

## Facilitating Cancer Research using Natural Language Processing of Pathology Reports

Hua Xu<sup>a</sup>, Kristin Anderson<sup>b</sup>, Victor R. Grann<sup>b</sup>, Carol Friedman<sup>a</sup>

<sup>a</sup> Department of Biomedical Informatics, Columbia University, NY, USA

<sup>b</sup> Division of Epidemiology, Mailman School of Public Health, Columbia University, NY, USA

### Abstract

Many ongoing clinical research projects, such as projects involving studies associated with cancer, involve manual capture of information in surgical pathology reports so that the information can be used to determine the eligibility of recruited patients for the study and to provide other information, such as cancer prognosis. Natural language processing (NLP) systems offer an alternative to automated coding, but pathology reports have certain features that are difficult for NLP systems. This paper describes how a preprocessor was integrated with an existing NLP system (MedLEE) in order to reduce modification to the NLP system and to improve performance. The work was done in conjunction with an ongoing clinical research project that assesses disparities and risks of developing breast cancer for minority women. An evaluation of the system was performed using manually coded data from the research project's database as a gold standard. The evaluation outcome showed that the extended NLP system had a sensitivity of 90.6% and a precision of 91.6%. Results indicated that this system performed satisfactorily for capturing information for the cancer research project.

### Keywords:

Natural Language Processing, Pathology

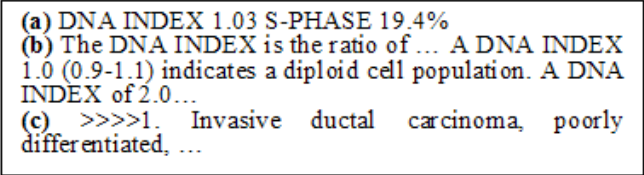
### Introduction

Many ongoing clinical research projects [1-2], such as projects involving studies associated with cancer, involve capturing information in pathology reports so that the information can be used to determine the eligibility of recruited patients for the study and to provide other information, such as cancer prognosis. The data is typically extracted and entered manually because pathology reports consist primarily of unstructured free-text and thus the information they contain are not usable for other automated processes that need to reliably access the information.

Natural Language Processing (NLP) offers an alternative to automated coding. Recently, NLP in the medical domain has begun to show promising results [3-8]. At Columbia-Presbyterian Medical Center (CPMC), MedLEE (Medical Language Extraction and Encoding System), a natural-language processor, has been successfully developed for automated encoding of the information content of text documents, including discharge summary, radiology, pathology and mammogram reports [9], but surgical pathology reports present a number of unique challeng-

es and are difficult to for MedLEE as well as for other NLP systems.

One difficulty is that many of the reports contain tabular data and are missing punctuation marks in many places. For example, in Figure 1(a), instead of a period, a space is used to separate the value of the finding *DNA INDEX* from the next finding *S-PHASE*, and there is no period following *S-PHASE*. The results of NLP systems generally depend on end of sentence markers to identify sentences as units of processing and lack of end of sentence marks results in errors. Another difficulty is that many of the reports have explanations that are interspersed with findings and the information in the explanations should not be considered as findings. Shown in Figure 1(b) is an explanation associated with the interpretation of *DNA INDEX* results. For NLP systems, it is difficult to differentiate between real findings and explanation statements. Another difficulty is that sometimes a pathology report contains special characters, such as '>' (shown as Figure 3 (c)), which will cause invalid XML output if the original text is preserved.



(a) DNA INDEX 1.03 S-PHASE 19.4%  
(b) The DNA INDEX is the ratio of ... A DNA INDEX 1.0 (0.9-1.1) indicates a diploid cell population. A DNA INDEX of 2.0...  
(c) >>>>1. Invasive ductal carcinoma, poorly differentiated, ...

Figure 1 - Sample findings of pathology reports.

A possible solution to the above difficulties would be to develop a preprocessor which can transform the original reports into a format that MedLEE or another NLP system can process more accurately. For the above *DNA INDEX* example (Figure 1(a)), the preprocessor can add periods in the appropriate places.

