Journal of Biomedical Informatics xxx (2012) xxx–xxx

Contents lists available at SciVerse ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Assertion modeling and its role in clinical phenotype identification

Cosmin Adrian Bejan [a,*], Lucy Vanderwende [b,a], Fei Xia [c,a], Meliha Yetisgen-Yildiz [a,c]

[a] Biomedical and Health Informatics, University of Washington, Seattle, WA 98195, United States
[b] Microsoft Research, Microsoft, Redmond, WA 98052, United States
[c] Department of Linguistics, University of Washington, Seattle, WA 98195, United States

ARTICLE INFO

Article history:
Received 14 June 2012
Accepted 5 September 2012
Available online xxxx

Keywords:
Natural language processing
Clinical information extraction
Assertion classification
Pneumonia identification
Statistical feature selection

ABSTRACT

This paper describes an approach to assertion classification and an empirical study on the impact this task has on phenotype identification, a real world application in the clinical domain. The task of assertion classification is to assign to each medical concept mentioned in a clinical report (e.g., *pneumonia*, *chest pain*) a specific assertion category (e.g., *present*, *absent*, and *possible*). To improve the classification of medical assertions, we propose several new features that capture the semantic properties of special cue words highly indicative of a specific assertion category. The results obtained outperform the current state-of-the-art results for this task. Furthermore, we confirm the intuition that assertion classification contributes in significantly improving the results of phenotype identification from free-text clinical records.

1. Introduction

There is a great amount of information captured in physicians' comments made during health care. Increasingly, researchers are finding valuable uses by mining and aggregating this data in clinical and translational studies which lead to improved patient care, but progress is slow since study cohorts first need to be constructed, and currently, this can only be done through manual annotation of the patient records. As a result, there is a great deal of interest in identifying critical illness phenotypes, and other medical concepts, automatically. The task cannot be handled by simple key word matching, since the phenotype may be mentioned even when the physician is ruling it out or specifying the conditions under which the phenotype is observed.

An important contribution to address this issue is the Informatics for Integrating Biology and the Bedside (i2b2)/Veteran's Affairs (VA) challenge. Each year, this challenge focuses on a specific problem to enable clinical information extraction. The 2010 i2b2/VA challenge introduced *assertion classification* as the shared task, formulated such that each medical concept mentioned in a clinical report (e.g., *asthma*) is associated with a specific assertion category [1]; 21 teams competed in this task, confirming that assertion classification is perceived to be an important task. The assertion categories that were included in the task are: *present*, *absent*, *conditional*, *hypothetical*, *possible*, and *not associated with the patient* (*not patient*, for short). As observed, because it requires the identification of both negated and uncertain medical concepts, the task of classifying medical assertions is closely related to both tasks of *negation detection* and *hedge* (or *speculation*) *identification* that were recently studied in [2,3].

While the systems report good results, error analysis suggests that the performance can still be improved, particularly for the categories that express hedging (i.e., *conditional*, *hypothetical*, and *possible*). In this paper, we introduce several novel features that explore the syntactic information encoded in dependency trees in relation to special cue words for these categories. We find that each of the feature classes added provides significant improvement.

With our new assertion classifier, we return to the task at hand, to automatically identify critical illness phenotypes, specifically pneumonia; while pneumonia identification is of great value to clinical researchers, there are many other phenotypes (e.g., *sepsis*) that can be identified using the same methodology. In this paper, we describe how we map from the six assertion categories to distinguishing positive from negative cases of pneumonia. We also experiment with adding assertion classification to various feature representations for this task, and find that assertion classification provides significant improvement in all experiments, underscoring the importance of semantic phenomena on the interpretation of clinical text.

2. Related work

The task of identifying an assertion category for a given medical concept has been extensively studied in the past several years due

* Corresponding author. Address: Department of Biomedical Informatics and Medical Education, School of Medicine, University of Washington, 1959 NE Pacific Street, HSB I-264, Box 357240, Seattle, WA 98195-7240, United States. Fax: +1 206 221 2671.

E-mail addresses: bejan@u.washington.edu (C.A. Bejan), Lucy.Vanderwende@microsoft.com (L. Vanderwende), fxia@uw.edu (F. Xia), melihay@u.washington.edu (M. Yetisgen-Yildiz).

1532-0464/$ - see front matter © 2012 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.jbi.2012.09.001

to its importance and direct application in the clinical domain. One of the earlier tools developed in this direction is NegExpander [4]. This tool employs rules based on part-of-speech tags with the purpose of distinguishing positive from negative expressions in clinical text. Later on, Mutalik et al. [5] proposed Negfinder – a tool relying on methodologies for parsing formal computer languages which encodes various rules used for extracting a large set of patterns associated with negated medical concepts.

A more simple but powerful regular expression-based tool for deciding on a negated concept is NegEx [6]. Since at the core of the NegEx algorithm lies a list of rules matching negative expressions which can be easily modified, this tool drew favorable attention to be enhanced with additional functionalities. In this regard, Golding and Chapman [7] incorporated machine learning techniques into the NegEx algorithm, Chapman et al. [8] proposed ConText, a generalized version of NegEx, and Uzuner et al. [9] implemented the Extended NegEx (ENegEx) tool in order to handle alter-association assertions. Uzuner et al. [9] also compared ENegEx with a machine learning approach for assertion classification.

