since there is considerable overlap of terminology among the domains, the lexicon is accumulative. In the new radiology domains, most of the entries that were added were specific body locations whereas in pathology most of the new entries were diagnoses.

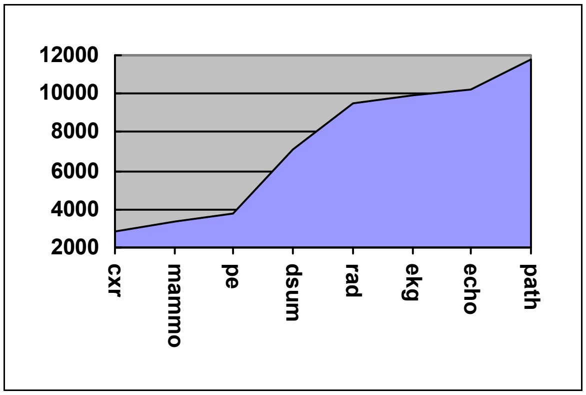

Figure 2. The total number of domain specific lexical entries in the lexicon after MedLEE was extended to each new domain. The lexical entries are accumulative because there is considerable overlap of the terminology used in the different domains.

Previously the number of grammar rules totaled 450 when the system covered only the cxr domain. The number was increased to 451 when it was extended to mammography and then to 730 for the extension to discharge summaries. The number of grammar rules now total 650, which is somewhat smaller than 730. This is due to a generalization that made the grammar simpler. Without the generalization the number of rules would have been approximately the same. Only a few new rules were added after the expansion to discharge summaries. These rules were aimed at improving specificity for the various domains.

Only three new semantic categories were added to the grammar. They correspond to new types of modifiers, such as diagnostic materials (i.e. *Indium 131*). Similarly, new frames were designed to represent the new types of information.

The extension for radiology took about one personyear whereas the extension to each of the other domains took about three months of one person's effort.

DISCUSSION

Once several extensions within radiology were performed, a manual review of sample output was obtained. After analyzing the output, we noted that a loss of accuracy occasionally occurred that was due to ambiguities from words and abbreviations that had multiple meanings. In order to improve the analysis, the preprocessing component was modified, and two new tables were added that were used by this component. One table consists of abbreviations where each abbreviation corresponds to a target form along with specific domains; abbreviations may also be specified as **general**, in which case they are applicable across clinical domains. This allows the preprocessor to choose the most specific abbreviation depending on the domain being processed.

The second table that was added consists of contextual disambiguation rules that are based on the local context of a word. For example, a rule was included to disambiguate *hr* depending on its local context (*hr* may mean: ***heart rate*** or **hour)**. Currently the disambiguation rules are manually created; it would be advantageous if automated machine learning techniques could be used to determine the appropriate word senses depending on local context. However, this would require substantial resources because training by an expert would be needed to determine and mark the correct senses from a training corpus.

A third table was also added for convenience. It serves the purpose of separating the names of the report sections from the program. Identifying sections is important because they are used in the grammar to constrain certain constructions and to improve the analysis. Sections also provide relevant contextual information for the structured output, which is valuable for applications using the output. We observed a large textual variation in section names (i.e. the table contains over 123 section names). To facilitate NLP, it would be useful if software applications and healthcare personnel responsible for generating the text reports would represent the reports in a standard format and use standardized section names.

Manual analysis of the output that was initially performed showed that there was some loss of accuracy caused by extending the grammar. In order to avoid maintenance overhead, only one grammar is maintained that covers all the different domains. A significant effort would be incurred if different grammars had to be developed and maintained for each domain.

However, having only one broad-coverage grammar necessitates that the grammar be over-general. To address this problem we added a context-dependent mechanism to allow or disallow certain constructs depending on the clinical domain and/or section of report. For example, a temporal modifier, such as *1 o'clock* is not allowed in the **specimen** section of pathology reports, but is allowed elsewhere. This mechanism improves accuracy because *1'clock* (when it occurs in a specimen section of a pathology report) in *biopsy incision at 1 o'clock* generally refers to a region and not the time of day.

In this paper, we described the changes that were made to the system in order to cover the desired domains. The system appeared to have a high sensitivity and specificity based on numerous manual analyses. However, performance has not been evaluated since the new extensions, but is critical in order to determine their effect on the system. Evaluation is a very costly process, and we anticipate that evaluations well be performed for all the extensions in the future. An evaluation requires that a reliable, accurate, and unbiased reference standard be available so that the automated system can be compared against the reference standard. It would be ideal if a reference standard could be obtained automatically, but so far this has not been possible.

There are currently a total of 15,306 lexical entries in the lexicon. This is a fairly small number for coverage of clinical terms in the domains we are discussing. The lexicon was trained based on the frequencies of words and phrases in sample domain corpora. Words that were unknown to the system and had a frequency of over four occurrences were classified and added to the lexicon. Ones that rarely occurred were not reported, and therefore unlikely to have been entered. This means that the lexicon contains most of the typical findings in patient reports, but is not yet complete. We plan on continuing to augment the lexicon in order to add the more atypical findings.

Originally MedLEE was developed for decision support applications, and it was shown that the final representation affected performance because the applications depended on queries that were written to retrieve relevant reports based on the output that was generated. We found that it was preferable to simplify the structured output as much as possible in order to facilitate the writing of queries.