The particular cancer research project that we are collaborating with assesses disparities and risks of developing breast cancer for minority women. Disparities associated with race/ethnicity have been observed in breast cancer incidence, mortality, and survival in the United States [2]. Although women of African and Hispanic descent have lower breast cancer incidence rates than women of European (non-Hispanic) descent in the United States, the difference is diminishing, and women of African descent have higher breast cancer mortality rates than other women. In this project, data that includes tumor markers obtained from pathology reports are used. Data from pathology reports in-

clude 13 types of findings: *procedure name, tumor stage, number of positive lymph nodes, expression of estrogen receptors, progesterone receptors and Her-2/Neu, nuclear grade, ploidy, DNA INDEX, quantitative S-Phase, qualitative S-Phase, G2-M and Proliferation Index*. In this study, we used all the findings except the *tumor stage* information.

Currently, the clinical research project has a data manager who is responsible for reading pathology reports manually, extracting relevant information, using rules to interpret them and entering the final results into an Access database. The rules, which were used to interpret the finding results and determine the final entries in the database, were developed by the research project team to specify details of the manual encoding process. Figure 2 shows the rules for the *Her-2/Neu* result. The result for *Her-2/Neu* can be “Positive”, “Negative” or “N/A”. For a result of “Negative”, there are three possible findings: (1) The Herceptest Score is “0” or “1+”; (2) The Her-2/Neu protein is < .1pg/cell by immunohistochemical quantification; (3) a narrative expression, such as “No over expression is identified”. If one of three findings is found, the result for Her-2/Neu will be “Negative” in the database. Similar rules for the “Positive” result of Her-2/Neu are also applied. For an automated NLP system that extracts and structures relevant clinical information in the reports, a postprocessor would be needed to implement those rules in order to obtain the same results as the manual coding.

Result	Conditions
Negative	No Over expression is identified
	Or Herceptest score of 0 or 1+
	Or Her-2/Neu protein < .1pg/cell
Positive	Over expression is identified
	Or Herceptest score of 2+ or 3+
	Or Her-2/Neu protein >= .1pg/cell
N/A	Not available

Figure 2 - Rules for Her-2/Neu Finding

This paper reports on the methods used to extend MedLEE for this application. Additionally, a study was performed to determine whether the extended MedLEE system could be used to facilitate the information needs of above project. The methods section will describe the development of the extended MedLEE system and the feasibility study. The results section will discuss the outcome of the evaluation.

## Methods

### Manual Analysis

Several samples of surgical pathology reports associated with breast cancer were retrieved from the data repository at Columbia-Presbyterian Medical Center (CPMC), and a manual analysis was performed. The analysis included several different steps. The first step identified the overall structure of the surgical pathology report and the information types in each section. The second step identified findings needed for the particular research project. The third step analyzed the sentences containing the findings to determine if MedLEE’s representational schema was adequate for capturing the types of information that were needed.

Based on the manual analysis of findings that were relevant, the findings were classified into two categories: tabular findings and narrative findings. Tabular findings, such as “DNA INDEX 1.09”, primarily have simple “Name Value” formats so that they are easy to capture automatically with high accuracy. Tabular findings include *DNA INDEX, G2-M, quantitative S-PHASE, Proliferation Index, Estrogen Receptors and Progesterone Receptor*. Narrative findings, such as “No Her-2/neu over expression is identified”, are in the form of narrative text. Narrative types of information are much more varied, and are usually more challenging to capture. We would expect that an NLP system would have somewhat lower performance in interpreting them than the tabular types of findings, and therefore we evaluated the results for the two categories separately. Narrative findings include *Procedure Name, Number of Positive Lymph Nodes, Her-2/Neu, Nuclear Grade, Ploidy and Qualitative S-Phase*. Therefore every report contains 6 tabular findings and 6 narrative findings for this study.

As a result of manual analysis, we determined that MedLEE’s target schema was adequate for representing the relevant information. The only modification that was needed was to add new lexical entries in order to recognize new types of information.

### System Development

An extended MedLEE system was developed. Figure 3 shows an overview of the system. It consists of a preprocessor for pathology reports, the MedLEE system with an expanded lexicon and a postprocessor for the research project. Fifteen new lexical entries for terms, such as “DNA INDEX”, “G2-M”, and “Her-2/Neu” were created and added to the existing MedLEE system. A pre-processor is used to transform the pathology report into a more suitable format for MedLEE. MedLEE processes the transformed report and generates structured XML output, and then the post-processor processes the XML output, extracts findings that are needed for the clinical research project, interprets them according to the rules formulated for the research project, and then maps the information into a data file that contains a tabular format where the information is in a form that can be directly used for the project.

