Improving Classification of Medical Assertions in Clinical Notes

Youngjun Kim
School of Computing
University of Utah
Salt Lake City, UT
youngjun@cs.utah.edu

Ellen Riloff
School of Computing
University of Utah
Salt Lake City, UT
riloff@cs.utah.edu

Stéphane M. Meystre
Department of Biomedical Informatics
University of Utah
Salt Lake City, UT
stephane.meystre@hsc.utah.edu

Abstract

We present an NLP system that classifies the assertion type of medical problems in clinical notes used for the Fourth i2b2/VA Challenge. Our classifier uses a variety of linguistic features, including lexical, syntactic, lexico-syntactic, and contextual features. To overcome an extremely unbalanced distribution of assertion types in the data set, we focused our efforts on adding features specifically to improve the performance of minority classes. As a result, our system reached 94.17% micro-averaged and 79.76% macro-averaged F_1-measures, and showed substantial recall gains on the minority classes.

1 Introduction

Since the beginning of the new millennium, there has been a growing need in the medical community for Natural Language Processing (NLP) technology to provide computable information from narrative text and enable improved data quality and decision-making. Many NLP researchers working with clinical text (i.e. documents in the electronic health record) are also realizing that the transition to machine learning techniques from traditional rule-based methods can lead to more efficient ways to process increasingly large collections of clinical narratives. As evidence of this transition, nearly all of the best-performing systems in the Fourth i2b2/VA Challenge (Uzuner and DuVall, 2010) used machine learning methods.

In this paper, we focus on the *medical assertions* classification task. Given a medical problem mentioned in a clinical text, an assertion classifier must look at the context and choose the status of how the medical problem pertains to the patient by assigning one of six labels: *present, absent, hypothetical, possible, conditional*, or *not associated with the patient*. The corpus for this task consists of discharge summaries from Partners HealthCare (Boston, MA) and Beth Israel Deaconess Medical Center, as well as discharge summaries and progress notes from the University of Pittsburgh Medical Center (Pittsburgh, PA).

Our system performed well in the i2b2/VA Challenge, achieving a micro-averaged F_1-measure of 93.01%. However, two of the assertion categories (*present* and *absent*) accounted for nearly 90% of the instances in the data set, while the other four classes were relatively infrequent. When we analyzed our results, we saw that our performance on the four minority classes was weak (e.g., recall on the *conditional* class was 22.22%). Even though the minority classes are not common, they are extremely important to identify accurately (e.g., a medical problem *not associated with the patient* should not be assigned to the patient).

In this paper, we present our efforts to reduce the performance gap between the dominant assertion classes and the minority classes. We made three types of changes to address this issue: we changed the multi-class learning strategy, filtered the training data to remove redundancy, and added new features specifically designed to increase recall on the minority classes. We compare the performance of our new classifier with our original

i2b2/VA Challenge classifier and show that it performs substantially better on the minority classes, while increasing overall performance as well.

2 Related Work

During the Fourth i2b2/VA Challenge, the assertion classification task was tackled by participating researchers. The best performing system (Berry de Bruijn et al., 2011) reached a micro-averaged F_1-measure of 93.62%. Their breakdown of F_1 scores on the individual classes was: *present* 95.94%, *absent* 94.23%, *possible* 64.33%, *conditional* 26.26%, *hypothetical* 88.40%, and *not associated with the patient* 82.35%. Our system had the 6th best score out of 21 teams, with a micro-averaged F_1-measure of 93.01%.

Previously, some researchers had developed systems to recognize specific assertion categories. Chapman et al. (2001) created the NegEx algorithm, a simple rule-based system that uses regular expressions with trigger terms to determine whether a medical term is *absent* in a patient. They reported 77.8% recall and 84.5% precision for 1,235 medical problems in discharge summaries. Chapman et al. (2007) also introduced the ConText algorithm, which extended the NegEx algorithm to detect four assertion categories: *absent, hypothetical, historical*, and *not associated with the patient*. Uzuner et al. (2009) developed the Statistical Assertion Classifier (StAC) and showed that a machine learning approach for assertion classification could achieve results competitive with their own implementation of Extended NegEx algorithm (ENegEx). They used four assertion classes*: present, absent, uncertain in the patient*, or *not associated with the patient*.

3 The Assertion Classifier

We approach the assertion classification task as a supervised learning problem. The classifier is given a medical term within a sentence as input and must assign one of the six assertion categories to the medical term based on its surrounding context.

3.1 Pipeline Architecture

We built a UIMA (Ferrucci and Lally, 2004; Apache, 2008) based pipeline with multiple components, as depicted in Figure 1. The architecture includes a section detector (adapted from earlier work by Meystre and Haug (2005)), a tokenizer (based on regular expressions to split text on white space characters), a part-of-speech (POS) tagger (OpenNLP (Baldridge et al., 2005) module with trained model from cTAKES (Savova et al., 2010)), a context analyzer (local implementation of the ConText algorithm (Chapman et al., 2001)), and a normalizer based on the LVG (Lexical Variants Generation) (LVG, 2010) annotator from cTAKES to retrieve normalized word forms.

