Does Negation Really Matter?

Ira Goldstein
University at Albany, SUNY
Albany, NY USA
ig4895@albany.edu

Özlem Uzuner
University at Albany, SUNY
Albany, NY USA
ouzuner@albany.edu

Abstract

We explore the role negation and speculation identification plays in the multi-label document-level classification of medical reports for diseases. We identify the polarity of assertions made on noun phrases which reference diseases in the medical reports. We experiment with two machine learning classifiers: one based upon Lucene and the other based upon BoosTexter. We find the performance of these systems on document-level classification of medical reports for diseases fails to show improvement when their input is enhanced by the polarity of assertions made on noun phrases. We conclude that due to the nature of our machine learning classifiers, information on the polarity of phrase-level assertions does not improve performance on our data in a multi-label document-level classification task.

1 Introduction

In the medical domain, a substantial amount of patient data is stored as free text in patient medical report narratives (Spat et al. 2008) and needs to be processed in order to be converted to more widely-useful structured information. These narratives contain a variety of useful information that can support syndromic surveillance (Shapiro 2004), decision support (Fiszman et al. 2000), and problem list generation (Sibanda et al. 2006).

Physicians often assert negative or speculative diagnoses in medical reports (Rao et al. 2003) to keep track of all potential diagnoses that have been considered and to provide information that contrasts with the positive diagnoses (Kim and Park 2006). The noun phrases (NP) associated with negative and speculative assertions in medical reports may be confused with positively asserted NPs, thereby adversely affecting automated classification system performance. In the medical domain, verbs often play a reduced role or are implied in assertions. We therefore focus our investigation of assertions on NPs.

In this paper, we describe the polarity of an assertion as being positive, speculative, or negative. Assertion classification is a generally accepted means for resolving problems caused by negation and speculation. Averbuch et al. (2004) use context to identify negative/positive instances of various symptoms. Mutalik et al. (2001) show that the Unified Medical Language System (UMLS) Metathesaurus can be used to reliably detect negated concepts in medical narratives. Harkema et al. (2009) develop ConText to determine not only positive and negative assertions, but also assertions referencing someone other than the patient.

The literature is filled with reports of systems which employ assertion classification (e.g., Google Scholar lists 134 documents citing Chapman et al.'s (2001) NegEx). However, few reports describe how much assertion classification contributes to the final system performance. Two exceptions are Goldstein et al. (2007) and Ambert and Cohen (2009).

Goldstein et al. develop a hand-crafted rule based system to classify radiological reports from the 2007 Computational Medicine Center (CMC) Challenge (Pestian et al. 2007). They show that negation and speculation play key roles in classifying their reports. Ambert and Cohen apply a machine learning (ML) approach to classifying discharge summaries from the 2008 i2b2 Obesity Challenge (Uzuner 2008). They report that due to "false negations," simply adding negation detection to their base system does not consistently improve performance. Prompted by these contradicting results in the literature, we explore the role assertion classification plays in the multi-label classification of medical reports from both the CMC and i2b2 challenges.

We attempt to improve document-level classification performance of two multi-label ML classifiers by identifying the polarity of assertions on NPs. We experiment with medical reports from two different corpora. We detect NPs which reference diseases. We then identify the polarity of the assertion made for each NP. We show that enriching reports with the polarity of the assertions does not improve performance for multi-label document-level classification of medical

Proceedings of the Workshop on Negation and Speculation in Natural Language Processing, pages 23–27,
Uppsala, July 2010.

reports into diseases in our corpora. Our findings imply that, despite common practice, the contribution of assertion classification may be limited when employing ML approaches to predicting document-level labels of medical reports.

2 Data

The data were provided by the CMC challenge (Pestian et al. 2007) and the i2b2 Obesity Challenge (Uzuner 2008). Both data sets had been de-identified (anonymized) and, where appropriate, re-identified with surrogates. Our task is to determine the presence of diseases in the patient based upon medical report narratives. The institutional review boards of the SUNY Albany and Partners HealthCare approved this study.