More recently, Elkin et al. [10] extended the Negfinder and NegEx tools by employing a negation ontology. In their study, Elkin et al. also performed an error analysis in order to determine the most common cases where negation is not properly identified. Also, Huang and Lowe [11] proposed an approach using regular expressions matching structured information that is encoded into syntactic parse trees. This way, they were able to identify not only the negated medical concepts situated into a close proximity from a negation signal (such as *no* or *not*), but also the ones located distantly from these signals.

The performances of the tools described above are, however, difficult to compare because most of these tools are not publicly available. Also, the access to the datasets on which these tools were evaluated is restricted due to the policies imposed by clinical data warehouses. On the other hand, the initiative of the i2b2/VA organizers to provide a standard, publicly available dataset of deidentified medical records gave to the researchers from the biomedical informatics field the opportunity to create a new generation of assertion classification tools which can be ranked on a common evaluation platform. The most successful approaches participating in the 2010 i2b2/VA challenge for the assertion classification task used support vector machine (SVM) classifiers and performed various feature selection strategies over a wide range of engineered features [1]. The best ranked system was designed by de Bruijn et al. [12] who proposed an architecture that solves the assertion classification task in two phases. First, predictions were assigned to every word from medical concepts; next, these word predictions were used by a second classifier to generate the final concept predictions. With this approach, de Bruijn et al. [12] obtained 93.62 micro-averaged *F*-measure (the primary metric considered for this task).

After the 2010 i2b2/VA challenge, two teams that participated in this competition improved their system's performance even further. Roberts and Harabagiu [13] enhanced their system's performance to 93.94 micro-averaged *F*-measure by devising various optimization methods in order to find a better set of features for this task. Kim et al. [14] implemented additional features by focusing on improving the performance of minority classes. Using this new set of features, their system reached 79.76 and 94.17 macro- and micro-averaged *F*-measures, respectively. Both teams used a single SVM classifier in their learning framework.

3. Classifying medical assertions

To enable the study of assertion classification, the i2b2/VA organizers provided a set of 826 documents annotated by medical experts, which was split into 349 and 477 training and test documents respectively, and 827 non-annotated documents. Overall, the training and test datasets contain 11,968 and 18,550 annotated concepts, respectively. The documents consist of deidentified discharge summaries and progress notes, collected from three different institutions. The set of annotated documents have the following distribution of assertion categories: approximately 69% of concepts are *present*, ~20% *absent*, ~4.5% *hypothetical* and *possible*, and <1% *conditional* and *not patient*. For the assertion classification task, the medical concepts were specified, and the documents were already segmented into sentences and tokenized.

Example 1. Example of an artificially created clinical report emphasizing assertion annotations of medical concepts.

s_1	*History of present illness:*
s_2	*For the past several years, the patient has continued to have significant exercise intolerance which has limited his ability to work.*
s_3	*He does become **slightly short of breath** when lifting weights.* (*conditional*)
...	
s_4	*Family history:*
s_5	*One of his brothers has had **a myocardial infarction.*** (*not patient*)
...	
s_6	*Physical examination:*
s_7	*On physical examination, the patient has no **fever.*** (*absent*)
...	
s_7	*Lungs revealed **occasional wheezing bilaterally** at both bases.* (*present*)
s_8	*Hospital course:*
...	
s_9	*He was also started on antibiotics for possible **bronchitis** or **community-acquired pneumonia.*** (*possible*)
s_{10}	*Discharge Medications:*
...	
s_{11}	*Lasix 40 mg Tablet Sig: One (1) Tablet PO once a day as needed for **shortness of breath** or **wheezing.*** (*hypothetical*)

In Example 1, we show an artificially created clinical report with sentences extracted from the i2b2/VA dataset. These sentences represent examples for each assertion category, in which the medical problems are emphasized in boldface. For instance, the medical problem from the sentence s_3 is annotated as *conditional* because it is experienced by the patient only under specific conditions (in this case, only when the patient performs intense physical activities).[1] As is also observed in Example 1, a clinical report can be structured into sections. Here s_1, s_4, s_6, s_8, and s_{10} represent examples of section headers which mark the beginning of a new section in the report. These types of section headers are usually seen in discharge summaries.

3.1. Features

As noted by many participating teams in the assertion classification task, an important factor to obtaining top results for this task lies in performing significant feature engineering. In concordance with this observation, we implemented a wide diversity of features that explore the contextual information of a medical concept at sentence and report level. To extract these features, we first pre-processed the documents using SPLAT [15], a state-of-the-art NLP platform which include lemmatization, Porter stemming,

[1] Additional examples for each assertion category are provided in the assertion annotation guidelines at https://www.i2b2.org/NLP/Relations/

Please cite this article in press as: Bejan CA et al. Assertion modeling and its role in clinical phenotype identification. J Biomed Inform (2012), http://dx.doi.org/10.1016/j.jbi.2012.09.001

Table 1
Features for assertion classification.