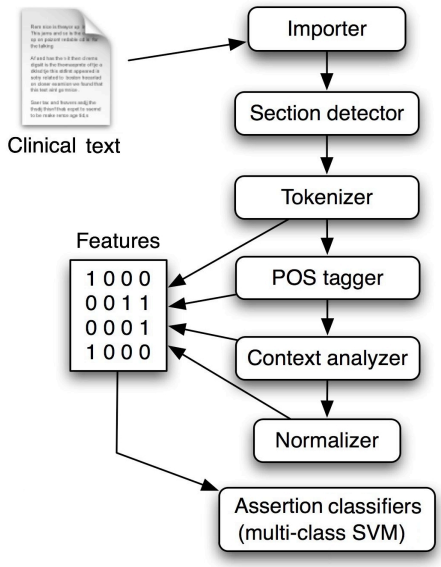

Figure 1: System Pipeline

The assertion classifier uses features extracted by the subcomponents to represent training and test instances. We used LIBSVM, a library for support vector machines (SVM), (Chang and Lin, 2001) for multi-class classification with the RBF (Radial Basis Function) kernel.

3.2 Original i2b2 Feature Set

The assertion classifier that we created for the i2b2/VA Challenge used the features listed below, which we developed by manually examining the training data:

Lexical Features: The medical term itself, the three words preceding it, and the three words following it. We used the LVG annotator in Lexical Tools (McCray et al., 1994) to normalize each word (e.g., with respect to case and tense).

Syntactic Features: Part-of-speech tags of the three words preceding the medical term and the three words following it.

Lexico-Syntactic Features: We also defined features representing words corresponding to several parts-of-speech in the same sentence as the medical term. The value for each feature is the normalized word string. To mitigate the limited window size of lexical features, we defined one feature each for the nearest preceding and following adjective, adverb, preposition, and verb, and one additional preceding adjective and preposition and one additional following verb and preposition.

Contextual Features: We incorporated the ConText algorithm (Chapman et al., 2001) to detect four contextual properties in the sentence: *absent* (negation)*, hypothetical, historical*, and *not associated with the patient.* The algorithm assigns one of three values to each feature: *true*, *false*, or *possible*. We also created one feature to represent the Section Header with a string value normalized using (Meystre and Haug, 2005). The system only using contextual features gave reasonable results: F_1-measure overall 89.96%, *present* 91.39%, *absent* 86.58%, and *hypothetical* 72.13%.

Feature Pruning: We created an UNKNOWN feature value to cover rarely seen feature values. Lexical feature values that had frequency < 4 and other feature values that had frequency < 2 were all encoded as UNKNOWNs.

3.3 New Features for Improvements

After the i2b2/VA Challenge submission, we added the following new features, specifically to try to improve performance on the minority classes:

Lexical Features: We created a second set of lexical features that were case-insensitive. We also created three additional binary features for each lexical feature. We computed the average tf-idf score for the words comprising the medical term itself, the average tf-idf score for the three words to its left, and the average tf-idf score for the three words to its right. Each binary feature has a value of *true* if the average tf-idf score is smaller than a threshold (e.g. 0.5 for the medical term itself), or *false* otherwise. Finally, we created another binary feature that is *true* if the medical term contains a word with a negative prefix.[1]

Lexico-Syntactic Features: We defined two binary features that check for the presence of a comma or question mark adjacent to the medical term. We also defined features for the nearest preceding and following modal verb and wh-adverb (e.g., where and when). Finally, we reduced the scope of these features from the entire sentence to a context window of size eight around the medical term.

Sentence Features: We created two binary features to represent whether a sentence is long (> 50 words) or short (<= 50 words), and whether the sentence contains more than 5 punctuation marks, primarily to identify sentences containing lists. [2]

Context Features: We created a second set of ConText algorithm properties for negation restricted to the six word context window around the medical term**.** According to the assertion annotation guidelines, problems associated with allergies were defined as *conditional*. So we added one binary feature that is *true* if the section headers contain terms related to allergies (e.g., "Medication allergies").

Feature Pruning: We changed the pruning strategy to use document frequency values instead of corpus frequency for the lexical features, and used document frequency > 1 for normalized words and > 2 for case-insensitive words as thresholds. We also removed 57 redundant instances from the training set. Finally, when a medical term co-exists with other medical terms (problem concepts) in the same sentence, the others are excluded from the lexical and lexico-syntactic features.

