Determining Assertion Status for Medical Problems in Clinical Records

Cheryl Clark, PhD, John Aberdeen, Matt Coarr, David Tresner-Kirsch, Ben Wellner, PhD, Alexander Yeh, PhD, Lynette Hirschman, PhD
MITRE Corporation, Bedford, MA 01730

Abstract

This paper describes the MITRE system entries for the 2010 i2b2/VA community evaluation "Challenges in Natural Language Processing for Clinical Data" for the task of classifying assertions associated with problem concepts extracted from patient records. Our best performing system obtained an overall micro-averaged F-score of 0.9343. The methods employed were a combination of machine learning (Conditional Random Field and Maximum Entropy) and rule-based (pattern matching) techniques.

1. Introduction

MITRE participated in the assertion classification subtask of the 2010 i2b2/VA community evaluation "Challenges in Natural Language Processing for Clinical Data." MITRE has been developing a system for detecting negated and uncertain or speculative information in clinical reports. Key features of the system are (1) the use of linguistic structure (i.e., cue scope) rather than proximity to determine whether a concept is influenced by a cue, (2) status rules that take into account the interaction of multiple cues to determine concept status, and (3) assertion status assignment to a variety of concept types. The system uses MITRE's Carafe conditional random field implementation to identify negation and uncertainty cues as well as their scopes. A regular expression-based document zoner identifies document section headings and section boundaries, and categorizes the sections. A rule-based module uses information regarding cues and scope received from the classifier to derive the assertion status of clinical concepts.

We made extensions to our system to address three assertion categories that it did not already address (*conditional*, *hypothetical*, and *not associated with the patient*). We also added a maximum entropy classifier to make the final assertion classification.

2. Methods

2.1 Section Identification

The MITRE document zoner consists of a set of manually generated regular expressions designed to match the headings of clinical reports. It marks section boundaries and assigns a section type to each section that it identifies. The document zoner was run on i2b2/VA training data and updated to improve its accuracy in identifying previously unseen headings.

2.2 Cue Identification and Scope Determination

2.2.1 Negation and Speculation Cues

The cue and cue scope taggers are Conditional Random Field (CRF) classifiers. Both classifiers were trained on radiology reports in the BioScope[1] corpus, a publicly available corpus of biomedical texts annotated for negation and uncertainty cues and cue scope.

The scope identifier uses the current word, the words between the current word and the corresponding cue phrase and the relative position (direction and distance) from the current word to the cue phrase.

2.2.2 Cues for *Conditional, Hypothetical,* and *Not Associated with Patient*

When the challenge was announced, we had not yet developed cue and scope identifiers to recognize contexts indicative of *conditional*, *hypothetical*, or *not associated with patient* assertion status. Given the limited time, we selected terms that appeared to function as cues, either individually or together with other terms, for each of these assertion classes. We then created features that represented the occurrence of these cues in the text, and these features were included as input to the assertion classifier.

2.3 Status Rules

The MITRE concept status module uses a set of rules implemented in Java. Status rules derive the status of concepts using information generated by the cue and scope modules. For the i2b2/VA challenge assertion task, the status module was not the final determiner of concept status. Instead, information generated by status rules was converted to features for the assertion classifier.

2.4 Final Assertion Status Module

We used a Maximum Entropy classifier to assign the final assertion category[2]. Maximum Entropy classifiers benefit from the simplicity of a single hyper-parameter, a zero-mean Gaussian prior over the parameter values. This serves as a regularizer that can prevent overfitting (lower variance values have a stronger regularization effect).

2.4.1 Features

The final feature set used by the system included features that represented words and word location, word classes, cues and their scope, and linguistic structure.

Word Features: Word features included concept unigrams; and for each other word in a sentence, features were generated that indicated the word and whether it occurred to the right or left of the concept of interest.

Semantic class features: Features were generated for words belonging to specific semantic classes as indicated by our lexicons, such as activity (e.g., *walking*), conditional (e.g., *with exertion*), hypothetical (e.g., *monitor for*), temporal (e.g., *when*), not patient (e.g., *mother*).

Additional features were assigned to negation and uncertainty cues recognized by the negation/uncertainty detection module.

A feature was generated for each concept annotation provided by the challenge (treatments, tests, and problems).

