Research Paper ■

A Rule-based Approach for Identifying Obesity and Its Comorbidities in Medical Discharge Summaries

NINAD K. MISHRA, MD, MS, DAVID M. CUMMO, JAMES J. ARNZEN, JASON BONANDER, MA

Abstract **Objective:** Evaluate the effectiveness of a simple rule-based approach in classifying medical discharge summaries according to indicators for obesity and 15 associated co-morbidities as part of the 2008 i2b2 Obesity Challenge.

Methods: The authors applied a rule-based approach that looked for occurrences of morbidity-related keywords and identified the types of assertions in which those keywords occurred. The documents were then classified using a simple scoring algorithm based on a mapping of the assertion types to possible judgment categories.

Measurements: Results for the challenge were evaluated based on macro F-measure. We report micro and macro F-measure results for all morbidities combined and for each morbidity separately.

Results: Our rule-based approach achieved micro and macro F-measures of 0.97 and 0.77, respectively, ranking fifth out of the entries submitted by 28 teams participating in the classification task based on textual judgments and substantially outperforming the average for the challenge.

Conclusions: As shown by its ranking in the challenge results, this approach performed relatively well under conditions in which limited training data existed for some judgment categories. Further, the approach held up well in relation to more complex approaches applied to this classification task. The approach could be enhanced by the addition of expert rules to model more complex medical reasoning.

■ **J Am Med Inform Assoc.** 2009;16:576–579. DOI 10.1197/jamia.M3086.

Introduction

The use of Natural Language Processing (NLP) to classify clinical records has been the subject of considerable research.[1–7] These studies include the classification of triage diagnoses according to illnesses, radiology reports according to the presence of pneumonia, and medical discharge summaries according to patient smoking status. The approaches taken often include applying machine learning techniques, such as naive Bayesian classifiers, decision trees, and support vector machines (SVMs), and sometimes augmenting these methods with human knowledge.[1–4] Several studies involve a combination of rule-based or statistical classifiers with multi-step information extraction processes that include various combinations of lexicons, expert-crafted rules, spell-checkers, syntactic and semantic parsers, and part-of-speech taggers, in addition to other components.[5–7] Such systems can be complex and can require substantial effort to design and build. A consideration with machine learning techniques is that their performance can be unstable when judgment categories contain few documents.[8]

This article describes our entry for the 2008 i2b2 Obesity Challenge. The contest was organized by Informatics for Integrating Biology & the Bedside (i2b2), a National Center for Biomedical Computing based at Partners Healthcare System in Boston, MA. The challenge involved a multi-class, multi-label document classification task focused on indicators for obesity and its co-morbidities found in medical discharge summaries.[9] We investigated the effectiveness of a relatively simple rule-based approach to the classification task. The approach involved keyword identification, negation detection, and simple scoring rules. We chose this approach because we expected it to require limited development effort, and we were interested in seeing how effectively it would perform on a clinical corpus containing few training documents in some judgment categories, which might be problematic for a typical machine learning approach.

Methods

The NLP task in the 2008 i2b2 Obesity Challenge involved classifying medical discharge summaries into several judgment categories for obesity and 15 associated co-morbidities. Two types of judgments were provided: textual judgments based on explicit indicators for the morbidities in the documents, and intuitive judgments based on what was implied about the morbidities in the documents. Teams had the option of classifying the documents based on either type of judgments or both.

Affiliations of the authors: Centers for Disease Control and Prevention (NKM, JB), Atlanta, GA; Northrop Grumman (DMC, JJA), Atlanta, GA.

The authors thank i2b2 for organizing the 2008 i2b2 Obesity Challenge. The authors thank Chapman et al for making NegEx available for non-commercial use, and thank Wendy Chapman, PhD, for suggesting ConText as an enhancement to our current approach.

Correspondence: Dr. Ninad Mishra Centers for Disease Control and Prevention, 1600 Clifton Rd, Mail Stop E76, Atlanta, GA 30333; e-mail: <nmishra@cdc.gov>.

Received for review: 12/01/08; accepted for publication: 04/07/09.

Data for the challenge were released in two sets: (1) a training set of 730 annotated discharge summaries for development and training purposes, and (2) a test set of 507 discharge summaries without annotations for the evaluation. In the textual training set, the Y and U categories were well represented with tens or hundreds of documents for most morbidities. In contrast, the N and Q categories contained fewer training documents, ranging from zero to 23 for each morbidity. Due to concerns that the limited amount of training data in two of the four judgment categories would substantially hinder the effectiveness of a machine learning algorithm, we opted to use a rule-based approach.

