Research Paper ■

Natural Language Processing Framework to Assess Clinical Conditions

HENRY WARE, CHARLES J. MULLETT, MD, PHD, V. JAGANNATHAN, PHD

Abstract **Objective:** The authors developed a natural language processing (NLP) framework that could be used to extract clinical findings and diagnoses from dictated physician documentation.

Design: De-identified documentation was made available by i2b2 Bio-informatics research group as a part of their NLP challenge focusing on obesity and its co-morbidities. The authors describe their approach, which used a combination of concept detection, context validation, and the application of a variety of rules to conclude patient diagnoses.

Results: The framework was successful at correctly identifying diagnoses as judged by NLP challenge organizers when compared with a gold standard of physician annotations. The authors overall kappa values for agreement with the gold standard were 0.92 for explicit textual results and 0.91 for intuited results. The NLP framework compared favorably with those of the other entrants, placing third in textual results and fourth in intuited results in the i2b2 competition.

Conclusions: The framework and approach used to detect clinical conditions was reasonably successful at extracting 16 diagnoses related to obesity. The system and methodology merits further development, targeting clinically useful applications.

■ **J Am Med Inform Assoc.** 2009;16:585–589. DOI 10.1197/jamia.M3091.

Introduction

Informatics for Integrating Biology at the Bedside (i2b2) Challenge

To stimulate further advancements in the field of healthcare data extraction from free text clinical documents, the group Informatics for Integrating Biology at the Bedside (i2b2) challenged teams to identify obesity and other co-morbidities from a corpus of dictated discharge summaries. This is their second Shared Task and Workshop Competition and the first one focused on de-identification of clinical documents and detecting smoking status.[1,2] We chose to participate as it offered us a chance to compare the success of our methods and techniques with others in the field, and, the physician among the authors hopes that NLP extraction techniques will allow free-text physician notes to "win out" over more restrictive templated physician notes for documentation in this era of computerized clinical decision support and automated billing and coding systems.

Background

By extracting data and information from dictated medical reports, NLP can potentially be used to facilitate and improve the process of medical care—particularly when coupled with contemporary clinical information systems. Academic investigators have largely focused on extraction of findings and problems,[3–7] while the healthcare industry has developed commercial business products.[8–11] Computer-assisted coding for the submission of billing claims is a thriving market.[12–15] Our group has previously published an assessment of the use of commercial NLP engines for medication information extraction.[16]

Methods

Research Corpus

The research corpus was supplied by the NLP Challenge organizers. It consisted of a training set of approximately 700 de-identified discharge summaries, each abstracted by two physician experts for the presence or absence of 16 different patient conditions (obesity, diabetes mellitus, etc). The abstractions were further broken down into "textual" and "intuitive." The textual analysis registered that the patient condition was directly mentioned in the text of the discharge summary, and each scoring for "textual" analysis had four possible values: Yes, No, Questionable, Unknown. The intuitive analysis identified those patient conditions that may not have been directly mentioned, but whose presence or absence could be deduced or inferred by the text of the document. Each scoring for "intuitive" analysis had three possible values: Yes, No, and Unknown. Therefore, the

Affiliations of the authors: MedQuist, Inc. (HW, VJ), Morgantown, WV; West Virginia University (CJM), Morgantown, WV.

The authors thank the organizers of the NLP challenge for providing us the opportunity to work with marked-up training data. Such corpora have been sorely lacking in this discipline. The results presented and discussed in this paper were also presented at the NLP challenge workshop organized by i2b2 concurrent with AMIA 2008. The authors also thank our sponsor for this project, Mark Ivié (CTO of MedQuist) for allowing us to participate and work on this motivating challenge.

Correspondence: V. Jagannathan, PhD, Vice President of Research, Medquist, Inc, 235 High Street, Suite 214, Morgantown, WV 26505; e-mail: <juggy@medquist.com>.

Received for review: 12/02/2008; accepted for publication: 04/07/2009.

training dataset that accompanied the 700 discharge summary documents contained roughly 16 × 2 × 700 individual data points.