f_1	Word, lemma, and stem uni/bi/tri-grams occurring before and after the medical concept
f_2	The sparse stem trigram to the right of the concept and with the wildcard placed immediately to the right of the concept
f_3	Binary features indicating the presence of special tokens (question mark, and comma) adjacent to the medical concept
f_4	The combination of the part of speech and stem associated with every token (except punctuations) from the sentence
f_5	Concatenation of all concept stems
f_6	The concept stems
f_7	The closest preposition to the left of the medical concept
f_8	Binary features indicating whether the tokens representing the concept and the ones occurring in a context window of size five around the concept start with a negative prefix
f_9	The output of the ConText algorithm [8] in a token context window of size six surrounding the concept
f_{10}	The section header
f_{11}	Binary feature that is *true* if the section header contains any of the *allergic*, *allergies*, or *allergy* words
f_{12}	Binary feature that is *true* if the section header contains the *family* and *history* words
f_{13}	Binary features indicating whether specific token *n*-grams for the *absent*, *conditional*, and *possible* classes occur around the concept
f_{14}	The closest negative cue in the left token context window delimited by the closest comma and the first token of the concept
f_{15}	The first assertion cue on the path in the dependency tree between the concept and root
f_{16}	The first verb on the path in the dependency tree between the medical concept and root
f_{17}	The modal auxiliary verb associated with the first verb on the path in the dependency tree between the medical concept and root
f_{18}	The distance on the dependency tree between the concept and the closest assertion cue
f_{19}	The sequence of part of speech labels between the closest assertion cue situated to the left and the concept

and constituent and dependency parsing. Then, based on an initial set of features, we trained an SVM multi-class classifier with default parameter settings [16] and employed an optimization method to decide which features from the initial set provide the best performing results. This optimization method is based on a greedy algorithm that iteratively selects the most salient features from the initial set. For determining the final set of features, we ran this feature selection method over the training set using a 10-fold cross validation evaluation scheme. The features from this final set are listed in Table 1.

In the remaining part of this section, we describe in detail the features from Table 1. For a better representation, we grouped the features based on their common characteristics and on the impact they have on improving the classifier's performance.

3.1.1. Basic features (f_{1-9})

This set of features encodes the surrounding contextual information of the medical concept at the sentence level. For instance, the feature f_2 associated with the concept from s_3 in Example 1 is "$*$ *lift weight*", which represents expressions occurring right after a concept such as *when*, *after*, or *while lifting weights*. The set of binary features f_3 tries to determine if the concept is part of a list, or if it is adjacent to a question mark, which, in medical language, often indicates uncertainty. The list of negative prefixes used for extracting the binary features f_8 is: *ab*, *de*, *di*, *il*, *im*, *in*, *ir*, *re*, *un*, *no*, *mel*, *mal*, and *mis*. This list is identical with the list used by Kim et al. [14]. One important feature is f_9, which is based on the output generated by the ConText algorithm [8]. This algorithm uses regular expressions to identify three clinical properties of a clinical condition: *negation*, *temporality*, and *experiencer*.

3.1.2. Section features (f_{10-12})

Since clinical records are often structured into sections, extracting features that identify the section headers proved to be useful for specific assertion categories. For instance, most of the medical concepts present in allergies sections are annotated as *conditional*; similarly, many concepts in the family history sections are *not associated with the patient*.

In this dataset, identifying section headers in clinical notes is relatively straightforward. Most of the section headers were selected from the sentences that end with a colon; however, the sentences that represent a special set of expressions (e.g., *instructions*, *family history*, *medications*, *discharge instructions*, etc.) or the ones that start with these expressions followed by a colon are also marked as section headers.

3.1.3. Category specific features (f_{13})

Due to the fact that the distribution of the assertion categories is highly non-uniform, the multi-class classifier has the tendency to favor the features covering the majority categories and is not able to assign sufficient weight to some of the features designed for instances corresponding to the minority categories. To address this issue, we built a simple analysis tool that extracts all possible adjacent *n*-grams to the left and right of the medical concepts from the training set and counts how many times each *n*-gram *g* is associated with every assertion category *c*. Based on these frequency counts, denoted as $f(g,c)$, we computed a *score* that measures how relevant an *n*-gram is for a specific category.[2] For instance, the left uni-gram *no* has a score of .99 ($\sim$1000/1009) for *absent* because it occurred 1000 times for this category, 3 times each for *present* and *not patient*, and 1 time each for *conditional*, *possible*, and *hypothetical*.

After computing the scores, we built a binary feature for each of the six categories. One such feature is true for a given concept if the concept has an adjacent *n*-gram *relevant* to the category associated with the feature. For an *n*-gram to be relevant to one of the six categories, it needs to be counted at least three times and the score for the category to be >.95.

In addition, we ran greedy optimization techniques for the most relevant features on the training set and used them as heuristics for predicting the instances associated with the minority categories. Examples of highly relevant features associated with the *conditional* category and used for this purpose are the bi-grams occurring to the left of the medical concept *on exertion* and *with exertion*.

3.1.4. Assertion focus features (f_{14-19})

This novel set of features tries to capture whether a medical concept is within the semantic influence of a special word expression that is able to signal a specific assertion category. The word expressions we identified as being able to signal an assertion category are the *negation cues* (e.g., *not*, *without*, and *absence of*) from the BioScope corpus [17], the *speculative* (or *hedge*) *cues* (e.g., *suggest*, *possible*, and *might*) from the same corpus, the *temporal sig-*

[2] The score of an *n*-gram *g* over a category *c* is given by $f(g,c)/\sum_c f(g,c)$. This is a simplification to the method that computes correlation scores based on statistical tests.

