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

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 obtains an overall micro-averaged F-score of 0.9343. The methods employed are a combination of machine learning (Conditional Random Field and Maximum Entropy) and rule-based (pattern matching) techniques.

1. Introduction

The 2010 i2b2/VA challenge was a three-tiered challenge designed to study (1) extraction of medical problems, tests, and treatments, (2) classification of assertions made on medical problems and (3) relations of medical problems, tests, and treatments. Challenge participants could choose to participate in any or all of the tiers. MITRE participated in the second tier, classification of assertions made on medical problems.

The goal of the assertion task was to determine for each occurrence of a problem concept (e.g., myocardial infarction, global hypoxic injury to the brain) what kind of assertion was made by the clinical report. Assertion categories included the following:

- Present (default category)
- Absent (problem does not exist in patient)
- Possible (patient may have problem but uncertainty expressed)
- Conditional (patient experiences problem only under certain conditions)
- Hypothetical (medical problem that the patient may develop)
- Not associated with the patient (problem associated with someone who is not the patient)

Furthermore, to resolve ambiguity there was a precedence order of categories, which was as follows (from highest to lowest priority):

- *Associated with someone else* overrode any other assertion.
- *Hypothetical* overrode *present, absent, possible*, or *conditional*
- *Conditional* overrode *present, absent*, or *possible*.
- *Possible* overrode *present* and *absent*
- *Absent* overrode *present*
- *Present* was to be used only used if there was no evidence of the other assertion types.

2. Methods

MITRE has been developing algorithms to improve the accuracy of detection of negated and uncertain or speculative information in clinical reports. We have built a medical information extraction system based on Mayo Clinic's cTAKES[1] system for identifying clinical concepts. Our system also 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 is used to identify document section headings and section boundaries, and categorize the sections. A rule-based module uses information regarding cues and scope received from the classifier to derive the assertion status of clinical concepts identified by cTAKES. 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.

Since ground truth for concept identification was provided by the challenge for the assertion task, we did not include the cTAKES extraction engine in the system we assembled, but we applied the other components of our system to the assertion task. These included a document zoner, a negation and uncertainty cue identification module, a cue scope determination module, and a rule-based assertion status module. We made extensions to our system to address three assertion categories that our system 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. This component took as input the found cues, outputs of the rule-based assertion status module, and additional concept and contextual features (discussed in detail in section 2.4).

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[2] corpus, which is a publicly available corpus of biomedical texts annotated for negation, uncertainty and the scope of negation and uncertainty cues in biomedical texts. The BioScope corpus contains 1,954 reports that were used for the clinical coding challenge organized by the Computational Medicine Center in Cincinnati, Ohio in 2007.

To identify cues, the cue tagger uses the words 3 to the left and 3 to the right ignoring case, together with a feature indicating whether the word appears in a list of known uncertainty and/or negation terms. In initial 10-fold cross-validation using BioScope biomedical abstracts and radiology reports, the cue tagger achieved an F-score of 0.93 for negation cues and 0.86 for speculation cues.

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. In initial 10-fold cross-validation using BioScope biomedical abstracts and radiology reports, the scope identifier achieved a token-based F-score of 0.82 for negation cue scopes on biomedical journal abstracts and 0.97 on radiology reports. For speculation cue scopes, it achieved accuracies of 0.88 on journal abstracts and 0.83 on radiology reports.

2.2.2 Cues for *Conditional, Hypothetical*, and *Associated with Someone Else*

We have not yet developed cue and cue scope identifiers that we could apply to recognize contexts indicative of *conditional, hypothetical*, or *associated with someone else* assertion status. Since the challenge data did not have cue scope annotated, and we did not have sufficient time to add such annotations to the data, we use a different approach to recognize these assertion categories. Based on our analysis of the data, we selected terms that appeared to function as cues, either individually or together with other terms, for each of these assertion classes. In other words, the meaning of these terms seemed to be the basis for a particular assertion interpretation. We also noted associations between assertion category and report section. Rather than writing rules to derive assertion status directly from the presence of these cues, we 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 use information generated by the cue and scope modules, to derive the status of concepts. 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 to be used by the assertion classifier.

2.4 Final Assertion Status Module

We used a Maximum Entropy classifier to assign the final assertion category[3]. 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). Our three separate submissions varied only in the setting of this hyper-parameter, using the values 1.0, 10.0 and 100.0 for submissions 1, 2, and 3, respectively.

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, a feature was generated that indicated the word and whether it occurred to the right or left of the concept whose assertion was to be classified. An additional feature was assigned to words that were within three tokens of the concept.

Semantic class features: Features were generated for words belonging to specific semantic classes as indicated by our lexicons:

activity (e.g., *walking*)
cause (e.g., *cause*)
conditional (e.g., *with exertion*)
hypothetical (e.g., *monitor for*)
negation (e.g., *without*)
speculation cue (e.g., *unlikely*)
temporal cue (e.g., *when*)
not patient (e.g., *mother*)

inherently negated (e.g., *afebrile*)
inherently conditional (e.g., *exertional angina*)

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).