Syntactic class features: Features were generated for words or phrases with specific syntactic functions as indicated by our lexicons such as clause boundary (e.g., *although*).

Cue scope: A feature was generated that indicated the number of negation and speculation cue scopes that enclosed the concept in question.

Document zone: A feature was generated that indicated the document section the concept in question occurred in.

Word class order: A feature was generated that identified the cue word class and order for each word to the left of the concept of interest. (No corresponding feature was generated for words to the right of the concept, but adding such a feature is under consideration.)

3. Submissions

We submitted three sets of classifier output for the i2b2/VA evaluation. The same features were used in all three runs. The submissions were distinguished only by the values of the hyper-parameter, which were 1.0, 10.0 and 100.0 for submissions 1, 2 and 3, respectively.

4. Results

Submission 1 achieved the highest F-measure (0.9343) for overall assertion classification, but scores for the three submissions were extremely close (Table 1).

Submission	F-Score
1 (prior = 1.0)	0.9343
2 (prior = 10.0)	0.9336
3 (prior = 100.0)	0.9308

Table 1. Overall assertion classification F-scores of three submissions.

Our system performed best on the *present* (0.96) and *absent* (0.94) categories, and achieved F-scores above 0.85 for *hypothetical* and *not associated with patient.* Poorest classification accuracy was obtained for *conditional* and *possible* (Table 2).

Assertion Category	Recall	Precision	F-Score
Present	0.9798	0.9370	0.9579
Absent	0.9202	0.9549	0.9372
Possible	0.5323	0.7718	0.6300
Hypothetical	0.8591	0.9235	0.8902
Conditional	0.2865	0.8033	0.4224
Not assoc. with patient	0.7793	0.9826	0.8692
Overall	**0.9343**	**0.9343**	**0.9343**

Table 2. Classification F-scores of best performing submission for individual assertion categories.

Although the two best performing categories were also the categories with the largest number of instances, assertion category frequency was not correlated with accuracy overall. The Pearson product moment correlation coefficient for assertion category frequency and F-score is 0.4747, which is not significant even at $p = 0.10$. Table 3 shows that our system performed better on *not associated with patient*, the smallest category, than on *conditional* or *possible*, both more frequent assertion categories. It also performed better on *hypothetical* than *possible*.

Assertion Category	Count	F-Score
Present	18,550	0.9579
Absent	3,609	0.9372
Possible	883	0.6300
Hypothetical	717	0.8902
Conditional	171	0.4224
Not assoc. with patient	145	0.8692
Overall	24,075	0.9343

Table 3. Assertion classification frequency and accuracy.

4.1 Errors

Of the total number of errors made by our system, 70.4% were false negatives for which the system annotation was *present* (the default category), and 21.6% were false positives for which the ground truth

assertion was *present*. Only 8.0% of the errors involved confusion between non-default assertion categories. (Table 4).

GT Sys	Pres	Abs	Pos	Hyp	Con	NAP
Pres	**12,762**	264	383	75	112	24
Abs	121	**3,321**	19	5	4	8
Pos	101	19	**470**	19	0	0
Hyp	31	3	11	**616**	6	0
Con	10	2	0	0	**49**	0
A/se	0	0	0	2	0	**113**

Table 4. Assertion category confusions. GT = ground truth assertion. Sys = system assertion.

4.1.1 *Conditional*

Our system performed most poorly on the *conditional* assertion category, with false negatives accounting for most of the errors. Analysis of these errors reveals multiple contributing factors. (1) Some of our semantic class lexica, such as the activity class lexicon, were incomplete, and consequently, relevant cues (e.g., *with palpation* in the sentence *Tone is normal , moving all limbs symmetrically , **irritable** with palpation of the scalp* .) were not identified and converted to features. Failure to recognize lexical indicators accounted for 63% of the false negatives. (2) In 24% of cases a relevant cue such as a medication was identified (TREATMENT concept) and represented with a feature, but the feature did not have sufficient weight to result in the proper classification. (3) Noisy features also appear to be a problem. In particular, the terms *allergy*, *allergies*, and *allergic* appear to have been annotated inconsistently in both the training and test data, sometimes receiving a *present* and sometimes receiving a *conditional* annotation. Example of *allergic* annotated as *present*:

> *He is **allergic** to Penicillin , Inderal , and also to Procan .*

Example of *allergic* annotated as *conditional*:

> *The patient is **allergic** to sulfa .*

There were twelve false positives. Several confusions with *present* involved context that appears to suggest *conditional* assertion status, and the basis on which a *present* classification was made is not obvious to us. An example follows:

> *… who presented to an outside hospital with a history of **left-sided chest pain** at rest.*

4.1.2 Possible

Possible was the second most challenging assertion classification for our system. Most of the errors were confusions with the *present* category, despite the fact that we had a cue and cue scope module designed to identify the relevant cues. Many false positives resulted from incorrect cue scoping, or interpretation of ambiguous cue scoping. For example, in the text *This showed lymphangitic spread of cancer in the chest , question of pulmonary nodules in the chest , pericardial effusion , multiple liver metastases , ...*, our system labeled *multiple liver metastases* as *possible*, rather than the intended *present*, because it assigned too large a scope for *question of*. Confusions with *present* accounted for 93% of the false negatives. We might be able to decrease this number significantly by implementing the precedence rules, as in ambiguous cases *possible* overrides *present*, and we did not enforce this.

4.1.3 *Not Associated with Patient*

For *not associated with patient* there were only two false positives. Both occurred in sentences listing immunization criteria, and in which a family cue occurred in close proximity to a problem but belonged to a separate criterion. The correct assertion was *hypothetical*.

> *Daycare during RSV season, a smoker in the household, neuromuscular disease, airway abnormalities or school age sibling, or 3 with chronic lung disease.*

Of the 32 false negatives, eight were cases that the system classified as *absent*. In fact, the problems were negated as well as being about someone other than the patient in all of these cases. Our system did not implement code to enforce the assertion category precedence (according to which *not associated with patient* takes priority over other assertion categories), and the classifier did not learn it.

> *Family History: no kids, no bleeds or strokes in other family members*

The remaining 24 cases were confusions with *present*, and all of these cases occurred outside the FAMILY HISTORY section. In most of these cases, a family cue was recognized and represented as a feature, but the feature set associated with the problem did not result in the desired classification.

> *Her brother had developed the typical rash on 9/3/9 .*

We believe our system performed as well as it did on this category due to its strong association with the section FAMILY HISTORY. Of the 113 concepts correctly classified as *not associated with patient* by our system, 104 had a FAMILY HISTORY section feature. Conversely, of the 127 concepts with the

FAMILY HISTORY section feature, 111 were true positives.

4.1.4 Hypothetical

We did not have a cue and scope module for *hypothetical* assertions and relied on features representing indicative terms and section types. We believe the relatively high classification accuracy (0.89) for this category is due to its strong association with specific section types. Of the 616 true positives in this category, 224 had an INSTRUCTIONS section feature, and 205 had a MEDCIATION section feature.

Of the 101 false negatives, 63 were associated with the system's failure to detect several lexical indicators that were not included in any lexicon; in twenty cases a hypothetical cue was detected but was not weighted sufficiently to result in hypothetical classification; in eight cases a relevant indicator was missed because it occurred in a preceding line of text. Finally, there were ten instances in which the gold annotation appeared to be inconsistent with annotation guidelines.

The 51 false positives were associated with a wide variety of factors, of which the two most frequent were (1) hypothesized change in status of a problem (20%) and location of hypothesized problem (24%).

> *If **problems with speech** or **weakness** worsen , go immediately to an emergency room for evaluation .*
>
> *Call for any fever , redness or drainage from **wounds** .*

4.1.5 Absent

Our system performed well classifying *absent* assertions, with accuracy second only to accuracy for the default assertion category, *present*. We were able to apply our negation cue and cue scope modules, and this category had more training examples than any category except the default category.

There were 121 false positives for which the ground truth classification was *present*. For roughly half of these, the scope that our system established for a negator incorrectly extended beyond a clause or phrase boundary. In other cases, proximity of negation cues led to a classification of *absent* despite the fact that scope determined by our negation scope module did not include the problem concept in question.