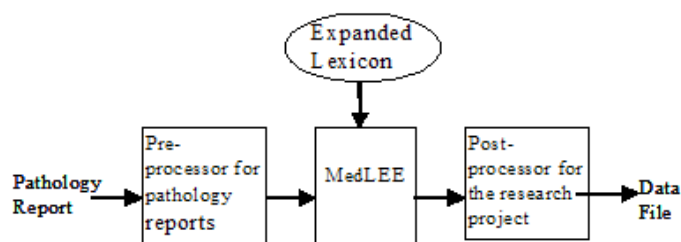


Figure 3 - Overview of the Extended MedLEE System.

The pre-processor was developed in Perl to transform certain text in a pathology report into a format that MedLEE could process more accurately. The preprocessor adds periods at the appropriate place (shown as Figure 4(a)), identifies and marks explanation statements with a special tag (“<ign>”) so that the information enclosed by the ignore tag will be ignored when processing the text (shown as Figure 4(b)), and removes special

characters, such as ‘>’ (shown as Figure 4(c)). Figure 4 shows the output after preprocessing the statements in Figure 1.

```
(a) DNA INDEX 1.03. S-PHASE 19.4%.
(b) <ign>The DNA INDEX is the ratio of ... A DNA
INDEX 1.0 (0.9-1.1) indicates a diploid cell population.
A DNA INDEX of 2.0...</ign>
(c) 1. Invasive ductal carcinoma, poorly differentiated,
...
```

Figure 4 - Sample outcomes after preprocessor.

The post-processor was developed in Perl by using the XPATH module from CPAN (Comprehensive Perl Archive Network). The postprocessor extracts relevant findings from the XML output of MedLEE, interprets them to obtain the final results based on the rules established for manual coding, enters the final results into a data file which is a tabular form that can be imported directly into the Access database of the research project. In the data file, 12 finding results from one report are placed in the same row in tabular form. This format was determined by the Access database schema associated with the research project. Figure 5 shows an example of a the value of the finding *Her-2/Neu*. Column 1 shows the original statement containing *Her-2/Neu* in the report, column 2 shows the corresponding XML output generated by MedLEE, and column 3 shows the final results that are stored in the corresponding field in the data file. In this example, the results for the final value of *Her-2/Neu* were based on the rules shown in Figure 2.

Original text	XML output	Result
No Her-2/Neu overexpression is identified.	<labtest v="Her-2/neu" idref="p1138"> <certainty v="no" idref="p1136"> </certainty> <sectname v="report addendum item"> </sectname> <sid idref="s6.19.3"> </sid> </labtest>	Negative

Figure 5 - Sample of information associated with *Her-2/Neu* before and after postprocessing.

### Evaluation

Manually coded data from 70 surgical pathology reports were obtained from the Access database of the breast cancer research project and saved into a data file. This data file of manual coding served as a gold standard in the study. This data was previously entered into the database as part of the operation of the research study, and was not prepared for this particular NLP study. The 70 original textual surgical pathology reports were also obtained from the data repository of CPMC. Each report was de-identified and the same study ID corresponding to the manually coded data ID was assigned. Twenty reports were used as a training set to develop the preprocessor and postprocessor, and to extend MedLEE. The remaining 50 reports were used as the test set to evaluate the performance of the extended MedLEE system.

After the extended MedLEE system was developed, it was evaluated by processing the 50 reports in the test set to obtain results.

Comparison was performed between the gold standard and the extended MedLEE system. The total number of matches and mismatches, which contained incorrect or incomplete finding results, were counted and used for analysis. Differences between the extended system and the gold standard were shown to the data manager of the research project for checking purpose. The data manager found a number of errors in the gold standard and then corrected the gold standard by fixing the errors. Comparison between the extended MedLEE system and the revised gold standard was performed as well.

### Results

Initially, the extended MedLEE system had an overall sensitivity of 90.6% and an overall precision of 91.6% when using the original gold standard (shown as Table 1). As we anticipated, performance was better for the tabular findings (sensitivity: 95.8% and precision: 95.4%) than for narrative findings (sensitivity: 86.0% and precision: 88.3%). The data manager of the research project checked the differences that were detected between the extended MedLEE system and the manual coding, and determined that a number of errors (39.2%(20/51)) were due to the manual coding. When the data manager corrected the gold standard, the extended MedLEE system had an overall sensitivity of 93.9% and an overall precision of 95.4% (shown as Table 2).