The challenge organizers focused on obesity and its comorbidities.

Overview of Procedure

Documents were first preprocessed to remove unnecessary elements. For each principal concept, the document was textually evaluated by a three step process for the concept itself and for any secondary concepts. First, each concept would be searched for by regular expression (explicit textual). Then, for each match found, the neighborhood of the match was examined for context. Third, the contextualized matches were brought together in an evaluation for that concept. Once all the principal and secondary concepts had been evaluated, a determination was made for the principal concept.

Preprocessing

Portions of the documents that were not germane to the task at hand were removed.

Many of the documents contained a family history section. While these sections did occasionally contain information about the current patient's current condition, they were mainly focused on the conditions of other family members. In early trials, we determined that these mentions of familial conditions confused the bag-of-words-based context finder. Therefore, this section was removed in a preprocessing step.

Some of the documents contained text identifying an override of the computerized clinical decision alert along with the responsible physician. An example is shown below:

> ECASA (ASPIRIN ENTERIC COATED) 325 MG PO QD
>
> Override Notice: Override added on 11/9/01 by
>
> FUDD, ELMER J., M.D.
>
> on order for COUMADIN PO (ref # 00944322)
>
> POTENTIALLY SERIOUS INTERACTION: ASPIRIN & WARFARIN

Reason for override: md aware

These medico legal snippets sometimes included references to medications the patient was not taking. In addition, these sections confused human readers. Consequently, they were removed in the preprocessing step.

Finally, when the first line of the document contained an admitting diagnosis, this line was removed, as this often represented a working hypothesis and not a confirmed diagnosis. Admitting diagnosis sections were not removed.

Feature Extraction

For feature selection we used a combination of guided and manual methods.

Most features were identified by medical relevance. We attempted to first answer the question "why did the physician grade the discharge summary this way?" and then to mimic that process.

Medical concepts were linked to the original 16 whenever they were suspected of affecting an annotation. For example, when it was conjectured that a patient's history of a myocardial infarction was used by the physician evaluators to establish a diagnosis of CAD, myocardial infarction was added as a concept, linked to CAD.

The most common class of the secondary concepts was medications. Medications were added semantically: with brand names, generic names, and common abbreviations. The concept would generally be at the level of a drug class, such as: steroids, non-steroidal anti—inflammatory drugs (NSAIDs), ACE inhibitors, loop diuretics, thiazide diuretics, β blockers, and antidepressants. A few concepts were more specific, such as nitroglycerin or albuterol. For example, use of inhaled albuterol suggests the diagnosis of asthma.

Treatments were also used as secondary concepts. For instance, gastric bypass surgery suggested obesity, and pressure stockings indicated venous insufficiency.

Additionally, word bigrams were considered with an inverse document frequency. This brought out a few phrases such as "GI Bleed" in a GERD context.

Each concept had a synonym list. For example, the list of synonyms for albuterol was "salbutamol", "albuterol", "ventolin", "proventil", and "proair". For medications, we used the Apelon terminology engine to provide these synonym sets. For a given drug class, this list might contain hundreds of synonyms.

Numerical features were also supported. For example, obesity could be inferred from a numeric BMI or from a numeric weight and height or from a numeric weight alone (for, e.g., >90 kg). Many other numerical features were recognized including ejection fraction, the lipid panel results, and hemoglobin A1c values.

Synonym lists were converted to regular expressions for matching against the clinical documents. The regular expressions were generally case insensitive and matched whole words, but these features could be overridden. For example, the synonym list for myocardial infarction was "MI", "ami", "imi", "myocardial infarc", "septal infarc", and "heart attack".

Concept Context

When a concept match was found, the neighborhood around the concept was examined. This approach follows NegEx and related work by Chapman et al.[17,18] The examination looked for the presence of key phrases within the neighborhood in order. Our implementation differed from Chapman's in two important ways. First, we processed at the level of characters rather than words—just regular expressions without preliminary lexing. Second we used a neighborhood bounded by any punctuation mark, rather than a five word window.

