Extracting Information for Generating A Diabetes Report Card from Free Text in Physicians Notes

Ramanjot S Bhatia
University of Ottawa Heart Institute
Ottawa, Ontario.
Rbhatia@ottawaheart.ca

Amber Graystone
McMaster University
Hamilton, Ontario.
amber.graystone@medportal.ca

Ross A Davies
University of Ottawa Heart Institute
Ottawa, Ontario.
RADavies@ottawaheart.ca

Susan McClinton
University of Ottawa Heart Institute
Ottawa, Ontario.
SMcClinton@ottawaheart.ca

Jason Morin
National Research Council Canada
Ottawa, Ontario.
jason.morin@nrc-cnrc.gc.ca

Richard F Davies
University of Ottawa Heart Institute
Ottawa, Ontario.
RFDavies@ottawaheart.ca

Abstract

Achieving guideline-based targets in patients with diabetes is crucial for improving clinical outcomes and preventing long-term complications. Using electronic heath records (EHRs) to identify high-risk patients for further intervention by screening large populations is limited because many EHRs store clinical information as dictated and transcribed free text notes that are not amenable to statistical analysis. This paper presents the process of extracting elements needed for generating a diabetes report card from free text notes written in English. Numerical measurements, representing lab values and physical examinations results are extracted from free text documents and then stored in a structured database. Extracting diagnosis information and medication lists are work in progress. The complete dataset for this project is comprised of 81,932 documents from 30,459 patients collected over a period of 5 years. The patient population is considered high risk for diabetes as they have existing cardiovascular complications. Experimental results validate our method, demonstrating high precision (88.8-100%).

1 Introduction

A standard practice for care providers is to record patient consults using voice dictation. The voice dictation record is transcribed into free text and stored electronically. The nature of this text is narrative with a possibility of containing headings marking the boundaries of the paragraphs. This remains the medium of choice for storing key patient information as opposed to structured tables due to time constraints, uncertainty about the use of codes, classification limitations, and difficulty with the use of computer systems. The information being stored in machine readable format is not amenable to any form of statistical analysis or review as it exists (Mcdonald 1997, Lovis et al. 2000). The usefulness of mining information from this text has been stressed by many including Heinze et al. (2001) and Hripcsak et al. (1995). The information unlocked from the free text could be used for facilitating patient management, researching disease symptoms, analyzing diagnoses, epidemiological research, book keeping, etc. The free text in these documents has been shown to be less ambiguous than text in general unrestricted documents (Ruch et al. 2001) making it feasible to successfully apply extraction techniques using tools from IE and NLP. Natural language processing has been used to analyze free text in

Proceedings of the NAACL HLT 2010 Second Louhi Workshop on Text and Data Mining of Health Documents, pages 8–14,
Los Angeles, California, June 2010.

medical domain for decision support (Chapman et al. 2005), classifying medical problem lists (Meystre and Haug 2005), extracting disease related information (Xu et al. 2004), building dynamic medications lists (Pakhomov et al. 2002), building applications for better data management, and for diagnosis detection. (Friedman et al. 2004, Roberts et al. 2008, Liu and Friedman 2004).

Our goal is to automatically generate diabetes report cards from the free text in physicians' letters. The report card can be used to detect populations at risk for diabetes mellitus and track their vital information over a period of time. Previous work in similar area has seen Turchin et al. (2005) identify patients with diabetes from the text of physician notes by looking for mention of diabetes and predefined list of medication names. They use a manually created list of negation tokens to detect false examples. They compare the process to manual chart review and billing notes and show the automatic system performs at par with manual review with the advantage of it being highly efficient.

In Turchin et al. (2006) the authors use regular expressions to extract blood pressure values and change of treatment for hypertension. They use a set of regular expressions to detect the presence of a blood pressure related tag, which predicts that the sentence is likely to contain a blood pressure value. The value itself is then extracted using regular expressions. They identify the strength of the process in it being relatively simple, efficient and quick to setup, while its weakness is its lack of generalization. Voorham and Denig (2007) solve a similar problem as in here and extract information regarding diabetes from free text notes using a number centric approach. They identify all positive numerical values and then attach respective labels to the values. They use a keyword based approach with a four word token window and apply a character sequence algorithm to check for spelling errors.