3.4 Multi-class Learning Strategies

Our original i2b2 system used a 1-vs-1 classification strategy. This approach creates one classifier for each possible pair of labels (e.g., one classifier decides whether an instance is *present* vs. *absent*, another decides whether it is *present* vs. *conditional*, etc.). All of the classifiers are applied to a new instance and the label for the instance is determined by summing the votes of the classifiers. However, Huang et al. (2001) reported that this approach did not work well for data sets that had highly unbalanced class probabilities.

Therefore we experimented with an alternative 1-vs-all classification strategy. In this approach, we

[1] Negative prefixes: ab, de, di, il, im, in, ir, re, un, no, mel, mal, mis. In retrospect, some of these are too general and should be tightened up in the future.

[2] We hoped to help the classifier recognize lists for negation scoping, although no scoping features were added per se.

create one classifier for each type of label using instances with that label as positive instances and instances with any other label as negative instances. The final class label is assigned by choosing the class that was assigned with the highest confidence value (i.e., the classifier's score).

4 Evaluation

After changing to the 1-vs-all multi-class strategy and adding the new feature set, we evaluated our improved system on the test data and compared its performance with our original system.

4.1 Data

The training set includes 349 clinical notes, with 11,967 assertions of medical problems. The test set includes 477 texts with 18,550 assertions. These assertions were distributed as follows (Table 1):

	Training (%)	Testing (%)
Present	67.28	70.22
Absent	21.18	19.46
Hypothetical	5.44	3.87
Possible	4.47	4.76
Conditional	0.86	0.92
Not Patient	0.77	0.78

Table 1: Assertions Distribution

4.2 Results

For the i2b2/VA Challenge submission, our system showed good performance, with 93.01% micro-averaged F_1-measure. However, the macro F_1-measure was much lower because our recall on the minority classes was weak. For example, most of the *conditional* test cases were misclassified as *present*. Table 2 shows the comparative results of the two systems (named 'i2b2' for the i2b2/VA Challenge system, and 'new' for our improved system).

	Recall		Precision		F_1-measure	
	i2b2	New	i2b2	New	i2b2	New
Present	97.89	98.07	93.11	94.46	95.44	**96.23**
Absent	92.99	94.71	94.30	96.31	93.64	**95.50**
Possible	45.30	54.36	80.00	78.30	57.85	**64.17**
Conditional	22.22	30.41	90.48	81.25	35.68	**44.26**
Hypothetical	82.98	87.45	92.82	92.07	87.63	**89.70**
Not patient	78.62	81.38	100.0	97.52	88.03	**88.72**
Micro Avg.	93.01	94.17	93.01	94.17	93.01	**94.17**
Macro Avg.	70.00	74.39	91.79	89.99	76.38	**79.76**

Table 2: Result Comparison of Test Data

The micro-averaged F_1-measure of our new system is 94.17%, which now outperforms the best official score reported for the 2010 i2b2 challenge (which was 93.62%). The macro-averaged F_1-measure increased from 76.38% to 79.76% because performance on the minority classes improved. The F_1-measure improved in all classes, but we saw especially large improvements with the *possible* class (+6.32%) and the *conditional* class (+8.58%). Although the improvement on the dominant classes was limited in absolute terms (+.79% F_1-measure for *present* and +1.86% for *absent*), the relative reduction in error rate was greater than for the minority classes: -29.25% reduction in error rate for *absent* assertions, -17.32% for *present* assertions, and -13.3% for *conditional* assertions.

	Present		Absent		Possible		Conditional		Hypothetical		Not patient	
	R	P	R	P	R	P	R	P	R	P	R	P
i2b2	**98.36**	93.18	94.52	95.31	48.22	**84.59**	9.71	**100.0**	86.18	**95.57**	55.43	**98.08**
+ 1-vs-all	97.28	94.56	95.07	94.88	57.38	75.25	27.18	77.78	**90.32**	93.33	**72.83**	95.71
+ Pruning	97.45	94.63	94.91	94.75	60.34	79.26	33.01	70.83	89.40	94.48	69.57	95.52
+Lex+LS+Sen	97.51	94.82	95.11	95.50	63.35	78.74	33.98	71.43	88.63	93.52	70.65	97.01
+ Context	97.60	**94.94**	**95.39**	**95.97**	**63.72**	78.11	**35.92**	71.15	88.63	93.52	69.57	96.97

Table 3: Cross Validation on Training Data: Results from Applying New Features Cumulatively (Lex=Lexical features; LS=Lexico-Syntactic features; Sen=Sentence features)

4.3 Analysis

We performed five-fold cross validation on the training data to measure the impact of each of the four subsets of features explained in Section 3. Table 3 shows the cross validation results when cumulatively adding each set of features. Applying the 1-vs-all strategy showed interesting results: recall went up and precision went down for all classes except *present*. Although the overall F_1-measure remained almost same, it helped to increase the recall on the minority classes, and we were able to gain most of the precision back (without sacrificing this recall) by adding the new features.

