Natural language processing to extract follow-up provider information from hospital discharge summaries

Martin C. Were, MD MS[1,2]; Sergey Gorbachev, MD[1,2]; Jason Cadwallader, MD[2]; Joe Kesterson, MA[1]; Xiaochun Li, PhD[1,2]; J. Marc Overhage, MD PhD[1,2]; Jeff Friedlin, DO[1,2]
[1]Regenstrief Institute Inc, Indianapolis, IN and [2]Indiana University School of Medicine, Indianapolis, IN

Abstract:

Objective: *We evaluate the performance of a Natural Language Processing (NLP) application designed to extract follow-up provider information from free-text discharge summaries at two hospitals.* ***Evaluation:*** *We compare performance by the NLP application, called the Regenstrief EXtracion tool (REX), to performance by three physician reviewers at extracting follow-up provider names, phone/fax numbers and location information. Precision, recall, and F-measures are reported, with 95% CI for pair-wise comparisons.* ***Results:*** *Of 556 summaries with follow-up information, REX performed as follows in precision, recall, F-measure respectively: Provider Name 0.96, 0.92, 0.94; Phone/Fax 0.99, 0.92, 0.96; Location 0.83, 0.82, 0.82. REX was as good as all physician-reviewers in identifying follow-up provider names and phone/fax numbers, and slightly inferior to two physicians at identifying location information. REX took about four seconds (vs. 3-5 minutes for physician-reviewers) to extract follow-up information.* ***Conclusion:*** *A NLP program had physician-like performance at extracting provider follow-up information from discharge summaries.*

Introduction:

The transition of care from the inpatient to outpatient setting is one of major patient safety concern.(1) Studies show that an alarming number of medical errors occur during this transition, and that these errors are largely a result of 'fumbled handoffs' between providers and institutions.(2-4) In fact, these errors are more likely to occur when there is need to contact the outpatient follow-up providers long after the patient is discharged from the hospital – as is often the case with test results returning after hospital discharge that require a change in patient management.(5)

The process for identifying the follow-up providers needs to be straightforward and efficient. Unfortunately, this is often not the case. In many cases, information about follow-up providers exists only within the patient's hospital discharge summary - the information is typically spread out in several areas within the summary, and is usually documented in an unstructured free-text format. To get to this follow-up information, busy providers have to manually peruse through most of the discharge summary. This is time-consuming, and the process has to be repeated every time this information is needed for a different patient. The coded follow-up information is also needed by several systems, including those that deliver discharge summaries to the follow-up providers and those used for healthcare quality measures.

Approaches are needed to efficiently extract follow-up provider information from free-text discharge summaries, and to accurately code this information into a database for easy referral and queries. Natural language processing (NLP) has been proven effective in extracting relevant clinical data from text reports.(6) NLP systems have been used to extract clinical data from discharge summaries, including diseases,(7) adverse event data (8), as well as smoking status and obesity co-morbidities.(9-10)

We hypothesized that NLP could be used to identify, extract, and categorize follow-up provider information from free-text discharge summaries. In this paper, we describe an advanced NLP technique that automates the process of extracting and classifying follow-up provider information from free-text discharge summaries. We also report on an evaluation of the accuracy of this NLP system in extracting follow-up information from discharge summaries from two hospitals. The primary outcome was accuracy of the NLP system at extracting follow-up provider names when compared to physician reviewers. Secondary outcomes were accuracy of the system at extracting follow-up phone/fax numbers, and follow-up location information.

Methods:

Setting

We performed this study at two large urban Midwestern hospitals (Hospital A and B). Hospital A is served by a comprehensive electronic health record system (EHRs) and discharge summaries are entered electronically as free-text within the EHRs. Hospital B also has an EHRs, but discharge summaries at this institution are dictated and later transcribed as free-text into an electronic format.