The first phrases searched for were obliterators. If an obliterator from a concept's obliterator list was found, the match was determined to be not a match at all. For example, for the concept osteoarthritis, the phrase "arthritis" was taken as a possible match. However, the phrase "rheumatoid arthritis" was used as an obliterator as it describes a different disease from osteoarthritis.

The next phrases searched for were pseudonegators such as "no further …"; The pseudonegators look like negations but are not. Then, we looked for hypothetics such as "evaluate for." These indicate that the condition is possible, but not certain. Next, history markers were considered. These are strings like "h/o"—meaning history of—that indicate the temporality of the condition. Finally, plain negaters were sought, such as "denies" or "not signs of."

Rules

We developed and inserted sets of rules to assist with the intuitive determinations. These rules were framed in a domain specific language (DSL) written in the scala programming language.

In the development of the rules, there was always a tension between what seemed to achieve better agreement with the annotators in the training set and what was more medically correct. Following the annotators too closely can lead to over fitting, matching noise rather than signal. By contrast, a rule which does not at first seem medically sound may proxy some real condition of the patient or thought process of the annotator.

An illustrative example is Congestive heart failure (CHF). When the concept is not directly mentioned in the text, several ways of inferring the patient's status were used.

This condition is unequivocally indicated by a low numeric ejection fraction; we took 50% as a threshold. Most patients with CHF will be treated with ACE inhibitors; however ACE inhibitors are also used to treat other conditions such as hypertension. In treating hypertension ACE inhibitors are often prescribed with a thiazide diuretic; this combination for CHF is rarer. So an ACE inhibitor without a thiazide diuretic is a possible treatment for CHF; but we would also like to see a condition associated with CHF. The findings we identified were a history of heart transplant, the presence of pulmonary edema, or the finding of a high numeric wedge pressure. Sample rules described in the DSL are shown below.

```
val chfMed = hasAce and! hasThiazide;

Val chfSymptom = hasHeartTransplant |

hasPulmonaryEdema |

hasHighWedgePressure:

Val chfRules = hasLowEjectionFraction |

(chfMed and chfSymptom);
```

Technologies and Support Systems Used

Expert and rule-base systems have a long history in healthcare, going back to the early efforts with Mycin[19,20] and Internist.[21] There are several commercial and open source solutions that implement forward chaining, backward chaining or even blackboard systems.[22] We chose to implement our solutions from scratch as our goals were more experimental. The difficult part of building the system was to develop the feature set—the implementation of the inference logic was more straight forward. We relied on Apelon terminology environment to determine relevant medications for conditions used in the challenge.[23]

Evaluation Metrics Used

We used Cohen's kappa metric to measure the progress of our evolving solution versus the physician gold-standard document annotators with the training set of documents. The kappa statistic for two person inter-rater agreement attempts to generate a measurement of agreement beyond that expected by chance. The main reason for the selection of this metric was that a good number of answers could simply be guessed correctly—i.e., "the person does not suffer from X" will be mostly true and sometimes overwhelmingly true.[24] Use of this measure allowed us to focus on optimizing a single metric, as opposed to a suite of metrics.

The NLP challenge organizers provided a utility to measure six metrics—micro and macro versions of precision, recall and F-measure. These measures are reported in the results section.

Results and Discussion

Results

The contest results were graded versus the gold standard using an F-measure score. This metric allows use of a single number to describe both precision and recall. The metric can be calculated by document to generate an F-Micro score and by category to generate an F-macro score. The F-macro score weights documents from rare classes much more heavily—in this case, documents in the "Questionable" category were very heavily weighted.

Table 1 and Table 2 show our results against the testing set of documents. Overall, our total average F-measure scores for the textual and intuitive measurements were very similar for the test set as they were for the training set, suggesting that either the two sets of documents were very similar, or that our system is reasonably robust.

Disease-based kappa scores comparing our system versus the gold standard of the annotators with adjustment for inter-annotator agreement are provided in the Jamia "online only" version of this article at http://www.jamia.org.