Extracting relevant information from free text represents a challenging problem since the task can be considered to be a form of reverse engineering and is above the mere presence of keywords or patterns. It is necessary to generate semantic representations to understand the text. The free text document may contain multiple values for the same label, and it's important to be able to distinguish and choose the correct value. These values could be:

- multiple readings (in which case a predefined rule may be enough, e.g. choosing the smallest mean arterial blood pressure value)
- potential target values (which may or may not be important)
- values taken over a period of time
- values taken at different locations
- values reflecting family history
- change in a value and not the actual value
- values influenced by some external reasons (e.g. take medication if the weight is above a certain value).

Friedman and Hripcsak (1999) discuss some of the many problems of dealing with free text in medical domain. One method to resolve these problems is to build a full grammar tree and assign semantic roles to accurately interpret the text. However, generating full parse trees for medical text requires specialized parsers developed for the clinical domain (Freidman, 2005). It has been shown that shallow syntactic approaches can yield similar results to the ones using full syntactic details (Gildea & Palmer, 2002).

In this work we use shallow syntactic and semantic features (manually created concept list and WordNet, Miller 1995) to tag information relating to the numerical values extracted from the text. We use machine learning tool WEKA (Hall et al. 2009) to build binary classifiers that pick positive values from the list of values extracted from the document. Our method allows us to build a robust and extendible system which should be easily portable to texts from different institutions and other medical domains.

2 Method

Our method extends Voorham's work in using the numeric value centered approach while developing a robust way to disambiguate between multiple values in the same document. The information extracted for the report card is divided into four categories: demographic information, numerical measurement values, medication list, and diagnoses. We currently have access to only one source of information, the free text in physicians' notes, hence all of the information needed for the report card is extracted from these notes. The extraction of demographic information is achieved using reg-

ular expressions/pattern matching based techniques. The demographic information extracted is year of birth, date of encounter and gender. The gender information is determined using a heuristic, which counts the number of third person masculine and feminine pronouns present in the text. Numerical measurement values extracted include blood pressure (systolic and diastolic), LDL, HDL, HbA1C, weight, total cholesterol, fasting glucose, glucose (unspecified) and creatinine. The medication list extraction process uses a manually created database of applicable medications. The diagnosis detection involves negation detection in the sentences that mention diabetes using the NegEx algorithm (Chapman et al. 2001).

In this study we use shallow syntactic and semantic attributes to build a system that extracts the physical examination and laboratory results data. The values are extracted as numeric value-label pairs. The system is divided into three main parts (Figure 1): preprocessing stage, extraction of the numeric value-label pairs, and testing the validity of the extracted pairs.

Preprocessing: The documents were originally stored in Microsoft Word format (WordML). They are converted to XML using XSLT transformation. All formatting information is stripped except for bold and italic font information and paragraph boundaries

The paragraphs in the document are further broken down into sentences and tokens. We use OPENNLP Maxent[1] library to do sentence boundary detection and tokenization. OPENNLP Maxent is based on maximum entropy algorithms described in Ratnaparkhi (1998) and Berger et al. (1996). The OPENNLP statistical tagger is used to assign syntactic tags to the tokens.

Data Extraction: In this phase the system extracts all potential numerical values and assigns them labels. The system loops through all of the tokens in the document, testing for numerical values. It tests each numerical token against a set of regular expressions and assigns them a list of potential labels based on the regular expression it matches. The system takes into account the presence of a measurement unit and revises the potential list of labels based on the unit. For each potential label, using a knowledge base, the system looks for concepts that validate the labels. The closest possible validation is accepted as pairing. The Edit distance algorithm is used to test for matching concepts in order to account for any spelling errors. The concepts are searched within the constraints of the sentence.