Table 1: Comparison of findings identified by MedLEE system and original manual Coding

Category	Sensitivity	Precision
Tabular Findings	95.8%(227/237)	95.4%(227/238)
Narrative Findings	86.0%(233/271)	88.3%(233/264)
Overall	90.6%(460/508)	91.6%(460/502)

Table 2: Comparison of findings identified by MedLEE system and revised manual Coding

Category	Sensitivity	Precision
Tabular Findings	96.2%(230/239)	96.6%(230/238)
Narrative Findings	91.9%(249/271)	94.3%(249/264)
Overall	93.9%(479/510)	95.4%(479/502)

For tabular findings, the system had a sensitivity of 96.2% and a precision of 96.6% with the revised gold standard. For narrative findings, the system had a sensitivity of 91.9% and a precision of 94.3% with the revised gold standard.

Comparison of processing time was also performed for the manual coding and the extended MedLEE system. As estimated by the data manager of the research project, it took approximately 10 minutes to manually code one report. Therefore 50 reports

would take 500 minutes for manual coding. For the extended MedLEE system, it took approximately 10 minutes to process 50 reports when using a Sun Blade 2000 workstation with two 900 MHz 64-bit UltraSPARC III CPUs and 2GB RAM.

## Discussion

In this study, we skipped one finding *tumor stage* that is also needed for the clinical research project because there are no direct findings related to *tumor stage* in the report. The *tumor stage* finding was determined by other two findings: *tumor size*, which is not captured in this study, and *number of positive lymph nodes*, which is one of the findings used in this study. The rules to determine the *tumor stage* are complex and have changed during the period of the clinical research project. In the future, the extended MedLEE system will be expanded for the operational phase in order to capture the finding of *tumor size* and implement the rules to determine the *tumor stage*.

During the development of the extended MedLEE system, another difficulty of surgical pathology reports was explored. Although it occurs infrequently, one report sometimes contains more than one specimen of interest for the research project, and findings associated with each different specimen are mentioned throughout the report. Figure 6 illustrates an example of this problem. The “Specimen” section refers to two specimens: A and B. Later, the “Microscopic Description” section mentions findings associated with specimen A and specimen B separately. It is difficult for an NLP system to match the findings with the particular specimens, especially since the specimens are not uniformly identified in the reports. Another preprocessing procedure was developed to solve this problem by creating separate report segments where findings are associated with the appropriate specimen. This procedure will be integrated into the preprocessor to improve the handling of these complex cases.

**Specimen:**  
A: Breast, left, modified radical mastectomy B: Breast, right, needle core biopsy

**Microscopic Description:**  
Slide "A" from the left mastectomy....  
Slide "B" shows ....

Figure 6: sample of multiple-specimens problem

Some errors of the extended MedLEE system appear easy to fix. For example, the system did not get results for the *Her-2/Neu* finding in a few reports because the reports contained the phrase “Herceptest staining score” instead of “Herceptest score”. To fix this, the term “Herceptest staining score” has to be added to the lexicon. Other errors would be more difficult to fix. For example, the current system did not capture the *number of positive lymph nodes* if the finding is expressed in two sentences as: “There were 16 lymph nodes. One of them was positive.”

By analyzing the errors in the manual coding, we obtained interesting findings: errors in manual coding seemed to increase when the interpretation rules were complicated or confusing. For example, in manual coding, the *Her-2/Neu* results were occasionally falsely interpreted as “Positive” when the “Herceptest

score” was “1+”. It is possible that “1+” may suggest a result “Positive”, though the rules defined it as “Negative”. This demonstrates that an automated system could eliminate some errors caused by human factors and produce more consistent results if the performance is satisfactory.

Although the extended MedLEE system generated results in less time than the manual coding process, the processing time was relatively long for an automated system because the XPATH module that was used was slow when the XML output of a report was large. To improve the efficiency of the system, there are two alternatives we will explore. One will involve reducing the size of XML input that is sent to the postprocessor: only relevant findings would be sent instead of output for the entire report. The other alternative would be to find a more efficient XML parser to replace XPATH.

Note that a relatively simple Perl script would have been adequate to extract the structured information needed for this project without use of MedLEE. However, the rest of data that was required (around 50%) occurred as free-text, and it would be too difficult for a simple Perl script to handle the variety found in free-text. Furthermore, additional effort would also be required to include code in the Perl program to generate output comparable with MedLEE’s XML output. By using a simple preprocessor to change the structured data to a form appropriate for MedLEE, the effort that was required for the overall system was minimized.