NLP System

We used the Regenstrief EXtraction tool (REX) to process patient discharge summary reports. REX is a NLP software tool developed at the Regenstrief Institute in Indianapolis, IN. Briefly, REX is a rule-based NLP system written in Java that has successfully extracted patient data and concepts from microbiology reports, admission notes, and radiology reports.(11-13) The impetus for its development was the need by Regenstrief researchers for accurate NLP technology that did not take weeks to train and develop. As such, REX is not designed to identify every patient concept present in a report; instead, its main use and function is the rapid deployment of NLP technology for specific, targeted NLP tasks. It allows a user to quickly enhance and/or customize the NLP algorithms for particular use cases through modification of a knowledge-base external to the REX program itself. These changes can be made using a simple word processor. No programming experience is necessary since no modifications of REX code are required, although some knowledge of regular expressions is needed.

REX processes medical reports through three main modules. The "*structure module*" analyzes the format and structure of the report to identify report header information and titled sections of the report, e.g. chief complaint, history of present illness, physical exam, and family history. It also parses the text into individual sentences and each sentence into individual words. The "*concept module*" of REX searches each sentence for words or phrases that signal the likely presence of a target finding or concept. REX uses a series of regular expressions and algorithmic rules to detect these concept words/phrases. The "*context module*" determines the context that the targeted concept occurs in. REX does this by using methods similar to those used in the concept detection phase i.e. by utilizing a series of regular expressions and algorithmic rules. An important aspect of this module is the examination by REX of a 'window' of words surrounding the concept phrase to determine context such as positive, negative, and historical. These windows of words can vary in length depending on type of report and sentence structure, (conjunctions and/or commas in the sentence can increase window size) but generally average 12 words (6 before the concept phrase and 6 after the concept phrase). We chose this design because it closely mimics how humans determine the meaning of freeform text.

Modifying REX to Extract Follow-Up Information

For this project, we added data extraction capabilities to REX. In previous projects, REX simply identified if a report contained mentions of a specific concept (such as heart failure or MRSA) and if the concept was found, REX determined the context (positive, negative, historical, etc) of that concept.(11-13) In this project, we needed REX to not only identify phrases indicating a concept, but also extract specific classes of information from the reports – i.e. provider names, phone/fax numbers and location information. Because REX did not have data extraction capabilities, we added this functionality to the core code of REX. **Figure 1** displays a flow diagram for how the enhanced REX NLP program processes discharge summaries to extract follow-up information.

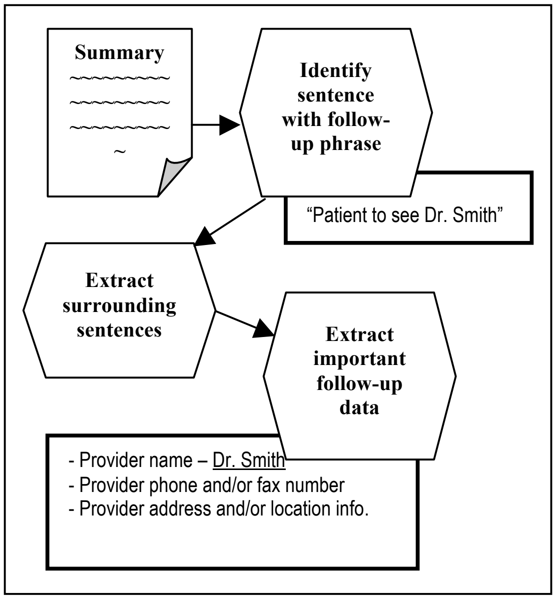
Summary
~~~~~~~~~
~~~~~~~~~
~~~~~~~~~
~
Identify sentence with follow-up phrase
"Patient to see Dr. Smith"
Extract surrounding sentences
Extract important follow-up data
- Provider name – Dr. Smith
- Provider phone and/or fax number
- Provider address and/or location info.

Figure 1: Processing of discharge summaries by REX

The first step in extracting follow-up provider information involved identifying the words and phrases clinicians typically use when referring to patient follow-up in discharge summaries. To do this, we randomly collected a training set of 100 discharge summaries from Hospital A and 372 discharge summaries from Hospital B prepared between July and September, 2007. We manually reviewed 90 of these to discover common follow-up words and phrases. Upon review of the reports, we discovered that follow-up information could be found in highly variable locations within each summary, and that REX would not be able to exploit section header information to identify where this information was located. In essence, REX needed to process the
~~~~~~~~~

entire summary, rather than focus on a specific section of it.