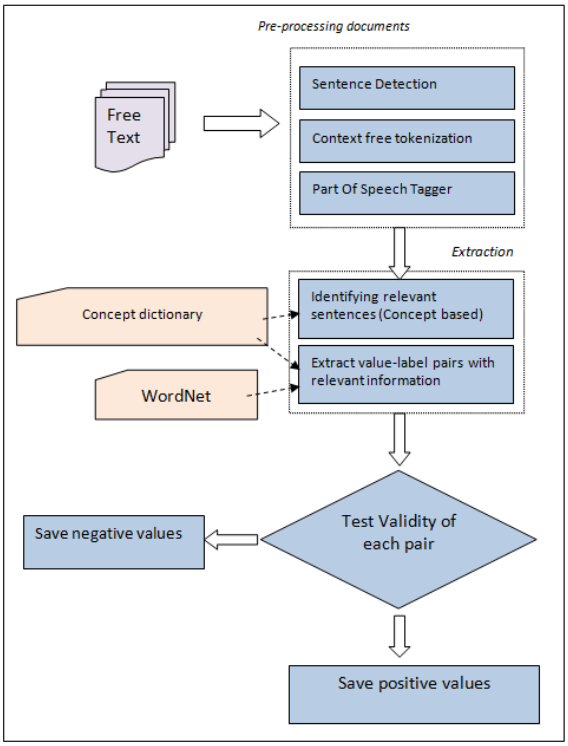

Figure 1 Process Flow diagram for the extraction process

In case multiple labels are validated because of the presence of multiple concepts in the same sentence, the label indicated by the closest concept is selected. For each pair, the system extracts a list of features which help to resolve for positive values. One exception to the sentence level boundary rule is: if no concepts are found in the sentence, and the sentence contains a third person singular inanimate pronoun, the search is extended to the previous sentence.

Testing Validity: The previous step extracts all possible label-numeric value pairs. As discussed earlier not all values are valid or of interest. In order to select positive values, binary classifiers were built for each label. The dataset used for training consisted of 900 documents (210 patients). The J48 (decision trees 4.5) and NBTree (Naïve Bayes decision trees) algorithms in WEKA were used to generate the machine learning classifiers.

[1] http://opennlp.sourceforge.net/

Features: The following is the list of features extracted for each pair.

a) Absolute distance between the label and the numerical value.
b) Label shared (Yes/No): Yes, if the same concept label is attached to another numerical value in the same document.
c) Closest verb token appearing left of the numerical value.
d) Presence of a modal verb (Yes/No)
e) Distance of numerical value from the modal verb (a positive value is assigned for the modal verb if it occurs before the numeric token, and a negative value when it appears after).
f) Conjunction present (Yes/No): If there is conjunction present between the label and numerical value or not.
g) Coreference present (Yes/No): If third person singular inanimate pronoun is present or not.
h) Negation concept present (Yes/No): True if there is any negation concept present in the vicinity of the numerical value/label. The negation concepts include not just negative statement markers, but also false cognates and other concepts collected by the domain experts. (e.g. systolic murmur or systolic volume do not indicate systolic pressure).
i) Locational Information token: The stemmed token is stored if it is recognized as a locational information token. The location information is deduced by generalizing each token and checking to see whether it resolves to one of many Locational cues in WordNet. The list of location indicators is presented in Figure 2. The cues are resolved against the WordNet hypernym definitions for that token.
j) Distance of numerical value from Locational token.
k) Temporal information token: Similar to (i), the stemmed token indicating temporal information. The temporal information token includes any tokens that indicate date or time. The list of temporal indicator cues in WordNet is shown in Figure 2.
l) Distance of the numerical value from the temporal token.

For features (c), (i) and (k) the tokens are stored in their uninflected form, achieved using Porter Stemmer. For the report card, in case of multiple positive values for the same label, the smallest value is selected. In the case of blood pressure, the smallest mean arterial pressure is selected.

Location
=>"Facility, installation" =>"Housing, lodging, living accommodations" =>"Facility, installation" =>"Passage" =>"Structure, construction" =>"Road, route" =>"Geographic point, geographical point" =>"Location"
Time
=>"time period, period of time, period" =>"time unit, unit of time" => "happening, occurrence, occurrent, natural event"

Figure 2 WordNet hypernym based generalization cues for location and time indicators

3 Evaluation

Evaluation was done using a test set consisting of 804 documents from 260 patients (50 percent had positive diagnosis for diabetes). The test set was created by a first year student at Michael G Degroote School of Medicine at McMaster University. The reviewer manually analyzed the notes and extracted final values that would appear on the report card along with a time stamp for each value to indicate the source document. The human reviewer took approximately 10 minutes per patient; in comparison the computer analyzed the data at 6.43 patients per minute.

Evaluation results testing the performance of the system using the manually coded test set are shown in Table 1 below.

	Value	Precision	Recall	F-measure
1	Blood Pressure	98.2	96.9	97.8
2	LDL	96.4	94.2	95.3
3	HDL	100	98.3	99.1
4	Creatinine	97.2	92.1	94.5
5	Weight	95.6	92.9	94.2
6	TC	93.1	98.1	95.5
7	Glucose	90.7	85.7	87.7
8	F Glucose	88.8	80.0	84.2
9	HbA1C	90.9	86.9	88.8

Table 1 Precision/Recall for numerical values

4 Results and Discussion

The precision, recall and f-measure for all nine label values extracted for the system along with the recall values for the human reviewer are listed in Table 1. The system demonstrates high precision in extracting and selecting positive numeric value-label pairs. Blood pressure is extracted with a precision of 98.2% and recall 96.9%. HDL and LDL values are easy to spot and extract as they usually occur without description. At the lower end of precision are fasting glucose, glucose and HbA1C where precision results are in the range of 88-90%. The majority of errors for all categories occurred due to problems in identifying numeric values because of typing errors.