Our team chose to participate in the classification task based on textual judgments since our approach looked for keywords directly related to each morbidity. The textual judgments consisted of the following four categories: (a) the morbidity is "present" (labeled "Y" for "yes"), (b) the morbidity is "absent" (labeled "N" for "no"), (c) occurrence of the morbidity is "questionable" (labeled "Q"), or (d) the morbidity is "unmentioned" (labeled "U").

Our approach involved three steps:

1. text preprocessing
2. identification of keyword occurrences and associated assertion types
3. document scoring and classification

The text preprocessing step was done by a custom script written in the Perl programming language. The script performed text clean-up and modification to improve the effectiveness of the keyword identification step. Text modifications were made using regular expression pattern matching and substitution. The preprocessing script made the following changes to each discharge summary: (a) removed the "Family History" section since text here is not relevant to the patient's current condition, (b) changed question marks ("?") to the word "questionable" to improve assertion type detection, and (c) changed commas (",") to periods (".") to restrict assertion modifiers to the most immediate terms they modify. A more detailed discussion of these modifications is available in a JAMIA online data supplement at http://www.jamia.org.

In the keyword identification step, the discharge summaries were examined for the occurrence of keywords associated with each morbidity, along with the type of assertion in which each keyword occurred. The possible assertion types were (a) positive (e.g., "Diabetes: diet controlled"), (b) negative (e.g., "no significant CAD"), and (c) questionable (e.g., "borderline HTN"). This step used the NegEx negation detection application developed by Chapman et al.[10] A key NegEx component is its dictionary of clinical terms and several types of negation phrases. We made substantial customizations to the NegEx dictionary to tailor it to this classification task. In particular, the default list of clinical terms was completely replaced with a custom list of keywords associated with the morbidities targeted in this task. In addition, the list of conditional possibility terms (e.g., "no history of", "might be ruled out for") was repurposed to store terms pertaining to other family members (e.g., "maternal", "father", "cousin") so keywords identified with this code could be ignored in the document scoring and classification step. A more detailed discussion of these modifications is available in a JAMIA online data supplement at http://www.jamia.org.

NegEx used the terms in its dictionary for identifying morbidities and assertion types in the discharge summaries. The NegEx output consisted of modified discharge summaries containing markup around any identified keywords. This markup indicated the assertion type in which each keyword occurred.

The output of the NegEx algorithm was passed through the document scoring and classification step. This step was performed by a custom Perl script that calculated the total number of positive, negative, and questionable assertions for each morbidity in each discharge summary. For each document, this resulted in three "scores" for each morbidity, one for each assertion type. The scoring process ignored any keyword occurrences pertaining to other family members as indicated by the terms assigned to the conditional possibility codes in the NegEx dictionary, as well as any keywords identified with a special code indicating keywords to be ignored. The discharge summaries were then assigned to a judgment category for each morbidity based on the assertion type with the highest total. Positive assertions corresponded to the "Y" judgment, negative assertions corresponded to the "N" judgment, questionable assertions corresponded to the "Q" judgment, and the absence of keyword occurrences corresponded to the "U" judgment.

For ties between several assertion types, three different tie-breaking rules were developed: (a) positive-weighted tie breaking, in which ties between positive and non-positive assertions resulted in a "Y" judgment, (b) negative-weighted tie-breaking, in which ties between negative and non-negative assertions resulted in an "N" judgment, and (c) questionable-weighted tie-breaking, in which a tie between questionable and non-questionable assertions resulted in a "Q" judgment (please see Tables 1, 2, and 3, available in a JAMIA online data supplement at http://www.jamia.org).

In scenarios in which the weighted assertion type did not participate, we made an arbitrary rule that the weighted judgment could not win the tie-breaker and we would default to the least positive judgment available. For example, in the positive-weighted scenario in which a tie exists between negative and questionable assertions and the positive assertion type is not involved, the resulting tie-breaker judgment is "N." The same outcome applies in the questionable-weighted scenario in which a tie exists between positive and negative assertions. In the negative-weighted scenario, a tie between positive and questionable assertions results in a tie-breaker judgment of "Q" since it is the least positive of the non-weighted judgments.

The initial training set for the challenge was released in mid-Mar 2008, allowing us time over several months to test and make adjustments to our approach before the release of the test set. Repeated runs against the training set were used to tune the various components of our classification system. Tuning tasks included adjusting the keyword lists for morbidities, negation terms, and questionable assertion terms, and trying different tie-breaker rules. The test set was released at the end of Jun, and teams were allowed three days to evaluate the data and submit their results.