Once the portion of the discharge summary with reference to patient follow-up was found, we needed to identify and extract pertinent follow-up data. To perform this task, we developed algorithms and created 20 regular expressions to extract the three classes of follow-up data of interest, namely: (1) Provider name; (2) Provider phone/fax #; and (3) Provider clinic location/addresses.

Through review of our training set, we discovered that the specific follow-up information of interest was generally found within +/- 3 sentences of the sentence containing the follow-up phrase. We programmed REX to identify this group of sentences and used it as the 'sentence window' from which to extract the three classes of follow-up data. The extracted follow-up information was output into a tab-delimited ASCII text file suitable for importing into a database or spreadsheet.

Evaluation

After modifying REX using the training set, we applied it to our test set, which consisted of 717 discharge summaries (333 summaries from Hospital A and 384 from Hospital B). Discharge summaries in the test set belonged to adult patients discharged from the two hospitals in January and February of 2009. The set had been randomly-selected from a sample of patients identified (through queries of the EHRs) as having test results returning after hospital discharge – as such, there was need to identify the follow-up provider(s) to whom results should be communicated.

To determine accuracy of the NLP program at extracting follow-up provider information from discharge summaries, we compared it to information abstracted from the same discharge summaries through manual reviews. Each discharge summary in the test set was reviewed by two of three physician investigators (MCW, JC, SG). Two of the reviewers (MCW, SG) were board-certified Internal Medicine physicians and the other one (JC) was a 3rd-year resident in Internal Medicine. Reviewers independently abstracted the names for follow-up providers, the relevant follow-up phone and fax numbers, and names and addresses of the follow-up clinic locations. If there were disagreements between the reviewers on whether a particular piece of follow-up information was mentioned, the reviewers discussed the case to achieve consensus. To create the 'gold standard' we used consensus information by human reviewers, and added any correct additional information that was only found by REX and not by human reviewers. This additional information was only included after thorough reviews by two investigators (MCW, JF) to confirm actual presence of the information in the discharge summaries. The study was approved by the institutional review board at Indiana University School of Medicine

Analysis

We calculated three standard measures - precision (positive predictive value), recall (sensitivity) and *F*-measure (giving equal weight to precision and recall i.e. β=1) to describe performance of human reviewers and REX at extracting follow-up provider information (i.e. provider names, phone/fax, and location/address)(**Table 1**).

Table 1: Measures to determine accuracy of NLP program

Measurement	Formula
Precision	$TP/(TP+FP)$
Recall	$TP/(TP+FN)$
F-measure	$(\beta^2+1)P.R/(\beta^2P)+R$ or $2(P.R)/(P+R)$

TP = True Positive; FP = False Positive; FN = False Negative;
P = precision R = Recall β = 1

The gold standard was used to determine if extracted information by the physician reviewers or by the NLP program were correct or incorrect. Our evaluation by nature could not categorize any concepts as true negatives (concept absent or 'negated' in document and not found by reviewers or NLP) as we had no way of knowing what follow-up information was missing from the discharge summaries. In addition, follow-up phrases that are negated (i.e. "patient will not follow-up with Dr. Smith") are not typically encountered in discharge summaries. As such, specificity could not be reported.

All measures are reported with a 95% confidence interval (CI) at an alpha of 5% ($p<0.05$). When pair-wise comparisons of CIs do not overlap, the difference in the measures is statistically significant.

Results:

A total of 161 (23%) of the reviewed summaries - 89 (27%) from Hospital A and 72 (19%) from Hospital B - did not have any follow-up information mentioned. The remaining 556 summaries were analyzed for this study. **Table 2** presents the accuracy measures (with the respective 95% CI) for each reviewer and for the NLP program. When identifying follow-up provider names, the NLP program's precision was the same as MD3 but slightly inferior to MD1 and 2. NLP had the same recall and *F*-measure as all physician reviewers for this concept.