Please cite this article in press as: Bejan CA et al. Assertion modeling and its role in clinical phenotype identification. J Biomed Inform (2012), http://dx.doi.org/10.1016/j.jbi.2012.09.001

nals (e.g., *after*, *while*, *on*, and *at*) from TimeBank [18], and a list of *kinship terms* (e.g., *father*, *mother*, *sister*, and *aunt*) from the Longman English Dictionary. We refer to all these word expressions as *assertion cues*.

Based on this list of special words, the features f_{14-19} extract various forms of knowledge that encode the connection between an assertion cue and a medical concept in order to help the classification algorithm decide whether or not the concept is within the *focus* of the assertion cue. To our knowledge, this is the first system to make use of syntactic features in the assertion classification task. The idea of assertion focus is inspired from the semantic theory of negation. According to Quirk et al. [19], the focus of negation is the most contrastive part from the negation scope whereas the negation scope is the part of the language that encodes the meaning of a negative item. A recent study on identifying the negation focus is described in [20].

Although most of these features explore the syntactic information encoded in dependency trees, we also extracted path features from constituent trees. However, those features did not end up being selected in the final feature set.

4. Assertion classification results

Table 2 shows the results achieved by our system on both training and test sets. The results for each category are reported in terms of precision (*P*) and recall (*R*) whereas the results measuring the performance over all assertion categories (listed as 'overall' in the table) are reported in terms of macro- and micro-averaged *F*-measure (macro*F* and micro*F* in the table). The micro-averaged *F*-measure is computed by averaging over all the medical assertions and therefore it assigns an equal weight to each data instance. In consequence, this measure is dominated by the results associated with the majority categories. On the other hand, the macro-averaged *F*-measure gives an equal weight for each assertion category by computing locally an *F*-measure for each such category and then averaging these measures to obtain the final score. Since the micro-averaged *F*-measure is considered as the primary measure for assertion classification, the feature selection optimization is performed based on this measure.

The first four rows in Table 2 show the cross validation results over the training instances. To measure the impact of each feature set, we split the experiments by cumulatively adding each set to the basic set of features. Also, we employed a randomization test based on stratified shuffling [21] to determine if the differences in performance between these experiments are statistically significant.

Even when considering only the features from the basic set, the system achieves a satisfactory 94.48 micro*F*. When adding the section features, a significant increase in performance is shown for *not patient*. This is consistent with our intuition since many of the instances from 'family history' sections are associated with this category. The category specific features contribute mostly in improving the performance of *conditional* and *possible* because most of the relevant *n*-grams were selected for these two categories. And finally, the syntactic features that capture the focus of the assertion cues have a positive impact on both precision and recall for all assertion categories.

The final test results obtained by our system are 79.96 and 94.23 macro- and micro-averaged *F*-measures respectively, which slightly outperformed the current best published results of Kim et al. [14] (last two rows in Table 2). While our system show improvement in both precision and recall for *present* (representing the majority class with 69% of medical assertions) and *possible*, the system proposed by Kim et al. [14] better predicts the *absent* category (which represents the second majority class with 20% annotated instances).

4.1. Error analysis

The worst performing results by our system are for the *conditional* category. From the total of 171 test instances annotated for this category 63.74% (109) were classified as *present*. In fact, these misclassifications accounted for 91.6% of the *conditional* false negatives. Our analysis of the system predictions over the training set revealed that most of the errors for the *conditional* category correspond to instances where the condition is expressed in text after the medical problem and is introduced by a preposition or subordinating conjunction followed by a construction expressing a physical activity (*with climbing*, *after exercise*, *when taking*, *while doing exercise*, *when getting up*, etc.). We tried to capture these cases by first looking for expressions that match a prepositional phrase or subordinating clause and having a (PP IN NP) or (SBAR WHADVP S) syntactic structure. Then, we tested whether the head of the noun phrase NP or of the clause S is a hyponym of the synset representing a human action or activity in WordNet [22]. However, this feature was not able to make a clear separation between *conditional* and *present* instances due to the fact that many *present* instances matched the above criteria even after we tried to narrow down the list of activity hyponyms.

Inconsistencies in annotations for the *conditional* category was another factor that negatively influenced the classifier's performance. One such annotation inconsistency is shown by Example 1 and Example 2 from the training set, where the former was annotated as *conditional* and the later as *present*.

(1) *The patient is **allergic** to Ciprofloxacin, morphine sulfate and Droperidol.*
(2) *The patient is **allergic** to prednisone and penicillin.*

A similar example of inconsistency for *allergic* was also presented in [23].

Table 2
Assertion classification results. The first four rows show the incremental improvements of the feature sets over the training set. The differences in performance between two consecutive experiments on this set are statistically significant at $p < .001$ (*). The last two rows compare the test results of our system against the results of Kim et al. [14].