2.1 CMC Data Set

The CMC data set consists of a training set of 978 radiology reports and a test set of 976 radiology reports. Each report is labeled with ICD-9-CM (National Center for Health Statistics 2010) standard diagnostic classification codes.

The reports have been hand labeled with 45 ICD-9-CM. Each code represents a distinct disease present in the patient. The codes reflect only the definite diagnoses mentioned in that report. At least one code is assigned to each report. Multiple codes per report are allowed. For each report in the test set, we predict which diseases are present in the patient and label the report with the ICD-9-CM code for that disease. Any code not assigned to a report implies that the corresponding disease is not present in the patient.

2.2 i2b2 Data Set

The i2b2 data set consists of a training set of 720 discharge summaries and a test set of 501 discharge summaries. These medical reports range in size from 133 words to more than 3000 words. The reports have been labeled for information on obesity and 15 of its most frequent co-morbidities. For each report, each disease is labeled as being present, absent, or questionable in the patient, or unmentioned in the narrative. Multiple codes per report are allowed.

Since we are interested in those diseases present in the patient, we retain the present class and collapse the absent, questionable, and unmentioned categories into a not present class. For each report in the test set we predict whether each of the 16 diseases is present or not present in the patient. We label each report with our prediction for each of the 16 diseases.

3 Methods

We preprocess the medical report narratives with a Noun Phrase Detection Pre-processor (NPDP) to detect noun phrases referencing diseases. We implement our own version of ConText (Harkema et al. 2009), enhance it to also detect speculation, and employ it to identify the polarity of assertions made on the detected NPs. We expand the text of the medical reports with asserted NPs. We conflate lexical variations of words. We train two different types of classifiers on each of the training sets. We apply labels to both the expanded and non-expanded reports using two ML classifiers. We evaluate and report results only on the test sets.

3.1 Noun Phrase and Assertion Detection

We detect noun phrases via an NPDP. We build our NPDP based on MetaMap (Aronson 2001). The NPDP identifies NPs which reference diseases in medical reports. We select 17 UMLS semantic types whose concepts can assist in the classification of diseases. First, NPDP maps NPs in the text to UMLS semantic types. If the mapped semantic type is one of the target semantic types, NPDP then tags the NP.

NPDP uses the pre-UMLS negation phrases of Extended NegEx (Sibanda et al. 2006) to identify adjectives indicating the absence or uncertainty of each tagged NPs. It differentiates these adjectives from all other adjectives modifying tagged NPs. For example, *possible* in *possible reflux* is excluded from the tagged NP, whereas *severe* in *severe reflux* is retained. We then identify the polarity of the assertion made on each NP. In order to distinguish the polarity of the assertions from one another, we do not modify the positive assertions, but transform the negative and speculative assertions in the following manner: Sentences containing negative assertions are repeated and modified with the NP pre-pended with "abs" (e.g., "Patient denies fever." is repeated as "Patient denies absfever."). Similarly, sentences containing speculative assertions are repeated and modified with the NP pre-pended with "poss". We refer to these transformed terms as *asserted noun phrases*. We assert NPs for the unmodified text of both the data sets. Table 1 provides a breakdown of the assertions for each of the detected NPs for each of the data sets.

We examine the performance of our enhanced implementation of ConText by comparing its results against CMC test set NPs manually annotated by a nurse librarian and author IG. Table 2

shows the performance for each of the three polarities. We find these results to be comparable to those reported in the literature: Mutalik et al.'s (2001) NegFinder finds negated concepts with a recall of .957; Chapman et al.'s (2001) NegEx report a precision of .8449 and a recall of .8241.

Assertion	CMC		i2b2	
	Training	Test	Training	Test
Positive	2,168	2,117	47,860	34,112
Speculative	312	235	3,264	2,166
Negative	351	353	8,202	5,654

Table 1 - Distribution of Asserted Noun Phrases for both the CMC and i2b2 data sets.

Assertion	Precision	Recall	F1-Measure
Positive	0.991	0.967	0.979
Speculative	0.982	0.946	0.964
Negative	0.770	0.983	0.864

Table 2 - Assertion Performance on the CMC test set.