For our system, the F-Micro represents largely the behavior of the synonym lists and rules; whereas the difference between the F-Micro and the F-macro scores largely indicate the success of the contextual algorithms. This is because most mentions of a term are in a positive context. So merely recognizing the concept by rules and synonyms results in a good F-Micro score. The F-macro, by contrast rests largely on correctly determining the context in which a concept appears, because instances in a non-positive context dominate the F-macro score.

Table 3 shows the computation of the overall kappa. Our overall kappa values for agreement with the gold standard were 0.92 for explicit textual results and 0.91 for intuited results.

Discussion

Textual

General textual performance of the system was good; F-Micro was over 0.98 for most diseases. For three categories, asthma, osteoarthritis, and obstructive sleep apnea, the system outscored the average estimated human versus the gold standard. One feature these conditions share is relatively little context: in particular they are unlikely to be negated.

Our least successful textual area by F-Micro was for coronary artery disease. We were unable to get a good handle on when a mention of a stenosed artery on cardiac catheterization was scored as textual CAD, when it was scored as intuitive CAD and when it was not read as indicative of any CAD. For the challenge, we went with a short synonym list with no linkages or presence of terms related to arteries. Our F-Micro in this area was 0.903.

Table 1 ■ Textual Analysis of the Testing Data Set

			Textual Judgement			
Disease	P-Micro	P-Macro	R-Micro	R-Macro	F-Micro	F-Macro
Obesity	0.9696	0.7373	0.9696	0.4910	0.9696	0.4891
Depression	0.9783	0.9407	0.9783	0.9757	0.9783	0.9572
Hypertriglyceridemia	0.9921	0.9960	0.9921	0.8000	0.9921	0.8730
Gallstones	0.9822	0.9896	0.9822	0.8057	0.9822	0.8695
OSA	0.9920	0.9817	0.9920	0.6651	0.9920	0.6565
Asthma	0.9881	0.8995	0.9881	0.8721	0.9881	0.8563
CAD	0.9034	0.8642	0.9034	0.8988	0.9034	0.8791
PVD	0.9921	0.9886	0.9921	0.9754	0.9921	0.9819
Gout	0.9861	0.9607	0.9861	0.9753	0.9861	0.9678
Diabetes	0.9682	0.9837	0.9682	0.7169	0.9682	0.7949
CHF	0.9335	0.7653	0.9335	0.8417	0.9335	0.7931
Venous Insufficiency	0.9803	0.7500	0.9803	0.9899	0.9803	0.8283
GERD	0.9861	0.9815	0.9861	0.7404	0.9861	0.7359
OA	0.9801	0.9541	0.9801	0.9788	0.9801	0.9659
Hypercholesterolemia	0.9681	0.9023	0.9681	0.7590	0.9681	0.8123
Hypertension	0.9461	0.8383	0.9461	0.7747	0.9461	0.8031
System	P-Micro	P-Macro	R-Micro	R-Macro	F-Micro	F-Macro
Textual	0.9718	0.8314	0.9718	0.7542	0.9718	0.7821

OSA = obstructive sleep apnea; CAD = coronary artery disease; PVD = peripheral vascular disease; CHF = congestive heart failure; GERD = gastroesophageal reflux disease; OA = osteoarthritis.

Intuitive

Our performance on diagnoses evaluated as "questionable" was, perhaps, our weakest overall area. Optimizing our system for this category was generally difficult, as there were relatively few questionable documents in the training set. However, we did not focus as much effort on this area as we could have. Across all sixteen intuitive categories, we correctly identified zero questionable documents in the test set. As a result, our F-macro scores were consistently higher in categories where no questionable documents were present.

Results otherwise were generally good but not as good as the textual results. Five of the 16 categories received an F-Micro of 0.98 or better. The best categories were hypertriglyceridemia and Gout, the least successful category was high cholesterol.