Table 5 ■ F-measure Results for Classification of the Test Set, Ranked by Macro F-measure. CDC Entries Shown within the Context of i2b2 Best Results and Overall Average Results

System	Micro F-Measure	Macro F-Measure
i2b2 best—textual task	0.9723	0.8052
CDC—positive-weighted tie-breaker	0.9704	0.7718
CDC—negative-weighted tie-breaker	0.9685	0.7391
CDC—questionable-weighted tie-breaker	0.9685	0.7383
i2b2 average—textual task	0.91	0.56

CDC = Centers for Disease Control.

Results and Discussion

Prior to the release of the test set, we were able to achieve overall macro F-measures between 0.74 and 0.76 against the training set (please see Table 4, available in a JAMIA online data supplement at http://www.jamia.org).

We submitted three entries to the challenge, each using a different tie-breaking rule. The micro and macro F-measure results are listed in Table 5, along with the overall best results and the average results for the textual classification task. The entry using the positive-weighted tie-breaking rule produced the best results of the entries we submitted. These results were similar to our results from evaluating the training set. Our best evaluation results were substantially better than the overall average for this task and within 0.04 of the macro F-measure of the overall best results. Our top entry ranked fifth out of the results submitted by 28 teams.

Detailed results for each morbidity are shown in Table 6. Out of the 16 morbidities involved in this year's challenge, our rule-based approach achieved macro F-measure scores above 0.8 for 12 of them (75%) and above 0.9 for five morbidities (31%). Our worst results occurred in trying to classify the discharge summaries for obesity, the primary morbidity of interest. Our system misclassified all the documents that should have been classified as "N" or "Q" for that morbidity (three documents for each of those judgments). Initially, we suspected this was caused by failing to include any keywords in our customized NegEx dictionary that occurred in any of the discharge summaries for these judgments. However, post-challenge analysis by a medical reviewer indicated several of the documents in the "Q" category contained multiple terms that *in combination* could indicate a possibility of obesity (e.g., dyslipidemia, NIDDM, hypertension, herniated disk). At least one of the documents in the "N" category included insulin-dependent diabetes as a morbidity, which could be an indicator that the patient is less likely to be obese, as opposed to a situation in which diabetes is non-insulin-dependent. For the "Q" judgments, our system was not able to assess that some terms, either individually or in combination, might indicate obesity was only *possible* rather than almost *certain*. Similarly, for the "N" judgments, our system did not include the concept that some terms could be contra-indicators for a morbidity. Our custom dictionary only consisted of clinical terms that were strong indicators of a particular morbidity. Due to macro-averaging, missing all the "N" and "Q" judgments had a substantial effect on our results for obesity. Based on the results of the challenge, poor performance on these two judgment categories for obesity appeared to be a common problem for other teams as well, with eight of the top ten entries having macro F-measures for obesity below 0.50.[11]

During the tuning process, the factor that increased the performance of this approach most substantially was customization of the morbidity keyword list in the NegEx dictionary, followed by customization of the lists of terms used for identifying questionable and negated assertions. More minor improvements were contributed by the text preprocessing steps.

There are several possibilities for future improvement of our rule-based approach. Based on the error analysis of the missed "N" and "Q" judgments for obesity, a prime area for improvement could be adding probabilities to the clinical terms in the dictionary to indicate if they are strong indicators or contra-indicators for a morbidity, or indicators of some questionable possibility. The post-processing scoring rules would also need to be updated to handle this new information. However, this would only apply to individual terms. Assessing multiple clinical terms in combination would require more complicated rules, possibly expert-crafted. This would require more substantial changes to our system. Rather than repurposing the conditional possibility code in NegEx to identify keywords related to other family members, an alternate approach would involve the use of the ConText[12] algorithm which includes the negation-detection features of NegEx but also allows the identification of other contextual features, such as to whom a keyword applies (i.e., the patient or someone else). The preprocessing step that changes question marks to the word "questionable" should be updated to use a text pattern that excludes question marks that occur at sentence boundaries. This would minimize the chance of having two sentences concatenated together to avoid the possibility of negation and questionable assertion terms crossing sentence boundaries to affect keywords to which they should not apply. However, the documents in this particular classification task

Table 6 ■ F-measure Results for Each Morbidity, Ranked by Macro F-measure (Positive-Weighted Entry)

Morbidity	Micro F-Measure	Macro F-Measure
Obesity	0.9757	0.4917
GERD	0.9841	0.6524
OSA	0.9901	0.6564
CHF	0.9173	0.7645
Hypertension	0.9501	0.8089
Hypertriglyceridemia	0.9901	0.8308
CAD	0.9014	0.8357
Hypercholesterolemia	0.9721	0.8607
Gallstones	0.9803	0.8684
Venous insufficiency	0.9862	0.8811
Diabetes	0.9722	0.8886
PVD	0.9724	0.9380
Asthma	0.9921	0.9434
Depression	0.9763	0.9566
OA	0.9781	0.9635
Gout	0.9861	0.9678

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

appeared to have few question marks at sentence boundaries, so this change may be unlikely to yield substantial performance improvements in this set of documents.