3.2 Lucene Classifier

We follow the k-Nearest Neighbor (Cover and Hart 1967) process previously described in Goldstein et al. (2007) to build our Lucene-based classifier. Classification is based on the nearest training samples, as determined by the feature vectors. This approach assumes that similar training samples will cluster together in the feature vector space. The nearest training samples are considered to be those that are most similar to the data sample.

We build our Lucene-based classifier using Apache Lucene (Gospodnetić and Hatcher 2005). We use the Lucene library to determine the similarity of medical report narratives. We determine which training reports are similar to the target report based upon their text. For each target report we retrieve the three most similar training reports and assign to the target report any codes that are used by the majority of these reports. In cases where the retrieved reports do not provide a majority code, the fourth nearest training report is used. If a majority code is still not found, a NULL code is assigned to the target report.

We first run the Lucene Classifier on lower case, stemmed text of the medical reports. We refer to this as the *Base Lucene Classifier* run. We next run the Lucene Classifier on the text expanded with asserted noun phrases. We refer to this as the *Asserted Lucene Classifier* run.

3.3 BoosTexter Classifier

BoosTexter (Schapire and Singer 2000) builds classifiers from textual data by performing multiple iterations of dividing the text into subsamples upon which weak decision-stub learners are trained. Among these weak learners, BoosTexter retains those that perform even marginally better than chance. After a set number of iterations, the retained weak learners are combined into the final classifier. BoosTexter classifies text using individual words (unigrams), strings of consecutive words (n-grams), or strings of non-consecutive words, without considering semantics.

We cross-validate BoosTexter (tenfold) on the CMC training set. We establish the optimal parameters on the CMC training set to be 1100 iterations, with n-grams of up to four words. We find the optimal parameters of the i2b2 training set to be similar to those of the CMC training set. For consistency, we apply the parameters of 1100 iterations and n-grams of up to four words to both data sets. In addition, we apply unigrams to BoosTexter in order to provide BoosTexter classifier results that are comparable to those of the Lucene classifiers.

We create two classifiers with BoosTexter using the lower case, stemmed text of the medical reports: one with unigrams and one with n-grams. We refer to these as *Base BoosTexter Classifier* runs. For each of unigrams and n-grams, we create runs on the text expanded with the asserted noun phrases. We refer to these as *Asserted BoosTexter Classifier* runs.

4 Evaluation

We evaluate our classifiers on both the plain text of the reports and on text expanded with asserted NPs. We present results in terms of micro-averaged precision, recall, and F1-measure (Özgür et al. 2005). We check the significance of classifier performance differences at α=0.10. We apply a two-tailed Z test, with $Z = \pm 1.645$.

5 Results and Discussion

Table 3 and Table 4 show our systems' performances. We predict ICD-9-CM codes for each of the 976 CMC test reports. We predict whether or not each of 16 diseases is present in the patient for each of the 501 i2b2 test set reports.

	Negative Reports			Positive Reports		
Run	Precision	Recall	F1-Measure	Precision	Recall	F1-Measure
CMC Base	0.991	0.993	0.992	0.717	0.664	0.690
CMC Asserted	0.991	0.992	0.992	0.712	0.668	0.690
i2b2 Base	0.905	0.886	0.896	0.612	0.660	0.635
i2b2 Asserted	0.904	0.890	0.897	0.618	0.651	0.634

Table 3 - Lucene Classifier's Performance.

The Asserted Lucene and BoosTexter Classifier runs show no significant difference in performance from their Base runs on either corpus. These results indicate that asserted noun phrases do not contribute to the document-level classification of our medical reports

5.1 Contribution of Asserted Noun Phrases

Through analysis of the Base and Asserted runs, we find enough similarities in the text of the training and test reports for a given class to allow our ML classifiers to correctly predict the labels without needing to identify the polarity of the assertions made on individual NPs. For example, for the CMC target report 97729923:

```
5-year-9-month - old female
with two month history of
cough. Evaluate for pneumonia.
No pneumonia.
```

the Base Lucene Classifier retrieves report 97653364:

```
Two - year-old female with
cough off and on for a month
(report states RSV nasal
wash).
No radiographic features of
pneumonia.
```

which allows the system to classify the target report with the ICD-9-CM code for cough. While identifying the polarity of the assertions for pneumonia strengthens the evidence for cough and not pneumonia, it cannot further improve the already correct document-level classification. These unenhanced assertions do not stand in the way of correct classification by our systems.