System configuration	*Absent*		*Not patient*		*Conditional*		*Hypothetical*		*Possible*		*Present*		Overall	
	P	*R*	*P*	*R*	*P*	*R*	*P*	*R*	*P*	*R*	*P*	*R*	macro*F*	micro*F*
Training set														
Basic	95.77	95.66	85.48	57.61	76.19	31.07	94.26	88.33	79.36	64.67	95.05	97.81	77.93	94.48
+Section	96.20	95.78	92.68	82.61	73.33	32.04	95.36	91.55	79.73	65.42	95.50	97.89	81.65	94.96*
+Category specific	96.50	95.78	92.94	85.87	80.39	39.81	95.51	91.55	84.87	72.34	95.97	98.16	84.55	95.55*
+Assertion focus	96.87	96.37	95.18	85.87	82.35	40.78	95.54	92.17	87.03	74.02	96.20	98.31	85.42	95.89*
Test set														
Kim et al.	**96.31**	**94.71**	**97.52**	81.38	**81.25**	**30.41**	92.07	**87.45**	78.30	54.36	94.46	98.07	79.76	94.17
Our results	95.71	93.88	91.79	**84.83**	80.00	**30.41**	**92.42**	86.75	**83.16**	**55.95**	**94.51**	**98.28**	**79.96**	**94.23**

Please cite this article in press as: Bejan CA et al. Assertion modeling and its role in clinical phenotype identification. J Biomed Inform (2012), http://dx.doi.org/10.1016/j.jbi.2012.09.001

Table 3
Results of the assertion classifier when mapping the assertion categories into *positive* and *negative* classes.

Mapping configuration		*Positive*			*Negative*		
		P	*R*	*F*	*P*	*R*	*F*
Before	$Hedge^-$	94.65	97.59	96.10	93.87	87.00	90.31
	$Hedge^+$	98.21	99.01	98.61	95.95	92.89	94.40
After	$Hedge^-$	94.51	98.28	96.36	95.53	86.55	90.82
	$Hedge^+$	98.40	98.93	98.67	95.70	93.66	94.67

Table 4
Top 5 most informative uni-grams, bi-grams and UMLS concepts according to *t*-statistic.

Uni-gram	Bi-gram	UMLS concept
Sputum	Sputum cx	Microbial sputum culture
Suctioning	Sputum culture	Sputum
h1n1	Continue lpv	Consolidation
Ventilatory	h1n1 Influenza	Infiltration
Consolidation	Acquired pneumonia	Influenza preparation

5. Positive and negative assertion categories

In practical clinical applications it is often required to identify a patient as being either *positive* or *negative* for a specific phenotype by analyzing the clinical information encoded into the patient reports. Therefore, in order to use our assertion classifier for such types of applications, we first need to determine what is the best way of mapping the six assertion categories into the *positive* and *negative* classes. From all $\sum_{k=1}^{5}\binom{6}{k} = 2^6 - 2 = 62$ ways of performing these mappings, we identified only two of them as more plausible. Since it is obvious that *present* belongs to *positive* and *absent* and *not patient* to *negative*, the two mappings encode whether the assertion categories that express hedging (i.e., *conditional*, *hypothetical*, and *possible*) are attached to the *positive* or *negative* class. We denote as $hedge^+$ and $hedge^-$ the mappings where the hedge categories are attached to *positive* and *negative*, respectively. Another aspect that we considered is whether to perform these two mappings *before* training the assertion classifier (i.e., building a binary predictor) or *after* learning the assertion model (i.e., mapping the assertion predictors).

Table 3 shows the results of the assertion classifier based on the mapping configurations described above on the i2b2/VA dataset. As can the noticed, the results when performing the two mappings after learning the assertion model, slightly outperform the results corresponding to the *before* configuration. Also, for both *after* and *before* configurations, the $hedge^+$ mapping is better predicted by the classifier than the $hedge^-$ mapping. Although mapping the hedge classes to the positive class gives slightly better performance, it remains of course to be seen which mapping results in the best performance in the context of the clinical application introduced in the following section.

6. A clinical case study

In this section, we extend our previous work described in [24] and investigate the role of assertion classification in identifying complex illness phenotypes such as pneumonia. Although our current study focuses on pneumonia identification, the methodology we propose for this task can be easily adapted for detecting other phenotypes as well.

Using all the reports associated with a patient, the task of pneumonia identification is to classify the patient as positive or negative for pneumonia. Because this type of application is resource intensive and, at the same time, requires real-time assessments, automatic methods for detecting different types of pneumonia are currently needed in clinical research and by hospitals for pneumonia surveillance.

6.1. Framework for pneumonia identification

The simplest method for deciding whether a patient is positive for pneumonia is to extract the pneumonia expressions from the patient clinical notes, and their assertion values, as described in Section 3. However, in some cases, pneumonia expressions are not even mentioned in the reports. When present, it is in fact often noted that clinical notes will use hedging terms even when there is a fairly high certainty for a patient to be positive for pneumonia, for reasons of liability or because conclusive test results are yet to arrive.

6.1.1. Statistical feature selection

To capture the relevant clinical information for pneumonia identification, we developed a supervised learning framework in which the features associated with a patient correspond to uni-grams and bi-grams of words and Unified Medical Language System (UMLS) concepts from the patient notes. For processing clinical data and classification we used the same NLP platform and SVM classifier as for assertion classification, and for extracting the UMLS concepts we used the 2011 version of MetaMap [25]. In this process, we configured MetaMap such that only the UMLS concepts having the highest mapping score for each match are considered.

Unlike a conventional learning framework, however, we also implemented a methodology to select only the most informative features for pneumonia identification. Specifically, this methodology uses statistical hypothesis testing to measure the association strength between each feature from the training set and the two categories of this task. As a result, the features will be ranked based on those values such that the ones with a strong association to the two categories will be on top. Finally, only the most relevant features that are within a specific threshold will be selected for training. This methodology, called statistical feature selection, has been successfully applied in text categorization [26,27].