Finally, we did observe cases for which the ground truth annotation appeared to contradict the assertion annotation guidelines. For example, the guidelines provided *his dyspnea resolved* as an example of an *absent* assertion for *his dyspnea*, but the evaluation data classified *loose bowel movements* as *present* in the following sentence:

> ***Loose bowel movements** - This problem was resolved and had been a viral syndrome on presentation .*

The system generated 264 false negatives for this category. For 156 of these instances, a feature representing negation was generated, but was not weighted enough to result in the *absent* assertion classification. In the remaining 108 cases, no negation cue was identified.

4.2 Contribution of Features

Our contribution to the i2b2 assertion task involves a number of different feature classes. Table 5, below, shows the effect of different features sets on accuracy, using the same training/test split as in the formal evaluation. Each successive run added one more feature class to the classes already used in the previous runs. The significance of the differences in F-measure was determined using a paired randomization test[3].

	F-Score	**Significance Level (p)**
Context uni-grams	0.9097	
+ Concept Uni-grams	0.9252	0.00001
+ Document Zones	0.9287	0.0003
+ Cue scope	0.9298	0.1865
+ Syntactic/Semantic	0.9342	0.00015

Table 5. F-measure results on the evaluation data as different feature sets are added to the classifier.

Our baseline system (context unigrams) achieved an F-measure of 0.9097. By adding features representing the concepts themselves, document zone (section) cue scope, and linguistic features of words, we were able to increase the F-score by 0.245. This accuracy increase represents a reduction of 27% of the baseline error. The impact of any specific feature set varied across assertion categories. The document zone feature was important for *hypothetical, conditional,* and *not associated with patient.* The concept itself was an important feature for *absent (*e.g., *nontender)* and *conditional (*e.g., *side effect).*

5. Discussion and Conclusions

The MITRE system that was developed to classify medical problem assertions combined machine learning algorithms with linguistic knowledge

represented in lexicons, regular expression-based patterns, and scope enclosure rules. To combine these, we fed output from our statistical scope module to a rule-based status module, and fed the output of that module, as well as features derived from other kinds of linguistic knowledge, to a final statistical system classifier. Our approach is an extension of the approach described in Clark et al.[4], where linguistic information such as negation status and temporal attributes was used as features for statistical classification of patient smoking status. We believe this is a good way to leverage rule-based and statistical techniques. When rule-derived information is converted to features and used as input to a machine learner, it is automatically weighted with respect to its contribution of true and false positives.

The MITRE assertion classification system achieved a baseline F-measure of 0.91 using a maximum entropy algorithm with context unigrams as features. System accuracy was improved (F-measure of 0.93) by adding features that represent linguistic attributes of the text, such as document structure, sentential structure, and semantic attributes of words in the sentence. Upon correction of an error in our lexical look-up and retraining after the challenge, our system was able to obtain an F-score of 0.94.

In analyzing our results, we noted that the cue scope module did not contribute as much to our system as we expected. We hypothesize that this is due to the fact that it was trained on data somewhat different from the challenge data. We plan to test this hypothesis by annotating i2b2 challenge data for cue and cue scope, and training our cue scope module on this data.

We expect that extending the coverage of our lexicons will make the associated features more reliable and further reduce the errors made by the current system. We plan to make an open source release of our system in the future.

Acknowledgements

The work presented in this paper was funded through MITRE's internal research and development program, the MITRE Innovation Program.

References

1. Vincze, V, Szarvas, G, Farkas, R, Móra, G and and Csirik, J. The BioScope corpus: annotation for negation, uncertainty and their scope in biomedical texts. BMC Bioinformatics 2008, 9(Suppl 11):S9
2. Wellner B, Huyck M, Mardis S, Aberdeen J, Morgan A, Peshkin L, Yeh A, Hitzeman J, Hirschman L. Rapidly retargetable approaches to de-identification in medical records. J Am Med Inform Assoc. 2007;14(5):564-73. http://sourceforge.net/projects/carafe
3. Yeh, A. More accurate tests for the statistical significance of result differences. 18th International Conference on Computational Linguistics (COLING 2000), 2000; 947-953
4. Clark, C, Good, K, Jezierny, L, Macpherson, M, Wilson, B, Chajewska, U. Identifying smokers with a medical extraction system, J Am Med Inform Assoc. 2008;15:36 –39