Connecting Medical Informatics and Bio-Informatics
R. Engelbrecht et al. (Eds.)
ENMI, 2005

Evaluation of Medical Problem Extraction from Electronic Clinical Documents Using MetaMap Transfer (MMTx)

Stéphane Meystre, Peter J Haug

Department of Medical Informatics, University of Utah School of Medicine, Salt Lake City, Utah, U.S.

Abstract

To improve the use and quality of the electronic Problem List, which is at the heart of the problem-oriented medical record in development in our institution (Intermountain Health Care, Utah, U.S.), we developed an Automated Problem List system using Natural Language Processing (NLP) technologies. A key part of this system is a module that automatically extracts potential medical problems from free-text clinical documents. The NLP module uses MMTx, developed at the U.S. National Library of Medicine. Negation detection was added to this application by adapting a negation detection algorithm called NegEx. To evaluate the adequacy of the performance of the NLP module for our Automated Problem List system, we evaluated it with 160 electronic clinical documents of different types. Two different data sets for MMTx were used: the default full UMLS data set and a customised subset adapted to detect the set of 80 medical problems we are interested in. With the default data set, we measured a recall of 0.74 (95% CI 0.68-0.8) and a precision of 0.76 (0.69-0.82). The customised subset had a significantly better recall of 0.9 (0.85-0.94), and a non-significantly different precision of 0.69 (0.63-0.75).

Keywords:
Program Evaluation; Natural Language Processing; Medical Records, Problem-Oriented.

1. Introduction

The Medical Problem List is an important piece of the medical record as well as a central component of the problem-oriented medical record in development in our institution. To serve the functions it is designed for, the Problem List has to be as accurate and timely as possible. The Problem List application that is used currently is typically incomplete or inaccurate, and is often unused. To address this deficiency, we extended this tool with components designed to make the Problem List easy and efficient to maintain. We developed an application using Natural Language Processing (NLP) to harvest potential Problem List entries from the multiple free-text electronic documents available in our EMR (Electronic Medical Record). These proposed problems drive an application designed for management of the Problem List, and are proposed to the physicians for addition to the official Problem List. Physicians then accept the problems proposed by changing their state to *active*, *inactive*, or *resolved*, or reject them by changing their state to *error*. The global aim of our project is to automate the process of creating and maintaining a Problem List for hospitalised patients and thereby to help guarantee the timeliness, accuracy and completeness of this information.

The problem-oriented, Computer-based Patient Record (CPR) and the Problem List have seen renewed interest as an organisational tool in the recent years [1,2]. Advantages to the

Problem List are that it can be the central place for clinicians to obtain a concise view of all patients problems, that it facilitates associating clinical information in the record to a specific problem, and that it can encourage an orderly process of clinical problem solving and clinical judgement. The Problem List in a problem–oriented patient record also provides a context in which continuity of care is supported, preventing both redundant and repeated actions [1].

Since problems of interest are frequently referenced in clinical documents collected electronically, we chose to supplement the practice of manually entering problems by developing an application built with Natural Language Processing tools. Several systems designed to automatically map clinical text concepts to standardised vocabularies have been reported, like MetaMap [3] and IndexFinder [4]. MetaMap was developed by the U.S. National Library of Medicine (NLM), and is used to index text or to map information in the analysed text to UMLS concepts. The mapped concepts are ranked, but no negation detection is performed. Five steps are needed, beginning with noun phrases identification using the SPECIALIST minimal commitment parser[1], followed by variants generation, candidate phrases retrieval, and computing of a score for each candidate by comparing it with the input phrase, and ending with the mapping and ranking using the computed score. MetaMap has been shown to identify most concepts present in MEDLINE titles [5]. It has been used for Information Extraction in biomedical text [5,6] and has been shown capable of extracting the most critical findings in 91% of the documents in a prior study [7]. Independent negation detection is required when using MetaMap. The application does not discriminate between present and absent concepts. In the medical domain this is important due to the fact that findings and diseases are often described as absent. A few negation detection algorithms have been developed, like NegEx, a computationally simple algorithm using regular expressions [8], or the more complex general-purpose Negfinder [9]. These algorithms have been evaluated and have shown good results. NegEx has been shown to have a sensitivity of 94.5% and a specificity of 77.8% [8].