To rank the set of features associated with a feature type (e.g., word bi-grams), we constructed a contingency table for each feature from the set and used the *t*-statistic[3] to determine whether there is an association between the feature and the two categories. For instance, Table 4 lists the top 5 uni-grams, bi-grams, and UMLS concepts ranked by this statistical test. As can be observed, many of these features are closely linked to the known causes, clinical signs, and symptoms of pneumonia.

Once all feature sets are ranked and their corresponding threshold values are established, the feature extractor is now able to build a feature vector for each patient. Specifically, given a fixed subset of relevant features from the ranked lists of features, the feature extractor considers in the representation of a patient's feature vector only the features from the subset of relevant features that are also found in the patient's reports. Therefore, the size of the feature space will be equal to the size of the relevant features subset whereas the length of each feature vector will be at most this value.

6.1.2. Assertion of pneumonia expressions

In order to determine whether the task of assertion classification plays a key role in pneumonia identification, we implemented a binary feature, called the *assert feature*, which assigns to each pa-

[3] We also experimented with χ^2 and Fisher exact test, but these measures did not perform as well as the *t* test.

Please cite this article in press as: Bejan CA et al. Assertion modeling and its role in clinical phenotype identification. J Biomed Inform (2012), http://dx.doi.org/10.1016/j.jbi.2012.09.001

Table 5
Baseline results for pneumonia identification.

System configuration	*P*	*R*	NPV	Spec.	*F*
Yetisgen-Yildiz et al.	73.90	38.60	87.32	96.90	50.70
Assert rule ($hedge^{-}$)	44.22	86.36	95.89	72.92	56.72
Assert rule ($hedge^{+}$)	34.11	100.0	100.0	55.73	50.87

tient a label corresponding to positive or negative pneumonia. For this purpose, we selected from the clinical reports only the medical expressions identified by MetaMap that have the same identifier as the pneumonia concept (CUI:C0032285) in the UMLS Metathesaurus [28]. For a more complete set, we also ran simple regular expressions to identify the word *pna*, an abbreviation often used by physicians for pneumonia but which is not tagged as a pneumonia concept in UMLS Metathesaurus yet. After we ran the assertion classifier for all the pneumonia concepts of a patient, we counted how many times each of the six assertion values were identified, and then we mapped the most frequent value to one of the two categories of pneumonia identification as described in Section 5. For the binary features corresponding to patients with no pneumonia concepts identified in their reports, we assigned a default value of negative pneumonia.

6.2. Pneumonia identification experiments

Although the *F*-measure is the primary measure for this task, we also report the negative predictive value (NPV) and specificity (Spec.), since accurately identifying patients for negative pneumonia is equally important in the clinical domain. Furthermore, we measured the statistical significance using the same randomization test as for assertion classification.

The dataset used in our study consists of 3442 reports (e.g., admit notes, discharge summaries, and daily progress notes) corresponding to a cohort of 236 patients. The annotation was performed at patient level by an experienced medical expert based on the information encoded in the patient reports. From the 236 patients, 44 were identified as positive and the remaining 192 as negative for pneumonia.

6.2.1. Baselines

A simple baseline we considered is a rule-based approach which uses as criteria for pneumonia identification the value of the assert feature. Since the dataset of 236 patients was also used by Yetisgen-Yildiz et al. [29] for the same task, we considered their approach as another baseline for our system. In their supervised approach, Yetisgen-Yildiz et al. [29] extracted word *n*-grams, UMLS concepts, and their corresponding semantic types, and represented the feature vector associated with a patient as a 'bag of words' from various combinations of the above mentioned feature types. Using a 5-fold cross validation scheme, their system reached the best performing results when only the word *n*-grams were considered. For an accurate comparison, we used the same evaluation scheme and dataset split as Yetisgen-Yildiz et al. [29] in all our experiments.

Table 5 shows the results of the two baselines described above. As observed, the results of the rule-based approach using both mapping configurations outperform the results of Yetisgen-Yildiz et al. [29]. Furthermore, these results indicate that mapping the hedge assertions to the negative category constitute a better choice for pneumonia identification. Based on this observation, we used the $hedge^{-}$ mapping for all the experiments in the next section involving the assert feature. Because the size of the dataset used in this study is relatively small, none of the baseline results are statistically significant from the other.

6.2.2. Experimenting with the assert feature

We performed a set of experiments to assess the benefit of using the assert feature in combination with other feature types. In this regard, we studied how the performance of our system evolves for these experiments when using various threshold values on the ranked word lists. Fig. 1 shows this experimental study. For each experiment, we considered 27 different values that capture the variation of the threshold for selecting from a range of 10 to 40,000 significant word *n*-grams. For instance, if the feature extractor selects the first 30 features from the ranked lists of word *n*-grams, our system will achieve 70.71 *F*-measure (left plot, *words* + *assert* experiment, threshold = 30). In all the experiments shown in this figure, we used a threshold value of 50 for the selection of the concept *n*-grams.

Although there exist noisy features in the ranked lists of words and concept *n*-grams, which cause some fluctuation in the performance across the range of threshold values, the experiments using the assert feature show substantial improvements over the other experiments. In Fig. 1, it is also worth observing that the results for the last threshold values are close to the results achieved by Yetisgen-Yildiz et al. [29] since, similar to the configuration of their system, our feature extractor selects almost the entire initial set of features.