Table 2: Performance Measures of NLP and Human Reviewers

	Precision	Recall	F-Measure
Provider			
MD1	1.00 (0.99 – 1.0)	0.93 (0.91 - 0.95)	0.96 (0.96 - 0.97)
MD2	1.00 (0.99 – 1.0)	0.91 (0.87 - 0.94)	0.95 (0.94 – 0.97)
MD3	0.99 (0.97 – 1.0)	0.89 (0.86 - 0.92)	0.94 (0.92 - 0.95)
NLP	0.96 (0.95 - 0.97)	0.92 (0.90 - 0.94)	0.94 (0.93 - 0.95)
Phone/Fax			
MD1	0.99 (0.97 – 1.0)	0.91 (0.87 - 0.94)	0.95 (0.93 - 0.97)
MD2	1.00 (0.96 – 1.0)	0.79 (0.71 - 0.87)	0.89 (0.84 - 0.93)
MD3	1.00 (0.97 – 1.0)	0.75 (0.68 - 0.81)	0.86 (0.81 – 0.90)
NLP	0.99 (0.97 – 1.0)	0.92 (0.89 - 0.95)	0.96 (0.94 - 0.97)
Location			
MD1	0.99 (0.96 – 1.0)	0.91 (0.84 - 0.95)	0.95 (0.92 - 0.98)
MD2	1.00 (0.94 – 1)	0.95 (0.89 - 0.99)	0.97 (0.94 - 1.0)
MD3	0.97 (0.83 – 1.0)	0.36 (0.26 - 0.47)	0.52 (0.41 - 0.63)
NLP	0.83 (0.76 - 0.89)	0.82 (0.74 - 0.88)	0.82 (0.77 - 0.87)

MD1, **MD2**, and **MD3** represent the human reviewers.
NLP represents the natural language processing program REX.
NOTE: 95% Confidence Intervals (95% CI) are shown in parenthesis. We used an alpha=5% (p<0.05). When pairwise comparisons of CIs don't overlap, the difference is statistically significant.

The NLP program had the same precision as the providers at identifying phone information, but had better recall and *F*-measure than MD2 and MD3. However, for identifying location information, the NLP program performed lower on all measures compared to MD2, and also lower on precision and *F*-measure compared to MD1. It however outperformed MD3 on recall and *F*-measure, and had the same precision as this reviewer.

Performance of the NLP program at the two institutions was the same on all measures when identifying provider and phone/fax information. The *F*-measure however was statistically different when identifying location information - 0.77 (CI:0.70-0.84) compared to 0.89 (CI: 0.84-0.95). Precision and recall were not statistically different between the two institutions. As is evident from Table 2, there was variable performance between the human reviewers across several measures in the three categories of extracted information.

It took about 3-5 minutes for a human reviewer to go through the summary, while it took about 4 seconds for REX to extract data from each summary.

Discussion

We present the use of an NLP program to identify, extract, and categorize follow-up provider information from free-text discharge summaries. In 2010, Ruud et al used SAS® Text Miner software to extract follow-up information from discharge summaries.(14) Compared to the Ruud study, we achieved similar to but slightly more accurate results. Interestingly, and perhaps not surprisingly, both studies revealed that location was the most difficult category (~83% accuracy) for NLP to extract. Our NLP program had physician-like performance at extracting follow-up provider names and phone numbers on all measures, and had better recall than one provider. In identifying follow-up provider names and numbers, our program performed above 92% on all measures.

Performance of the NLP program was slightly lower than two of the reviewers when it came to extracting information about the location for follow-up, even though it performed better on recall and *F*-measure than the third reviewer. Manual reviews of errors related to location information revealed that the multitude of ways in which location information is recorded in summaries were not fully represented in our pattern-matching approach. We also observed that false positives could sometimes be attributed to our three sentence window before and after the concept of interest. Luckily, the evaluated discharge summaries, unlike regular notes, were decent at following grammatical rules, and this reduced the problem experienced in other studies where notes disregarded many grammatical conventions.(15)

One big advantage of our NLP program compared to humans was the speed with which it was able to perform the data extractions. Further, it was able to code the derived data into columns appropriate for easy queries. We did not observe much variability in performance by the NLP program between the two institutions across almost all the measures (except for the *F*-measure to identifying location information). This suggests the potential generalizability of our tool across institutions. Of note, there are significant differences between reviewers on several of the measures, and this signifies that relative to some reviewers, the NLP tool could actually be better.