Figure 3 shows an example of the level of complexity resolved using the algorithm developed here. The clinical documents frequently have multiple values for weight and blood pressure in a single document. The lab values do not have the same level of multiplicity but it can occur. In this example, the extraction step extracts all five values, and the classifier successfully rejects values #3 and #5. To comply with the report card's output requirements the lowest mean arterial pressure of the remaining three values is adopted, which is the correct response. This approach is extendible to build a slot-filler system for the values, which would allow the system to reason on its choice.

In previous work, the disambiguation of the values is only based on the presence of negation concepts within a pre-specified boundary. We extend this to include a simple need based co-reference, location and temporal information, and a heuristic approach to include the head verb (it only takes into account the closest verb, which may or may not be the governing verb). The system can successfully detect negative values such as target values, previous values, change in value or values measured elsewhere.

The information extracted is stored in a structured MySQL database. The system allows multiple views on this information. Figure 4 shows the output for blood pressure and creatinine for a patient that was created from the information extracted from the free text.

Blood pressure initially was 196/92 in the left arm and 194/90 in the right arm. Usually, the patient states, at home it is 140-150/80. The blood pressure subsequently decreased to 160/88 when I waited several minutes and had him calm down……Target blood pressure should be below 140/90.

Values Extracted:

1) 196/92 = {Loc: left arm, Verb: was}
2) 194/90 = {Loc: right arm, Verb: was, Conjunction: True}
3) 140-150/80 = {Loc: home, Co-ref: It, Verb: is}
4) 160/88 = {Verb: Decreased}
5) 140/90 = {Verb: be, Modal: should}

Figure 3 Example 1

At this time we have not evaluated the contribution of each feature individually, as this requires building a comprehensive test set; it remains as future work.

5 Conclusion

Our preliminary results demonstrate that the system performs with high precision and recall at the task of extracting numerical values. It also shows the ability to build a patient-chart abstractor within the restricted domain. The use of semantic and syntactic features enables the system to tag the values which permit the overall extraction process to generate more informative numeric value-label pairs. The use of machine learning algorithms coupled with a large enough learning dataset produces a robust system that should work reliably on similar data from any source. We plan to test the system on a dataset obtained from the free text notes of endocrinologists at a different health institution to validate the generalization of the algorithm. The next step for the Diabetes Report Card is to extract the list of medications and track any changes in medication, dosage and frequency.

Acknowledgments

A special thanks to Michael Domenic Corbo for doing the manual review and creating the gold standard dataset.

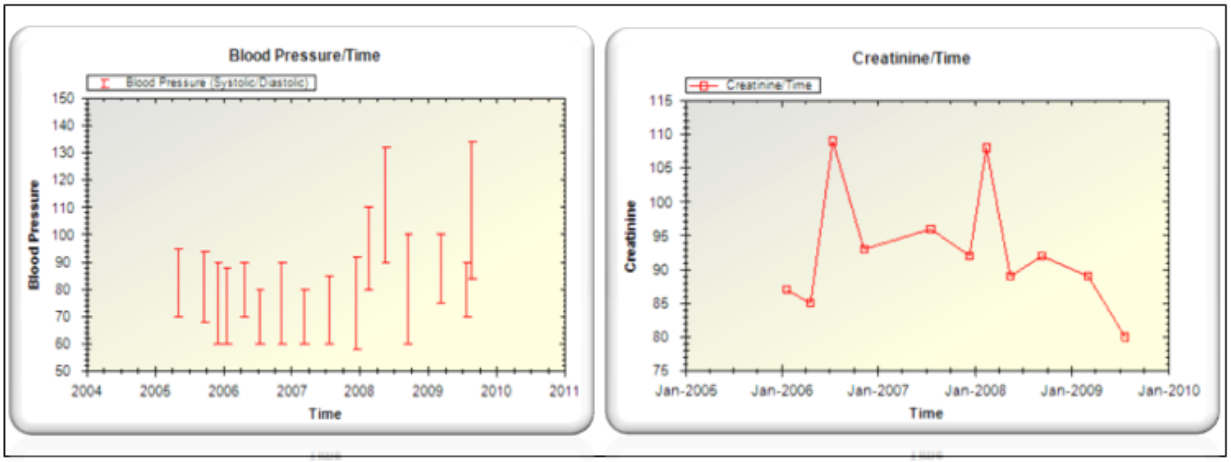

Figure 4 System Output: Automatically generated graphs for blood pressure and creatinine values for a patient

6 References

Berger, A. L., Pietra, V. J., & Pietra, S. A. (1996). A maximum entropy approach to natural language processing. *Computational Linguistics* , 39-71.