The new lexical features including negative prefixes and binary tf-idf features primarily increased performance on the *absent* class. Using document frequency to prune lexical features showed small gains in all classes except *absent*. Sentence features helped recognize *hypothetical* assertions, which often occur in relatively long sentences.

The *possible* class benefitted the most from the new lexico-syntactic features, with a 3.38% recall gain. We observed that many *possible* concepts were preceded by a question mark ('?') in the training corpus. The new contextual features helped detect more *conditional* cases. Five allergy-related section headers (i.e. "Allergies", "Allergies and Medicine Reactions", "Allergies/Sensitivities", "Allergy", and "Medication Allergies") were associated with *conditional* assertions. Together, all the new features increased recall by 26.21% on the *conditional* class, 15.5% on *possible*, and 14.14% on *not associated with the patient*.

5. Conclusions

We created a more accurate assertion classifier that now achieves state-of-the-art performance on assertion labeling for clinical texts. We showed that it is possible to improve performance on recognizing minority classes by 1-vs-all strategy and richer features designed with the minority classes in mind. However, performance on the minority classes still lags behind the dominant classes, so more work is needed in this area.

Acknowledgments

We thank the i2b2/VA challenge organizers for their efforts, and gratefully acknowledge the support and resources of the VA Consortium for Healthcare Informatics Research (CHIR), VA HSR HIR 08-374 Translational Use Case Projects; Utah CDC Center of Excellence in Public Health Informatics (Grant 1 P01HK000069-01), the National Science Foundation under grant IIS-1018314, and the University of Utah Department of Biomedical Informatics. We also wish to thank our other i2b2 team members: Guy Divita, Qing Z. Treitler, Doug Redd, Adi Gundlapalli, and Sasikiran Kandula. Finally, we truly appreciate Berry de Bruijn and Colin Cherry for the prompt responses to our inquiry.

References

Apache UIMA 2008. Available at http://uima.apache.org.

Jason Baldridge, Tom Morton, and Gann Bierner. 2005. OpenNLP Maxent Package in Java, Available at: http://incubator.apache.org/opennlp/.

Berry de Bruijn, Colin Cherry, Svetlana Kiritchenko, Joel Martin, and Xiaodan Zhu. 2011. Machine-Learned Solutions for Three Stages of Clinical Information Extraction: the State of the Art at i2b2 2010. J Am Med Inform Assoc.

Chih-Chung Chang and Chih-Jen Lin, LIBSVM: a Library for Support Vector Machines, 2001. Available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

Wendy W. Chapman, Will Bridewell, Paul Hanbury, Gregory F. Cooper, and Bruce G. Buchanan. 2001. A Simple Algorithm for Identifying Negated Findings and Diseases in Discharge Summaries. Journal of Biomedical Informatics, 34:301-310.

Wendy W. Chapman, David Chu, and John N. Dowling. 2007. ConText: An Algorithm for Identifying Contextual Features from Clinical Text. BioNLP 2007: Biological, translational, and clinical language processing, Prague, CZ.

David Ferrucci and Adam Lally. 2004. UIMA: An Architectural Approach to Unstructured Information Processing in the Corporate Research Environment. Journal of Natural Language Engineering, 10(3-4): 327-348.

Tzu-Kuo Huang, Ruby C. Weng, and Chih-Jen Lin. 2006. Generalized Bradley-Terry Models and Multiclass Probability Estimates. Journal of Machine Learning Research, 7:85-115.

i2b2/VA 2010 Challenge Assertion Annotation Guidelines. https://www.i2b2.org/NLP/Relations/assets/Assertion%20Annotation%20Guideline.pdf.

LVG (Lexical Variants Generation). 2010. Available at: http://lexsrv2.nlm.nih.gov/LexSysGroup/Projects/lvg.

Alexa T. McCray, Suresh Srinivasan, and Allen C. Browne. 1994. Lexical Methods for Managing Variation in Biomedical Terminologies. Proc Annu Symp Comput Appl Med Care.:235–239.

Stéphane M. Meystre and Peter J. Haug. 2005. Automation of a Problem List Using Natural Language Processing. BMC Med Inform Decis Mak, 5:30.

Guergana K. Savova, James J. Masanz, Philip V. Ogren, Jiaping Zheng, Sunghwan Sohn, Karin C. Kipper-Schuler, and Christopher G. Chute. 2010. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. J Am Med Inform Assoc., 17(5):507-513.

Özlem Uzuner and Scott DuVall. 2010. Fourth i2b2/VA Challenge. In http://www.i2b2.org/NLP/Relations/.

Özlem Uzuner, Xiaoran Zhang, and Sibanda Tawanda. 2009. Machine Learning and Rule-based Approaches to Assertion Classification. J Am Med Inform Assoc., 16:109-115.