2. Materials and Methods

As mentioned earlier, the Automated Problem List (APL) system extracts potential medical problems from free-text medical documents, and uses NLP to achieve this task. The APL system is made of two main components: a background application and the Problem List management application. The background application does all the text processing and analysis and stores extracted problems in the central clinical database, called CDR (Clinical Data Repository). These problems can then be accessed by the Problem List management application integrated in our Clinical Information System. We are currently evaluating a prototype in which the background application looks for 80 different problems, principally diagnoses, that were selected based on their frequency of use in our field of evaluation (cardiovascular and general medicine). Document processing starts with detection of sections and sentences, and is followed by a text restructuring step, to prepare it for the NLP module. This module uses MMTx, the Java™ version of MetaMap. Some processing to reduce ambiguity is also required, to avoid confusion with common acronyms (e.g. "Mr." detected as mitral regurgitation). Since we are currently interested in 80 different problems, and not in the entire UMLS Metathesaurus content (UMLS version 2004AA contains over 1 million concepts[2]), we created a subset of the Metathesaurus adapted to our system. The selection process resulted in a reduction to about 0.25% of the original data set (from more than a million to about 2,500 concepts). This reduction made the NLP module

[1] http://ii.nlm.nih.gov/MTI/phrasex.shtml
[2] http://www.nlm.nih.gov/pubs/factsheets/umlsmeta.html

more than 3 times faster, and also improved accuracy. The process of selecting relevant concepts first consisted in the use of MetamorphoSys[3], an application provided with the UMLS that allows filtering the Metathesaurus based on source vocabularies, semantic types and other filters. To subset it further, we loaded the data into a MySQL[4] database for subsequent processing. A mapping table was manually built to link the 80 selected concepts with all related subconcepts (e.g. *Right Bundle Branch Block* was mapped to *Incomplete Right Bundle Branch Block*, *Complete Right Bundle Branch Block*, and *Other or unspecified Right Bundle Branch Block*). This table was built manually and was used to select relevant concepts in the UMLS subset. The final step was the creation of the MMTx data files. A tool called MMTx Data File Builder[5] is provided with MMTx and allows the creation of these files from UMLS Metathesaurus subsets or from custom built data sets.

As mentioned above, MMTx lacks negation detection: concepts are mapped the same way whether they are described as present or as absent. We therefore had to add this feature. To this end, the NLP module actually works in two steps: a first step uses MMTx to extract each potential medical problem, and a second step infers the state of each of those problems. During the first step, phrases recognised by MMTx are replaced by a tag containing the corresponding UMLS identifier (CUI). This transformed version of the sentence is then passed to a negation detection algorithm that infers the state of the mapped concept. In this second phase, we adapted the NegEx algorithm described above, and implemented it in Java. We used the improved version of NegEx, called NegEx 2[6].

Study design:

To determine the success of our approach, we conducted a descriptive laboratory resource's function study. Two standard measures were used to evaluate the accuracy of this Natural Language Processing system: precision, the proportion of problems found that were correct (equivalent to positive predictive value here), and recall, the proportion of problems present in the documents that were actually found (equivalent to sensitivity here). Another typical value combining precision (P) and recall (R) – the F-measure (equal to $(\beta^2+1)PR / (\beta^2P)+R$) – was also calculated.

A medical problem was considered present if mentioned in the text as probable or certain in the present or the past (e.g. "the patient has asthma"; "past history positive for asthma"; "pulmonary oedema is probable"), and considered absent if negated in the text or not mentioned at all (e.g. "this test excluded diabetes…"; "he denies any asthma").

We randomly selected 160 clinical documents from a study population of adult inpatients seen in a cardiovascular unit of the LDS Hospital during the year 2002. Subjects were restricted to those who had stayed for at least 48 hours. These clinical documents were of various types, like radiology reports, procedure reports, history and physical exam reports, pathology reports, discharge summaries, progress notes, etc. They were processed by our background application, using the MMTxAPI and MMTx version 2.3.C. The system processed each document twice: once with the default data set, and once with our customised subset. The extracted problems and a transformed XML version of the document were then stored in a local database. This transformed version of the document uses an information model described in another publication [10]. It is used by the web-based review application mentioned below. A reference standard for problems present in these documents was created using an electronic chart review by physicians. Two independent physicians reviewed each electronic document using a web-based review