5.2 Approach, Data, and Task

Hand-crafted rule-based approaches usually encode the most salient information that the experts would find useful in classification and would therefore benefit from explicit assertion classification subsystems, e.g., Goldstein et al., (2007). On the other hand, ML approaches have the ability to identify previously undetected patterns in data (Mitchell et al. 1990). This enables ML approaches to find patterns that may not be obvious to experts, while still performing correct classification. Therefore, the contribution of asserted NPs appears to be limited when applied to ML approaches to document-level classification of medical reports. This is not to say that an ML approach to document-level classification will never benefit from identifying the polarity of NPs; only that on our data we find no improvement.

		Negative Reports			Positive Reports		
Run		Precision	Recall	F1-Measure	Precision	Recall	F1-Measure
CMC unigram	Base	0.993	0.995	0.994	0.812	0.747	0.778
	Asserted	0.993	0.996	0.995	0.837	0.767	0.800
CMC ngram	Base	0.995	0.996	0.996	0.865	0.812	0.838
	Asserted	0.995	0.996	0.996	0.866	0.812	0.839
i2b2 unigram	Base	0.970	0.973	0.917	0.902	0.889	0.895
	Asserted	0.970	0.975	0.973	0.908	0.891	0.899
i2b2 ngram	Base	0.971	0.976	0.974	0.911	0.895	0.903
	Asserted	0.974	0.977	0.975	0.914	0.903	0.908

Table 4 - BoosTexter Classifier's Performance.

The CMC and i2b2 data sets can each be described as being homogenous; they come from a relatively small communities and limited geographic areas. In these data, variation in vocabulary that might arise from the use of regional expressions would be limited. This would be especially true for the CMC data since it comes from a single medical department at a single hospital. It would not be surprising for colleagues in a given department who work together for a period of time to adopt similar writing styles and to employ consistent terminologies (Suchan 1995).

Our task is one of multi-label document-level classification. Working at the document level, each negative and speculative assertion would play only a small role in predicting class labels.

The homogeneity of the text in our data sets, and the task of document-level classification may have been factors in our results. Future research should examine how the characteristics of the data and the nature of the task affect the role of assertion classification.

6 Conclusion

Identifying the polarity of phrase-level assertions in document-level classification of medical reports may not always be necessary. The specific task and approach applied, along with the characteristics of the corpus under study, should be considered when deciding the appropriateness of assertion classification. The results of this study show that on our data and task, identifying the polarity of the assertions made on noun phrases does not improve machine learning approaches to multi-label document-level classification of medical reports.

References

Kyle H. Ambert and Aaron M. Cohen. 2009. A System for Classifying Disease Comorbidity Status from Medical Discharge Summaries Using Automated Hotspot and Negated Concept Detection. *Journal of the American Medical Informatics Association* 16(4):590-95.

Alan R. Aronson. 2001. Effective Mapping of Biomedical Text to the UMLS Metathesaurus: The Metamap Program. *Proceedings of the AMIA symposium*. 17-21.

Mordechai Averbuch, Tom H. Karson, Benjamin Ben-Ami, Oded Maimon, and Lior Rokach. 2004. Context-Sensitive Medical Information Retrieval. *Medinfo. MEDINFO* 11(Pt 1):282-86.

Wendy W. Chapman, Will Bridewell, Paul Hanbury, Gregory F. Cooper, and Bruce G. Buchanan. 2001. A Simple Algorithm for Identifying Negated Findings and Diseases in Discharge Summaries. *Journal of Biomedical Informatics* 34(5):301-10.

Thomas M. Cover and Peter E. Hart. 1967. Nearest Neighbor Pattern Classification. *IEEE Transactions on Information Theory* 13(1):21-27.

Marcelo Fiszman, Wendy W. Chapman, Dominik Aronsky, and R. Scott Evans. 2000. Automatic Detection of Acute Bacterial Pneumonia from Chest X-Ray Reports. *Journal of the American Medical Informatics Association* 7:593-604.