Although this particular application only needs limited clinical information, MedLEE, which was previously trained in the pathology domain, captures a much larger amount of clinical information in pathology reports. It means that this system can also be used for other applications requiring other types of information in pathology reports. More than that, the model described here can serve as a prototype for other research studies. For example, other types of reports, which may contain tables, explanations that are not findings, or present similar difficulties for existing NLP systems, can use straightforward preprocessing programs to reduce or eliminate the difficulties. Meanwhile, this model broadens the application fields for NLP.

This study had a few limitations. The gold standard we used was obtained from the manual coding of one data manager and could be a source of bias. In fact, the data manager noted that the gold standard contained errors, and different results may have been obtained if the gold standard was determined by a different method. However, it is very difficult to obtain a gold standard. By using the manual coding that already existed as a gold standard, we saved a substantial amount of time and effort, and also used data that was realistically obtained for an ongoing clinical study.

Our results demonstrated that the expanded system performed well and that the clinical research project would benefit from it. In the future, we hope to integrate the extended MedLEE system into an operational mode so that it will be used to populate the database of the research project. Integration issues will also involve linking the final results to the relevant places in the original reports so that, if desired, the data manager of the project would be able to manually validate the results obtained using the system. Since NLP will most likely generate some errors, it will

be important for clinical researchers to be able to conveniently validate the results for quality control.

## Conclusion

We have identified how certain information in pathology reports, which is troublesome for NLP systems, can be captured accurately through use of a preprocessor that adjusts the text and eliminates some of the difficulties. Additionally, we also demonstrated that post-processing was generally needed. A postprocessor was used to incorporate rules formulated by the clinical project in order to obtain the information in a form that was compatible with that of the project. The extended NLP system was evaluated using a gold standard established as part of the operation of the clinical research project. Results showed that the extended MedLEE system had a sensitivity of 90.6% and a precision of 91.6% with the original gold standard. After the data manager revised the gold standard to correct errors she felt she made, the system had a sensitivity of 93.9% and a precision of 95.4%. This study demonstrated how an existing NLP system could be used to facilitate clinical research projects with only minor modification.

## Acknowledgments

This work was supported by grants LM06274 and LM7659 from the National Library of Medicine. We would like to thank Johanna Alves and Judith R Jacobson from the Department of Epidemiology in Mailman School of Public Health of Mailman School for their help of obtaining pathology reports, Lyudmila Shagina of the Department of Biomedical Informatics for her technical assistance with running MedLEE.

## References

- [1] Rosenberg DJ, Neugut AI, Ahsan H, Shea S. Diabetes mellitus and the risk of prostate cancer. *Cancer Invest.* 2002; 20(2): 157-65.
- [2] Grann, V.R. and Jacobson, J.S. Health Insurance and Cancer Survival, *Arch Intern Med*; (In press Oct 2003)
- [3] Baud RH, Rassinoux AM, Scherrer JR. Natural language processing and semantical representation of medical texts. *Methods Inf Med* 1992; 31:117-125.
- [4] Haug PJ, Ranum DL, Frederick PR. Computerized extraction of coded findings from free-text radiologic reports: work in progress. *Radiology* 1990; 174:543-548.
- [5] Hripcsak G, Friedman C, Alderson PO, DuMouchel W, Johnson SB, Clayton PD. Unlocking clinical data from narrative reports: a study of natural language processing. *Ann Intern Med* 1995; 122:681-688.
- [6] Sager N, Lyman M, Bucknall C, Nhan N, Tick LJ. Natural language processing and the representation of clinical data. *J Am Med Inform Assoc* 1994; 1:142-160.
- [7] Zweigenbaum P, Bouaud J, Bachimont B, Charlet J, Boisivieux JF. Evaluating a normalized conceptual representation produced from natural language patient discharge summaries. *Proc AMIA Annu Fall Symp* 1997; 590-594.
- [8] Fisman M, Chapman WW, Aronsky D, Evans RS, Haug PJ. Automatic detection of acute bacterial pneumonia from chest x-ray reports. *J Am Med Inform Assoc* 2000; 7:593-604.
- [9] Friedman C, Knirsch C, Shagina L, Hripcsak G. Automating a severity score guideline for community-acquired pneumonia employing medical language processing of discharge summaries. *Proc AMIA Symp.* 1999;:256-60.

## Address for correspondence

Carol Friedman, PhD  
 Department of Biomedical Informatics  
 Columbia University  
 VC-5, 622 W. 168 Street, New York, NY 10032  
 E-mail: <friedman@dbmi.columbia.edu>.