Chapman, W. W., Christensenb, L. M., Wagnera, M. M., Haugb, P. J., Ivanova, O., Dowlinga, J. N., et al. (2005). Classifying free-text triage chief complaints into syndromic categories with natural languages processing. *Artificial Intelligence in Medicine* , 31-40.

Chapman, W., Bridewell, W., Hanbury, P., Cooper, G., & Buchanan, B. (2001). Evaluation of negation phrases in narrative clinical reports. *Proc AMIA Symp* , 105-114.

Freidman, C. (2005). Semantic Text Parsing for Patient Records. In *Medical Informatics* (pp. 423-448). Springer US.

Friedman, C., & Hripcsak, G. (1999). Natural Language Processing and Its Future in Medicine. *Acad Med* , 890-895.

Friedman, C., Shagina, L., Lussier, Y., & Hripcsak, G. (2004). Automated Encoding of Clinical Documents based on Natural Language Processing. *Journal of American Medical Informatics Association* .

Gildea, D., & Palmer, M. (2002). The Necessity of Syntactic Parsing for Predicate Argument Recognition. *Association for Computational Linguistics*, (pp. 239-246).

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The WEKA Data Mining Software: An Update. *SIGKDD Explorations* .

Heinze, D. T., Morsch, M. L., & Holbrook, J. (2001). Mining free-text medical records. *AMIA*, (pp. 254-258).

Hripcsak, G., Friedman, C., Alderson, P., DuMouchel, W., Johnson, S., & Clayton, P. (1995). Unlocking clinical data from narrative reports: a study of natural language processing. *Ann Intern Med* , 681-689.

Liu, H., & Friedman, C. (2004). CliniViewer: a tool for viewing electronic medical records based on natural language processing and XML. *MedInfo* , 639-643.

Lovis, C., Baud, R. H., & Plancheb, P. (2000). Power of expression in the electronic patient record: structured data or narrative text? *International Journal of Medical Informatics* , 101-110.

Mcdonald, C. J. (1997). The Barriers to Electronic Medical Record Systems and How to Overcome Them. *Journal of the American Medical Informatics Association* , 213-221.

Meystre, S., & Haug, P. J. (2005). Automation of a problem list using natural language processing. *BMC Medical Informatics and Decision Making* , 5-30.

Miller, G. A. (1995). WordNet: A Lexical Database for English. *Communications of the ACM , 38*, 39-41.

Pakhomov, S. V., Ruggieri, A., & Chute, C. G. (2002). Maximum entropy modeling for mining patient medication status from free text. *Proceedings of the American Medical Informatics*, (pp. 587–591).

Ratnaparkhi, A. (1998). *Maximum Entropy Models for Natural Language Ambiguity Resolution.* Phd Thesis.

Roberts, A., Gaizauskas, R., Hepple, M., & Guo, Y. (2008). Mining clinical relationships from patient narratives. *Natural Language Processing in Biomedicine (BioNLP) ACL Workshop.*

Ruch, P., Baud, R., Geissbuhler, A., & Rassinoux, A.-M. (2001). Comparing general and medical texts for information retreival based on natural language processing: An inquiry into lexical disambiguation., (pp. 261-266).

Turchin, A., Kohane, I., & Pendergrass, M. (2005). Identification of patients with diabetes from the text of physician notes in the electronic medical record. *Diabetes Care* , 1794-1795.

Turchin, A., Kolatkar, N., Grant, R. W., Makhni, E. C., Pendergrass, M. L., & Einbinder, J. S. (2006). Using regular expressions to abstract blood pressure and treatment intensification information from the text of physician notes. *Journal of the American Medical Informatics Association* , 691-696.

Voorham, J., & Denig, P. (2007). Computerized Extraction of Information on the Quality of Diabetes Care from Free Text in Electronic Patient Records of General Practitioners. *The Journal of the American Medical Informatics Association* , 349-354.

Xu, H., Anderson, K., Grann, V. R., & Friedman, C. (2004). Facilitating Research in Pathology using Natural Language Processing. *Proc AMIA Symp*, (p. 1057).