Ira Goldstein, Anna Arzumtsyan, and Özlem Uzuner. 2007. Three Approaches to Automatic Assignment of ICD-9-CM Codes to Radiology Reports. *Proceedings of the AMIA symposium*. 279-83.

Otis Gospodnetić and Erik Hatcher. 2005. *Lucene in Action*. Greenwich, CT: Manning.

Henk Harkema, John N. Dowling, Tyler Thornblade, and Wendy W. Chapman. 2009. Context: An Algorithm for Determining Negation, Experiencer, and Temporal Status from Clinical Reports. *Journal Of Biomedical Informatics* 42(5):839-51.

Jung-Jae Kim and Jong C. Park. 2006. Extracting Contrastive Information from Negation Patterns in Biomedical Literature. *ACM Transactions on Asian Language Information Processing (TALIP)* 5(1):44-60.

Tom Mitchell, Bruce Buchanan, Gerald DeJong, Thomas Dietterich, Paul Rosenbloom, and Alex Waibel. 1990. Machine Learning. *Annual Review of Computer Science. Vol.4*. Eds. Joseph F. Traub, Barbara J. Grosz, Butler W. Lampson and Nils J. Nilsson. Palo Alto, CA: Annual Reviews.

Pradeep G. Mutalik, Aniruddha Deshpande, and Prakash M. Nadkarni. 2001. Use of General-Purpose Negation Detection to Augment Concept Indexing of Medical Documents: A Quantitative Study Using the UMLS. *Journal of the American Medical Informatics Association* 8(6):598-609.

National Center for Health Statistics. 2010. *ICD - ICD-9-CM - International Classification of Diseases, Ninth Revision, Clinical Modification*. Accessed: May 1, 2010. <www.cdc.gov/nchs/icd/icd9cm.htm>.

Arzucan Özgür, Levent Özgür, and Tunga Güngör. 2005. Text Categorization with Class-Based and Corpus-Based Keyword Selection. *ISCIS 2005*. Eds. Pınar Yolum, Tunga Güngör, Fikret Gürgen and Can Özturan. Istanbul, Turkey: Springer. 606-15 of *Lecture Notes in Computer Science*.

John P. Pestian, Christopher Brew, Pawel Matykiewicz, D. J. Hovermale, Neil Johnson, K. Bretonnel Cohen, and Włodzisław Duch. 2007. A Shared Task Involving Multi-Label Classification of Clinical Free Text. *ACL:BioNLP*. Prague: Association for Computational Linguistics. 97-104.

R. Bharat Rao, Sathyakama Sandilya, Radu Stefan Niculescu, Colin Germond, and Harsha Rao. 2003. Clinical and Financial Outcomes Analysis with Existing Hospital Patient Records. *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*: ACM Press New York, NY, USA. 416-25.

Robert E. Schapire and Yoram Singer. 2000. Boostexter: A Boosting-Based System for Text Categorization. *Machine Learning* 39(2):135-68.

Alan R. Shapiro. 2004. Taming Variability in Free Text: Application to Health Surveillance. *MMWR. Morbidity And Mortality Weekly Report* 53 Suppl:95-100.

Tawanda Carleton Sibanda, T. He, Peter Szolovits, and Özlem Uzuner. 2006. Syntactically-Informed Semantic Category Recognition in Discharge Summaries. *Proceedings of the AMIA symposium*. 714-8.

Stephan Spat, Bruno Cadonna, Ivo Rakovac, Christian Gütl, Hubert Leitner, Günther Stark, and Peter Beck. 2008. Enhanced Information Retrieval from Narrative German-Language Clinical Text Documents Using Automated Document Classification. *Studies In Health Technology And Informatics* 136:473-78.

Jim Suchan. 1995. The Influence of Organizational Metaphors on Writers' Communication Roles and Stylistic Choices. *Journal of Business Communication* 32(1):7-29.

Özlem Uzuner. 2008. Second I2b2 Workshop on Natural Language Processing Challenges for Clinical Records. *Proceedings of the AMIA symposium*:1252-53.