[3] http://www.nlm.nih.gov/research/umls/meta6.html
[4] http://www.mysql.com
[5] http://mmtx.nlm.nih.gov
[6] http://web.cbmi.pitt.edu/chapman/NegEx.html

application. When the two reviewers disagreed, a third physician determined the presence or absence of the disputed problem. To reduce the potential disagreement between reviewers, they were trained and tested on selected sample cases before the formal review, and were provided a set of standardised instructions. We also used a medical record review technique called *structured implicit review* that focuses the reviewers' attention on specific issues (our list of selected problems) on which judgement is to be based [11]. This technique is associated with higher inter-rater reliability than *implicit review*, where reviewers use only their knowledge or beliefs to make judgements. This focus was achieved in the review application by displaying the document to review beside a list of the 80 problems to look for. Reviewers checked the problems present in the document and submitted these as an initial review. To improve the quality of the review, these first results were compared in real-time with the results of the NLP module. Reviewers were then asked whether they wanted to keep a problem that wasn't found by NLP, or add a problem that was found only by NLP. During this second phase, the document was displayed with the sentences containing the problem text highlighted in red for faster reading. After responding to the suggestions from the application, the reviewers submitted their final list of the problems present. The two reviews of each document were compared and, if disagreement was found, a third reviewer used the same web-based application to select the disputed problems he considered present in the document.

3. Results

Eight physicians participated in the review process to create the reference standard. Five were board-certified physicians, and three were residents with two or more years of training. With the web-based application described above, reviewers spent an average of 93 to 189 seconds per document. Reviewers' overall agreement was almost perfect, with a Cohen's kappa of 0.9 and a Finn's R of 0.985. This latter value is more representative of the agreement, our agreement table being strongly skewed, with far more true negatives than true positives.

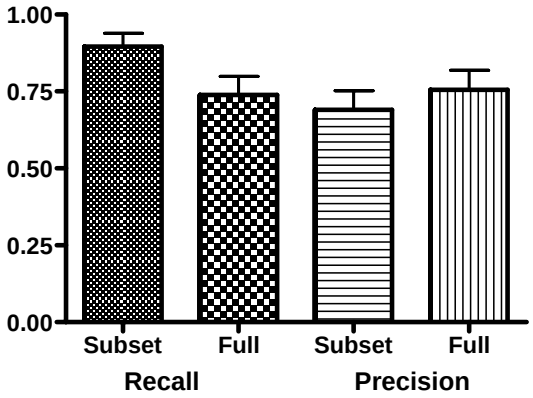

Figure 1: Graphical display of recall and precision
(Subset is customised subset; Full is full data set)

Recall and precision were measured with means and 0.95 confidence intervals. The F-measure was calculated with a value of 1 (same weight for recall and precision), and a value of 2, to give more importance to the recall, the most important feature for our system. Indeed, our aim is to detect as many medical problems present as possible.

Statistical evaluation of these results showed that the recall was significantly higher (two-

tailed $p < 0.0001$) for the customised subset than the full data set. Precision was not significantly different (two-tailed $p = 0.0797$) (Figure 1 and Table 1). When only considering certain types of clinical documents analysed with the customised data subset, no significant difference was found, except a significantly higher precision with discharge summaries when compared to history and physical exam reports. The test used for this analysis was Mann-Whitney U-statistic, to compensate for lack of normality.

Table 1: NLP module evaluation results

Measurements	Full default data set	Customised data subset	Radiology reports	H&P reports	Discharge summaries
Recall	0.740 * (0.680-0.8)	0.896 * (0.854-0.949)	0.802 (0.667-0.937)	0.918 (0.854-0.982)	0.878 (0.786-0.971)
Precision	0.756 (0.694-0.819)	0.691 (0.63-0.752)	0.624 (0.477-0.771)	0.648 □ (0.524-0.772)	0.821 □ (0.714-0.928)
F-Measure ($\beta = 1$)	0.748	0.78	0.701	0.76	0.849
F-Measure ($\beta = 2$)	0.743	0.846	0.759	0.847	0.866

* Extremely significant difference ($p < 0.0001$); □ Significant difference ($p = 0.0368$)

4. Discussion

To reduce biases and improve the generalisability of this evaluation, we tried respecting criteria for effective evaluation of NLP systems [12]. Most criteria were respected. However, the developer of the system also participated in the evaluation. To minimise this problem, documents were randomly selected after the system was frozen for evaluation, and reviewers did their task fully independently.

The results of this evaluation show good recall and satisfying precision, both at a level that fulfils our requirements for the NLP module of our Automated Problem List system in a clinical setting. A sufficient recall is required to significantly improve the quality of the Problem List, and a sufficient precision is desirable to avoid overloading the Problem List with false positives. Our application was developed to maximise recall, to the expense of a lower precision. We suppose that about 30% of false positives will be acceptable by users of the problem list. Our results compare favourably with another evaluation of MMTx, where a recall of 53% was reported. However, this latter result has to be considered cautiously because of small sample size and other reasons [13]. Also, our system only extracted a limited set of concepts, and all children of those concepts were matched to the parent ones, therefore improving the recall. Other NLP systems extracting UMLS concepts from free-text have been reported, like MetaMap with exact-match recall of 52.8% and exact-match precision of 27.7% [5]; this study evaluated detection of all biomedical concepts in title phrases. MedLEE has been recently evaluated when extracting UMLS concepts from medical text documents, achieving 83% recall and 89% precision [14].