Some limitations of this study deserve mention. The study was done at only two institutions and on summaries from one type of clinical team – as such, the system's performance might not be as good across institutions or teams. Our human reviewers were not blinded, but we made sure that disagreements were adjudicated, and that developers of the NLP system were not involved in any of the chart reviews. It also required about thirty hours of programming to modify and enhance REX to perform the current tasks.

We plan to further refine REX to especially improve its ability to extract information about follow-up locations. We will also evaluate the performance of the tool using summaries from services other than the medicine teams. In the future, we will make the NLP tool available to users to help automate look-up and extraction of follow-up provider information from discharge summaries. Eventually, we plan to augment the extracted information through queries against administrative databases. This will enable us to add additional follow-up details not documented in discharge summaries.

Conclusion

A natural language processing tool successfully extracted and classified follow-up information about providers, phone/fax numbers, and locations that were contained in free-text discharge summaries. It also performed this function many times faster than human reviewers.

Acknowledgements

This research was supported in part by the Indiana CTSI through grant 5KL2RR025760 from NIH/NCRR/CTSA. The project was also supported by grant K08HS018539 from AHRQ. The content of this manuscript are solely the responsibility of the authors and do not necessarily represent the official views of the funding agencies.

References

1. Forster AJ, Murff HJ, Peterson JF, Gandhi TK, Bates DW. The incidence and severity of adverse events affecting patients after discharge from the hospital. Ann Intern Med. 2003 Feb 4;138(3):161-7.

2. Kripalani S, LeFevre F, Phillips CO, Williams MV, Basaviah P, Baker DW. Deficits in communication and information transfer between hospital-based and primary care physicians: implications for patient safety and continuity of care. JAMA. 2007 Feb 28;297(8):831-41.

3. Moore C, Wisnivesky J, Williams S, McGinn T. Medical errors related to discontinuity of care from an inpatient to an outpatient setting. J Gen Intern Med. 2003 Aug;18(8):646-51.

4. Coleman EA. Falling through the cracks: challenges and opportunities for improving transitional care for persons with continuous complex care needs. J Am Geriatr Soc. 2003 Apr;51(4):549-55.

5. Roy CL, Poon EG, Karson AS, Ladak-Merchant Z, Johnson RE, Maviglia SM, et al. Patient safety concerns arising from test results that return after hospital discharge. Ann Intern Med. 2005 Jul 19;143(2):121-8.

6. Friedman C, Shagina L, Lussier Y, Hripcsak G. Automated encoding of clinical documents based on natural language processing. J Am Med Inform Assoc. 2004 Sep-Oct;11(5):392-402.

7. 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 Oct;34(5):301-10.

8. Melton GB, Hripcsak G. Automated detection of adverse events using natural language processing of discharge summaries. J Am Med Inform Assoc. 2005 Jul-Aug;12(4):448-57.

9. Uzuner O, Goldstein I, Luo Y, Kohane I. Identifying patient smoking status from medical discharge records. J Am Med Inform Assoc. 2008 Jan-Feb;15(1):14-24.

10. Uzuner O. Second i2b2 workshop on natural language processing challenges for clinical records. AMIA Annu Symp Proc. 2008:1252-3.

11. Friedlin J, Grannis S, Overhage JM. Using natural language processing to improve accuracy of automated notifiable disease reporting. AMIA Annu Symp Proc. 2008:207-11.

12. Friedlin J, McDonald CJ. Using a natural language processing system to extract and code family history data from admission reports. AMIA Annu Symp Proc. 2006:925.

13. Friedlin J, McDonald CJ. A natural language processing system to extract and code concepts relating to congestive heart failure from chest radiology reports. AMIA Annu Symp Proc. 2006:269-73.

14. Ruud KL, Johnson MG, Liesinger JT, Grafft CA, Naessens JM. Automated detection of follow-up appointments using text mining of discharge records. Int J Qual Health Care. 2010 Jun;22(3):229-35.

15. Barrows Jr RC, Busuioc M, Friedman C. Limited parsing of notational text visit notes: ad-hoc vs. NLP approaches. Proc AMIA Symp. 2000:51-5.