Table 6 lists the best performing results as well as the aggregate results over all the threshold values considered. For the *concepts* and *concepts* + *assert* experiments, we used a limited set of threshold values (up to 10,000), since the total number of concepts extracted is relatively smaller than the number of word *n*-grams. The thresholds for the word and concept lists are denoted in this table as w_{th} and uc_{th}, respectively. To compute the aggregate results, we used one contingency table with the results generated when considering all threshold values. For example, a total of 236×27 system predictions were considered for all the experiments involving word *n*-grams. For this configuration, the results when adding the assert feature are statistically significant at

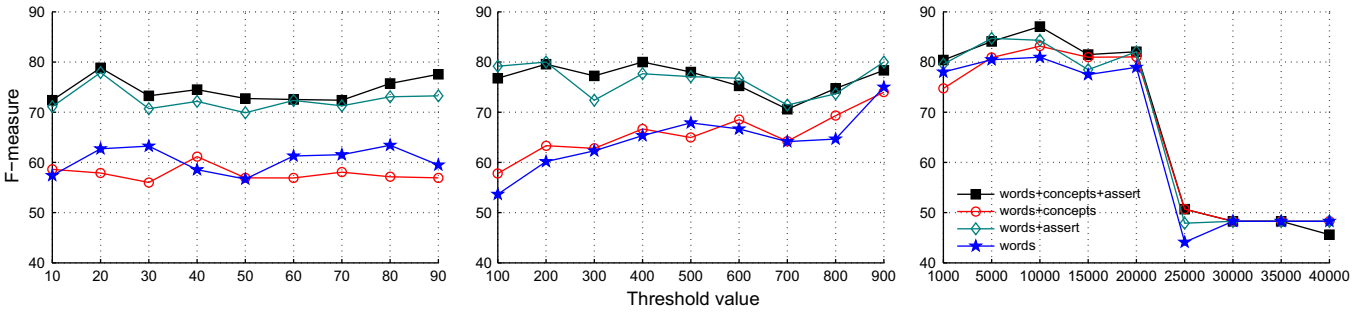

Fig. 1. A study on the impact of the performance results of feature type combinations when using various threshold values on the ranked list of word *n*-grams.

Please cite this article in press as: Bejan CA et al. Assertion modeling and its role in clinical phenotype identification. J Biomed Inform (2012), http://dx.doi.org/10.1016/j.jbi.2012.09.001

Table 6
The best performing results and the aggregate results for pneumonia identification. The differences in performance for the aggregate results when adding the assert feature are statistically significant at $p < .001$ (*).

Feature set	Best results							Aggregate results					
	w_{th}	uc_{th}	P	R	NPV	Spec.	F	uc_{th}	P	R	NPV	Spec.	F
Concepts	–	5000	60.71	77.27	94.44	88.54	68.00	–	38.76	66.60	90.34	74.80	49.00
Concepts + assert	–	5000	67.31	79.55	95.11	91.15	72.92	–	60.20	70.55	92.98	89.31	64.97*
Words	10,000	–	85.00	77.27	94.90	96.88	80.95	–	56.19	72.98	93.11	86.52	63.49
Words + assert	5000	–	87.80	81.82	95.90	97.40	84.71	–	71.35	74.83	94.17	93.11	73.05*
Words + concepts	10,000	50	82.22	84.09	96.34	95.83	83.15	50	54.82	74.66	93.60	85.77	63.22
Words + concepts + assert	10,000	50	90.24	84.09	96.41	97.92	87.06	50	71.06	77.53	94.74	92.77	74.15*

$p < .001$. Also, all the best performing results are statistically significant when compared with the baseline results.

7. Discussion

Our study of assertion classification is fundamentally a study of semantics. Although a medical concept may not at first be conceived of as a predication, it may be viewed as a state that is true to some degree, or not, for a specific patient. Thus, our problem is related to determining the truth-conditionality of predication, which is a fundamental process in understanding text. What we have found in our experiments is that indeed, the truth-conditions of the medical concept matter when trying to determine whether or not the patient exhibits that phenotype.[4] In fact, as we observe in Fig. 1, the assertion feature helps for all possible representations for modeling the phenotype classification. Currently, the assertion feature is present only for medical concepts denoting pneumonia, but as there are other concepts that are also predictive of pneumonia, we expect that expanding our assertion feature to all concepts will show yet further improvements.

Previous studies have used simple word-based features for assertion classification, usually where the cue words are found within some window of words before and after the medical concept. Table 2 captures our experiments that show that each of the sets of features we propose in addition to the baseline features provides a statistically significant performance improvement. To our knowledge, this is the first study to include syntactic features for modeling assertion classification, which had a positive impact on predicting all assertion categories. As we mention in the text, modeling the *conditional* category is particularly challenging, so it is encouraging to see improvement using syntactic features.

Lastly, we find that it is the application in which the semantic feature is used, that ultimately determines the best mapping from the theoretically motivated six assertion classes to the practically-oriented two classes, positive and negative. In the context of our application for predicting pneumonia, it is better to model the phenotype by mapping the hedge classes to a negative instance of pneumonia.