General Observations

We identified concepts in clinical documents using a collection of synonymous terms. If these concepts were found to be in the right context, they provided a reasonable basis for

Table 2 ■ Intuitive Analysis of Test Data

			Intuitive Judgment			
Disease	P-Micro	P-Macro	R-Micro	R-Macro	F-Micro	F-Macro
Obesity	0.9821	0.9812	0.9821	0.9824	0.9821	0.9818
Depression	0.9706	0.9472	0.9706	0.9709	0.9706	0.9584
Hypertriglyceridemia	0.9959	0.9789	0.9959	0.9789	0.9959	0.9789
Gallstones	0.9857	0.9916	0.9857	0.9562	0.9857	0.9729
OSA	0.9939	0.9855	0.9939	0.6659	0.9939	0.6589
Asthma	0.9597	0.8928	0.9597	0.9703	0.9597	0.9259
CAD	0.9585	0.9733	0.9585	0.6377	0.9585	0.6387
PVD	0.9785	0.9742	0.9785	0.6377	0.9785	0.6392
Gout	0.9920	0.9692	0.9920	0.9954	0.9920	0.9818
Diabetes	0.9749	0.9704	0.9749	0.9704	0.9749	0.9704
CHF	0.9521	0.9678	0.9521	0.6409	0.9521	0.6376
Venous Insufficiency	0.9578	0.8776	0.9578	0.7536	0.9578	0.8013
GERD	0.9249	0.9455	0.9249	0.5746	0.9249	0.5903
OA	0.9481	0.9508	0.9481	0.6210	0.9481	0.6192
Hypercholesterolemia	0.9072	0.9107	0.9072	0.9162	0.9072	0.9070
Hypertension	0.9484	0.9271	0.9484	0.9207	0.9484	0.9239
System	P-Micro	P-Macro	R-Micro	R-Macro	F-Micro	F-Macro
Intuitive	0.9654	0.6410	0.9654	0.6399	0.9654	0.6404

OSA = obstructive sleep apnea; CAD = coronary artery disease; PVD = peripheral vascular disease; CHF = congestive heart failure; GERD = gastroesophageal reflux disease; OA = osteoarthritis.

Table 3 ▪ Kappa Metric Overall for the Test Set

Test All Textual				
	Y	N	Q	U
Y	2098	9	5	94
N	10	34	0	22
Q	6	0	8	1
U	78	22	4	5653
K1				NA
K2				92.4
Test All Intuitive				
	Y	N	Q	U
Y	2131	102	5	22
N	10148	4998	9	0
Q	6	0	0	0
U	0	0	0	0
K1				NA
K2				91.4

X-Axis: Gold Scoring; Y-Axis: Program Scoring.
K1: Annotator Kappa agreement overall – not available.
K2: Overall results for program versus gold Kappa on the Test set.

making conclusions regarding the patient condition. Whether our methods are scalable to other clusters of diagnoses with only a reasonable amount of work is another question. We expect that analyzing large volumes of dictated data and using supervised learning techniques to guide the selection and creation of rules could allow us to successfully and semi-automatically scale this current approach.

For a variety of applications, ranging from documenting conditions present on admission to the identification of variations in care from published clinical guidelines, the approach investigated in this i2b2 Obesity Challenge appears promising.

Conclusions

The results show that our methods were successful as we finished fourth in the intuitive scoring and placed third in the textual scoring in the NLP challenge contest. We conclude that our method of optimizing the matching produced excellent results and provides a viable framework for future applications.