The excellent agreement among reviewers allows a reference standard of good quality, therefore giving accurate results for the set of randomly selected test documents. A limitation is the fact that the developer of the system also designed and led its evaluation, therefore reducing the generalisability of this study. The limited set of 80 targeted medical problems also limits the generalisability of the study. The sample size was sufficient to show significantly different recall between the two data sets used with MMTx, but a larger sample could have reduced confidence intervals and possibly allowed detection of differences in precision.

5. Conclusion

We developed tools to automate the Problem List using NLP to extract potential medical problems from free-text documents in a patient's EMR. This system's goal is to improve the Problem List's quality by increasing its completeness, accuracy and timeliness. We have evaluated the NLP module developed with MMTx and shown reasonable performance for clinical use in our system. The effect of our system on the quality of actual Problem Lists will be evaluated. We anticipate that, by automatically proposing appropriate additions to the list of problems, we will see an increased proportion of correct problems, a reduced proportion of incorrect problems, and a reduced time between problem identification and addition to the Problem List. This will help to guarantee the quality of this central component for our problem-oriented Electronic Medical Record. Further analysis of the NLP module is planned including a comparison to other Natural Language Processing tools present in our laboratory.

6. Acknowledgements

This work is supported by a Deseret Foundation Grant (Salt Lake City, Utah, U.S.).

7. References

[1] Bayegan E, Tu S. The helpful patient record system: problem oriented and knowledge based. *Proc AMIA Symp* 2002:36-40.
[2] Elkin PL, Mohr DN, Tuttle MS, Cole WG, Atkin GE, Keck K, et al. Standardized problem list generation, utilizing the Mayo canonical vocabulary embedded within the Unified Medical Language System. *Proc AMIA Annu Fall Symp* 1997:500-4.
[3] Aronson AR. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. *Proc AMIA Symp* 2001:17-21.
[4] Zou Q, Chu WW, Morioka C, Leazer GH, Kangarloo H. IndexFinder: A Method of Extracting Key Concepts from Clinical Texts for Indexing. *Proc AMIA Symp* 2003:763-7.
[5] Pratt W, Yetisgen-Yildiz M. A Study of Biomedical Concept Identification: MetaMap vs. People. *Proc AMIA Symp* 2003:529-33.
[6] Weeber M, Klein H, Aronson AR, Mork JG, de Jong-van den Berg LT, Vos R. Text-based discovery in biomedicine: the architecture of the DAD-system. *Proc AMIA Symp* 2000:903-7.
[7] Shadow G, McDonald C. Extracting structured information from free text pathology reports. In: *Proc AMIA Symp* 2003. p. 584-588.
[8] Chapman WW, Bridewell W, Hanbury P, Cooper GF, Buchanan BG. A simple algorithm for identifying negated findings and diseases in discharge summaries. *J Biomed Inform* 2001;34(5):301-10.
[9] Mutalik PG, Deshpande A, Nadkarni PM. Use of general-purpose negation detection to augment concept indexing of medical documents: a quantitative study using the UMLS. *J Am Med Inform Assoc* 2001;8(6):598-609.
[10] Meystre S, Haug PJ. Medical problem and document model for natural language understanding. *Proc AMIA Symp* 2003:455-9.
[11] Ashton CM, Kuykendall DH, Johnson ML, Wray NP. An empirical assessment of the validity of explicit and implicit process-of-care criteria for quality assessment. Med Care 1999;37(8):798-808.
[12] Friedman C, Hripcsak G. Evaluating natural language processors in the clinical domain. *Methods Inf Med* 1998;37(4-5):334-44.
[13] Divita G, Tse T, Roth L. Failure Analysis of MetaMap Transfer (MMTx). *Medinfo*2004;2004:763-7.
[14] Friedman C, Shagina L, Lussier Y, Hripcsak G. Automated Encoding of Clinical Documents Based on Natural Language Processing. *J Am Med Inform Assoc* 2004.

Address for correspondence:

Stéphane Meystre
Department of Medical Informatics
University of Utah School of Medicine, Room AB194
Salt Lake City, UT 84132-2913, USA
s.meystre@utah.edu or smeystre@bluewin.ch