One way to simplify the output for decision support applications is to treat a compositional multi-word phrase as an atomic unit. For example, a phrase such as *left ventricular hypertrophy* is regarded as a unit instead of as a combination of words. When a user wants to retrieve reports associated with a specified condition, the user will compose a query based on a list of target output terms that MedLEE generates. A compositional term, like the example above, is easy to identify and select for the query because the entire term could be found on the list of target terms. In contrast, if a multi-word term is processed as a sequence of individual words, the output would consist of a finding **hypertrophy** with modifiers. The output for the phrase would then be less obvious because it would not be directly on the target list, although the individual components would be. That means the user would have to identify and select a primary finding, modifier types and values, and then compose a more complex query.

Another way to simplify the structured output is to reduce the variety of values for certain modifier types by mapping them to a limited set of terms appropriate for the particular type. For example, currently there are 228 different words and phrases associated with degree information but when they are processed by MedLEE, they are mapped into a set consisting of only 23 different values such as **low degree**, **moderate degree,** or **high degree**. With this approach, the output for *extensive sinus bradycardia* would be the same as the output for *severe sinus bradycardia* and *marked sinus bradycardia*. A reduction in variety considerably simplifies querying the structured output and appears to simplify coding applications, but the effectiveness of this approach for other types of applications is unknown. For example, for certain summarization applications there may be a loss of granularity.

For an application that involves the mapping of terms from one vocabulary system to another, a decompositional approach to multi-word phrases appears to be more effective than a compositional one. For example, in this mode, the phrase *left ventricular hypertrophy* would not be processed as a unit but as individual words. MedLEE functions in this mode by ignoring lexical entries for compositional multi-word phrases. This mode is useful in a mapping application because the vocabulary terms in each terminology can be structured using MedLEE, and then the terms can be matched based on structural similarities. Even if there is no exact structural match, it may still be possible to obtain a close match based on the structures. We are planning on developing and evaluating this approach further.

MedLEE can also be used to facilitate vocabulary development. A large corpus from the domain can be collected and processed, and the output organized in order to obtain information about terms used in the domain. A decompositional mode of processing is also needed for this type of application. The method we developed to facilitate vocabulary development is based on XML, and is discussed by Liu[9].

SUMMARY AND CONCLUSIONS

We have described successive extensions of a natural language processing system, called MedLEE, to multiple domains and applications. Originally the system was designed for decision support applications, but is now being used for vocabulary development and encoding into ICD-9, SNOMED, and the UMLS. Previously, the system was trained for chest radiological reports, mammography reports, and discharge summaries. With the current extensions, MedLEE also can process radiology, electrocardiography, echocardiograpy, and pathology. The system relies on a grammar and lexicon, as well as other knowledge-based components. The grammar, a complex rule-based component, required few changes for the new extensions. In contrast, the lexicon grew substantially. This is promising because lexical development is fairly straightforward whereas work on the grammar is more complex. It will be important to evaluate the system in the new domains. If performance is satisfactory, it will demonstrate that the incremental method used to extend the system is effective, and that further extension is likely to involve primarily lexical work. Most importantly it will demonstrate that it is possible to achieve broad coverage within one general system.

Acknowledgements

This publication was supported in part by grants LM06274 from the National Library of Medicine, IIS-98-17434 from the National Science Foundation, and by the Columbia CAT supported by the New York State Science and Technology Foundation.

References

1. Sager N, Lyman M, Nhan NT, TIck LJ. Medical language processing: applications to patient data representation and automatic encoding. Meth Inform Med 1995;34:140-46.
2. Baud RH, Rassinoux AM, Wagner JC, Lovis C. Representing clinical narratives using conceptual graphs. Meth Inform Med 1995;1/2:176-86.
3. Friedman C, Alderson PO, Austin J, Cimino JJ, Johnson SB. A general natural language text processor for clinical radiology. JAMIA 1994;1:161-74.
4. Haug PJ, Ranum DL, Frederick PR. Computerized extraction of coded findings from free-text radiologic report. Radiology 1990; 174:543-48.
5. Gundersen ML, Haug PJ, Pryor TA, van Bree R, Koehler S, Bauer K *et al.* Development and evaluation of a computerized admission diagnoses encoding system. Computers and Biomedical Research 1996;29:351-72.
6. Friedman, C., Knirsch, CA., Shagina, L., Hripcsak, G. Automating a severity score guideline for community-acquired pneumonia employing medical language processing of discharge summaries. In Lorenzi N, ed. Proc AMIA Symp. Phil. Hanley & Belfus. 1999;256-260.
7. Blanquet A and Zweigenbaum, P. A lexical method for assisted extraction and coding of ICD-10 diagnoses from free text patient discharge summaries. . In Lorenzi N, ed. Proc AMIA Symp. Phila. Hanley & Belfus.1999;1029.
8. Jain, NL. and Friedman, C. Identification of findings suspicious for breast cancer based on natural language processing of mammogram reports. In Masys, DR, ed. Proc AMIA Symp. Phila. Hanley & Belfus.1997;829-833.
9. Liu H and Friedman, C. A method for vocabulary development and visualization based on medical language processing and XML. Submitted to AMIA Fall Symp 2000.
10. Lussier Y, Shagina, L., and Friedman, C. Automated ICD-9 encoding using medical language processing: a feasibility study. Submitted to AMIA Fall Symp 2000.
11. Friedman C, Starren J, Johnson SB. Architectural requirements for a multipurpose natural language processor in the clinical environment, In: Gardner RM, ed. Proce of SCAMC 1995. Phil: Hanley & Belfus; 1995;347-51.
12. Friedman C, Hripcsak G, DuMouchel W, Johnson SB, Clayton PD. Natural language processing in an operational clinical information system. Nat.Lang.Eng. 1995;1:83-108.
13. Friedman, C. Towards a comprehensive medical language processing system: methods and issues. In Masys, DR. Proc AMIA Symp. Phil, Hanley & Belfus. 1997;595-599.