References ▪

1. Uzuner O, Luo Y, Szolovits P. Evaluating the State-of-the-art in automatic de-identification. J Am Med Inform Assoc 2007;14(5), September/October:550–63.
2. Uzuner O, Goldstein I, Luo Y, Kohane I. Identifying patient smoking status from medical discharge records. J Am Med Inform Assoc 2008;15(1), Jan/Febr:14–24.
3. Sager N, Friedman C, Lyman MS. Medical Language Processing: Computer Management of Narrative Data, Addison-Wesley, 1987.
4. Friedman C. “Towards a comprehensive medical language processing system: methods and issues”, Proceedings of AMIA, Annual Fall Symposium 1997:595–9.
5. Hripcsak G, Kuperman GJ, Friedman C. Extracting findings from narrative reports: Software transferability and sources of physician disagreement. Methods Inf Med Jan 1998;37(1):1–7.
6. Meystre S, Haug PJ. Automation of a problem list using natural language processing. BMC Med Inform Decis Mak Aug 31 2005;5:30.
7. Hazlehurst B, Frost HR, Sittig DF, Stevens VJ. MediClass: A system for detecting and classifying encounter-based clinical events in any electronic medical record. J Am Med Inform Assoc September–Oct 2005;12(5):517–29.
8. Mamlin BW, Heinze DT, McDonald CJ. Automated extraction and normalization of findings from cancer-related free-text radiology reports. AMIA Annu Symp Proc 2003:420–4.
9. Heinze DT, Morsch ML, Holbrook J. Mining free-text medical records. Proc AMIA Symp 2001:254–8.
10. Denecke K, Bernauer J. Extracting specific medical data using semantic structures: In: Bellazzi R, Abu-Hanna A, and Hunter J, editors. Lecture Notes in Computer Science: New York: Springer; 2007, LNAI 4594, p. 257–64.
11. Friedman C, Shagina L, Lussier Y, Hripcsak G. Automated encoding of clinical documents based on natural language processing. J Am Med Inform Assoc September–Oct 2004;11(5): 392–402.
12. Heinze D, Morsch M, Sheffer RE, et al. LifeCode®—A deployed application for automated medical coding. AI Mag 2001, Summer.
13. Morris W, Heinze D, Warner Jr HR, et al. Assessing the accuracy of an automated coding system in emergency medicine. Proceedings of the 2,000 AMIA Annual Fall Symposium, Jamia, November 2000.
14. Resnik P, Niv M, Nossal M, et al. Using intrinsic and extrinsic metrics to evaluate accuracy and facilitation in computer-assisted coding. Perspectives in Health Information Management Computer Assisted Coding Conference Proceedings, Fall 2006.
15. Clark C, Good K, Jezierny L, et al. Identifying smokers with a medical extraction system. J Am Med Inform Assoc 2008;15(1), Jan/Febr:26–39.
16. Jagannathan V, Mullett CJ, Arbogast JG, et al. Assessment of commercial NLP engines for medication information extraction from dictated clinical notes. Int J Med Inform 2009 Apr;78(4): 284–91.
17. Chapman WW, Bridewell W, Hanbury P, et al. “Evaluation of Negation Phrases in Narrative Clinical Reports”, Proceedings of AMIA Symposium 2001:105–9.
18. Chapman WW, Dowling JN, Chu DL. Context: An algorithm for identifying contextual features from clinical text. In: BioNLP Workshop of the Association for Computational Linguistics Prague, Czech Republic, pp 81–8, 2007.
19. Shortliffe EH, Axline SG, Buchanan BG, et al. An artificial intelligence program to advise physicians regarding antimicrobial therapy. Comput Biomed Res 1973;6:544–60.
20. Shortliffe EH, Davis R, Axline SG, et al. Computer-based consultations in clinical therapeutics: Explanation and rule acquisition capabilities of the MYCIN system. Comput Biomed Res 1975;8:303–20.
21. Miller RA, Pople HE, Myers JD. Internist-1, an experimental computer-based diagnostic consultant for general internal medicine. N Engl J Med 1982 Aug 19;307(8):468–76.
22. Jagannathan V, Dodhiawala RT, Baum LS (eds.). Blackboard architecture and applications. In: Perspectives in Artificial Intelligence Series, Academic Press, 1989.
23. Distributed Terminology System. [Online], 2006. Available at: http://www.apelon.com/products/white%20papers/DTS%20White%20Paper%20V34.pdf. Accessed December 2008.
24. Dawson B, Trapp RG. Basic and Clinical Biostatistics, 2nd edn, Norwalk, CT: Appleton & Lange, 1994.