A practical consideration is that the customized lists of morbidity keywords and terms for identifying assertion types were manually created. This can involve a substantial amount of manual effort. Future research could investigate automated feature selection techniques to augment the creation of these keyword lists to reduce the level of human effort.

In the system we developed, representations of expert knowledge primarily exist in the lists of clinical terms and the various types of assertion modifiers (e.g., negation terms, pseudo-negation terms, etc) stored in our customized NegEx dictionary. Domain-specific knowledge does not exist as expert-crafted rules in our system. Rules in the pre-processing step are tied to domain-specific considerations at a superficial level, but they do not represent deep expert knowledge. The addition of domain-specific rules built on expert medical knowledge has the potential for enhancing the performance of this approach, particularly for situations requiring the assessment of multiple clinical terms to arrive at a judgment.

Conclusions

We applied a relatively simple rule-based approach in our entry to the 2008 i2b2 Obesity Challenge. The classification strategy involved looking for morbidity keywords and the types of assertions in which they occurred, and then classifying the documents based on scores assigned to the various morbidity/judgment combinations. As indicated by its fifth-place ranking and its performance relative to the averages for the textual classification task, this strategy performed reasonably well on the i2b2 Obesity Challenge data containing widely varying numbers of documents across morbidity/judgment categories, with few documents in some judgment categories. The overall results also indicate that this relatively simple approach held up well in comparison to more complicated strategies applied in other entries to the challenge. The approach relies substantially on the NegEx negation detection algorithm and tailoring keyword lists to this particular task. The keyword list creation and customization relies on manual inspection of training data to identify terms to be used to classify documents. These keyword lists may need to be customized or recreated to use this strategy on different classification tasks. However, this strategy also has the potential to perform reasonably well when limited example data are available for training a machine learning algorithm. The approach could be enhanced further by the addition of domain-specific rules designed to model the reasoning of a medical expert.

References ■

1. Wilcox AB, Hripcsak G. The role of domain knowledge in automating medical text report classification. J Am Med Inform Assoc 2003;10:330–8.
2. Chapman WW, Cooper GF, Hanbury P, et al. Creating a text classifier to detect radiology reports describing mediastinal findings associated with inhalational anthrax and other disorders. J Am Med Inform Assoc 2003;10(5):494–503.
3. Olszewski RT. Bayesian classification of triage diagnoses for the early detection of epidemics. Proc of the 16th Int FLAIRS Conference, pp 412–6, 2003.
4. Clark C, Good K, Jezierny L, et al. Identifying smokers with a medical extraction system. J Am Med Inform Assoc 2008;15:36–9.
5. Fiszman M, Chapman WW, Aronsky D, Evans RS, Haug PJ. Automatic detection of acute bacterial pneumonia from chest x-ray reports. J Am Med Inform Assoc 2000;7:593–604.
6. Chapman WW, Christensen LM, Wagner MM, et al. Classifying free-text triage chief complaints into syndromic categories with natural language processing. Artif Intell Med 2005;33(1):31–40.
7. Zeng QT, Goryachev S, Weiss S, et al. Extracting principal diagnosis, co-morbidity and smoking status for asthma research: Evaluation of a natural language processing system. BMC Med Inform Decis Mak 2006;6:30.
8. Dumais ST, Platt J, Heckerman D, Sahami M. Inductive learning algorithms and representations for text categorization. In: Proceedings of the 7th ACM International Conference on Information and Knowledge Management. New York: ACM Press; 1998. p. 148–155.
9. Uzuner O, Szolovits P, Kohane I, Second i. 2b2 shared-task and workshop [internet]. i2b2: Informatics for integrating biology and the bedside, 2008. Available at: https://www.i2b2.org/NLP/. Accessed: Aug 15, 2008.
10. 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;34:301–10.
11. Uzuner O. Recognizing obesity and co-morbidities in sparse data. J Am Med Inform Assoc 2009;16:560–70.
12. Chapman WW, Chu D, Dowling JN. ConText: An algorithm for identifying contextual features from clinical text. In: Proceedings of the 2007 ACL Workshop on Biological, Translational, and Clinical Language Processing (BioNLP); 2007 Jun 29; Prague, Czech Republic. Madison, WI: Omnipress; 2007, pp 81–8.