References

[1] Uzuner O, South BR, Shen S, DuVall SL. 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. J Am Med Inform Assoc 2011;18(5):552–6.
[2] Morante R, Sporleder C. Proceedings of the workshop on negation and speculation in natural language processing; 2010.
[3] Farkas R, Vincze V, Móra G, Csirik J, Szarvas G. The CoNLL-2010 shared task: learning to detect hedges and their scope in natural language text. In: Proceedings of the fourteenth conference on computational natural language learning; 2010. p. 1–12.
[4] Aronow DB, Fangfang F, Croft WB. Ad hoc classification of radiology reports. J Am Med Inform Assoc 1999;6(5):393–411.
[5] Mutalik PG, Deshpande A, Nadkarni PM. Use of general-purpose negation detection to augment concept indexing of medical documents: a quantitative study using the UMLS. J Am Med Inform Assoc 2001;8(6):598–609.
[6] 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(5):301–10.
[7] Goldin IM, Chapman WW. Learning to detect negation with 'Not' in medical texts. In: Proceedings of the workshop on text analysis and search for bioinformatics at the 26th annual international ACM SIGIR conference; 2003.
[8] Harkema H, Dowling JN, Thornblade T, Chapman WW. Context: an algorithm for determining negation, experiencer, and temporal status from clinical reports. J Biomed Inform 2009;42(5):839–51.
[9] Uzuner O, Zhang X, Sibanda T. Machine learning and rule-based approaches to assertion classification. J Am Med Inform Assoc 2009;16(1):109–15.
[10] Elkin PL, Brown SH, Bauer BA, Husser CS, Carruth W, Bergstrom LR, et al. A controlled trial of automated classification of negation from clinical notes. BMC Med Inform Decis Mak 2005; 5 (13).
[11] Huang Y, Lowe HJ. A novel hybrid approach to automated negation detection in clinical radiology reports. J Am Med Inform Assoc 2007;14(3):304–11.
[12] de Bruijn B, Cherry C, Kiritchenko S, Martin J, Zhu X. Machine-learned solutions for three stages of clinical information extraction: the state of the art at i2b2 2010. J Am Med Inform Assoc 2011;18(5):557–62.
[13] Roberts K, Harabagiu SM. A flexible framework for deriving assertions from electronic medical records. J Am Med Inform Assoc 2011;18(5):568–73.
[14] Kim Y, Riloff E, Meystre S. Improving classification of medical assertions in clinical notes. In: Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies; 2011. p. 311–6.
[15] C. Quirk, P. Choudhury, J. Gao, H. Suzuki, K. Toutanova, M. Gamon, W. et al. MSR SPLAT, a language analysis toolkit. In: Proceedings of NAACL HLT 2012 demonstration session; 2012. <http://research.microsoft.com/projects/msrsplat>.
[16] Fan R-E, Chang K-W, Hsieh C-J, Wang X-R, Lin C-J. LIBLINEAR: a library for large linear classification. J Mach Learn Res 2008;9:1871–4.
[17] Vincze V, Szarvas G, Farkas R, Móra G, Csirik J. The BioScope corpus: biomedical texts annotated for uncertainty, negation and their scopes. BMC Bioinformatics 2008; 9(Suppl. 11):S9.
[18] Pustejovsky J, Hanks P, Sauri R, See A, Gaizauskas R, Setzer A, et al. The TimeBank corpus. In: Corpus linguistics; 2003. p. 647–56.
[19] Quirk R, Greenbaum S, Leech G, Svartvik J. A comprehensive grammar of the english language. New York: Longman; 1985.
[20] Blanco E, Moldovan D. Semantic representation of negation using focus detection. In: Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies; 2011. p. 581–9.
[21] Noreen E. Computer-intensive methods for testing hypotheses. New York: John Wiley & Sons; 1989.
[22] Fellbaum C. WordNet: an electronic lexical database. MIT Press; 1998.
[23] Clark C, Aberdeen J, Coarr M, Tresner-Kirsch D, Wellner B, Yeh A, et al. MITRE system for clinical assertion status classification. J Am Med Inform Assoc 2011;18(5):563–7.
[24] Bejan CA, Xia F, Vanderwende L, Wurfel MM, Yetisgen-Yildiz M. Pneumonia identification using statistical feature selection. J Am Med Inform Assoc 2012;19(5):817–23.
[25] Aronson AR. Effective mapping of biomedical text to the UMLS metathesaurus: the MetaMap program. In: Proceedings of the American medical informatics association symposium; 2001. p. 17–21.
[26] Yang Y, Pedersen JO. A comparative study on feature selection in text categorization. In: Proceedings of the 14th international conference on machine learning; 1997. p. 412–20.
[27] Mladenić D, Grobelnik M. Feature selection on hierarchy of web documents. Decis Support Syst 2003;35(1):45–87.
[28] Bodenreider O. The unified medical language system (UMLS): integrating biomedical terminology. Nucl Acids Res 2004;32(Database issue):D267–70.
[29] Yetisgen-Yildiz M, Glavan BJ, Xia F, Vanderwende L, Wurfel MM. Identifying patients with pneumonia from free-text intensive care unit reports. In: Proceedings of the ICML workshop on learning from unstructured clinical text; 2011.
[30] Goldstein I, Uzuner O. Does negation really matter? In: Proceedings of the workshop on negation and speculation in natural language processing; 2010. p. 23–7.

[4] Although our findings clearly indicate that assertion classification plays a key role in pneumonia identification, previous studies did not find this task helpful in improving various other clinical applications [30].

Please cite this article in press as: Bejan CA et al. Assertion modeling and its role in clinical phenotype identification. J Biomed Inform (2012), http://dx.doi.org/10.1016/j.jbi.2012.09.001