Potentially misleading phrases:

A separate set of features was generated for phrases that contained terms that would otherwise be treated as negation or speculation cues but which were not relevant to assertion of problems in this context:

NegEx[4] pseudo negation (e.g., *not certain if*)
MITRE pseudo negation (e.g., *almost no*)
pseudo speculation (e.g., e*tiology likely from*)

Syntactic class features: Features were generated for words or phrases with specific syntactic functions as indicated by our lexicons:
clause boundary (e.g., although)
phrase boundary (e.g., despite)
NegEx scope terminator (e.g., as a source for)

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

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 whose assertion was to be classified. (No corresponding feature was generated for words to the right of the concept, but the addition of 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 accuracy (0.9343) for overall assertion classification, but accuracy 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 accuracy 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 *associated with someone else.* 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
Assoc. with Someone Else	0.7793	0.9826	0.8692
Overall	**0.9343**	**0.9343**	**0.9343**

Table 2. Classification accuracy 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 = .10. It can be seen from Table 3 that our system performed better on *associated with someone else*, the smallest category, than on *conditional* or *possible*, both more frequent assertion categories.

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
Assoc. with Someone Else	145	0.8692
Overall	24,075	0.9343
Pearson r = 0.4747, p > 0.10 df = N-2 = 4		

Table 3. Assertion classification frequency and accuracy.

Our system also achieved higher accuracy for the *hypothetical* category than the more frequent *possible* category.

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	A/se
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. (2) In a large number of cases (~37) 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. Examples of allerg* annotated as *present*:

> He is **allergic** to Penicillin , Inderal , and also to Procan .
>
> no known drug allergies , but has **shellfish allergy** , and question of an Iodine allergy

Examples of allerg* annotated as *conditional*:

> The patient is **allergic** to sulfa .
>
> She has **allergies** to Morphine , Percocet , Codeine , Penicillin , Xanax and Toradol .

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:

> *The patient is a 63 - year-old man with a history of hypertension and prior history of atrial fibrillation secondary to hyperthyroidism 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. Preliminary analysis shows that some of these errors result 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.*

4.1.3 *Hypothetical*

Our system achieved its third highest accuracy score for the *hypothetical* assertions. We did not have a cue and scope module for this category, and relied on features representing indicative terms and section types. We believe the relatively high classification accuracy 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.

4.1.4 *Associated with Someone Else*

For *associated with someone else* 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 *associated with someone else* 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 *associated with someone else* by our system, 104 had a FAMILY HISTORY section feature.

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 also had more training examples than any category except for 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 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, provides a set of different runs using the same training/test split as was used for the formal evaluation. Each successive run added one more feature class to the classes already used in the previous runs.

	F-measure
Context uni-grams	**0.9097**
+ Concept Uni-grams	**0.9252**
+ Document Zones	**0.9287**
+ Cue scope	**0.9298**
+ Syntactic/Semantic/Word	**0.9342**

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.

Discussion and Conclusions

The MITRE system that was developed to classify medical problem into assertion categories combined machine learning algorithms with linguistic knowledge represented in lexicons, with regular expression-based patterns, and with scope status rules. Rather than develop distinct rule-based and statistical systems as has been done in other NLP systems, 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.[5], 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.

Our 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 in the future by annotating i2b2 challenge data for cue and cue scope, and training our cue scope module on this data. If we succeed in obtaining error reductions, we may extend this approach to other cues such as those associated with hypothetical and other classes of temporal assertions.

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.

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. We expect to achieve additional improvements in accuracy by improving the accuracy of the individual modules that generate the information upon which these linguistic features are based.

Acknowledgements

The 2010 i2b2/VA challenge and the workshop are funded in part by the grant number U54-LM008748 on Informatics for Integrating Biology to the Bedside from National Library of Medicine. This challenge and workshop are also supported by resources and facilities of the VA Salt Lake City Health Care System with funding support from the Consortium for Healthcare Informatics Research (CHIR), VA HSR HIR 08-374 and the VA Informatics and Computing Infrastructure (VINCI), VA HSR HIR 08-204. MedQuist, the largest transcription technology and services vendor, generously co-sponsored the 2010 i2b2/VA challenge meeting at AMIA.

References

1. Savova GK, Kipper-Schuler KC, Buntrock JM and Chute CG. UIMA-based clinical information extraction system. LREC 2008: Towards enhanced interoperability for large HLT systems: UIMA for NLP.
2. 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
3. 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
4. Chapman WW, Bridewell W, Hanbury P, Cooper G, Buchanan B. Evaluation of negation phrases in narrative clinical reports. Proc AMIA, 2001;105-109.
5. 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