Extending NegEx with Kernel Methods for Negation Detection in Clinical Text

Chaitanya Shivade[†], **Marie-Catherine de Marneffe**[§], **Eric Fosler-Lussier**[†], **Albert M. Lai**[*]
[†]Department of Computer Science and Engineering,
[§]Department of Linguistics,
[*]Department of Biomedical Informatics,
The Ohio State University, Columbus OH 43210, USA.
shivade@cse.ohio-state.edu, mcdm@ling.ohio-state.edu
fosler@cse.ohio-state.edu, albert.lai@osumc.edu

Abstract

NegEx is a popular rule-based system used to identify negated concepts in clinical notes. This system has been reported to perform very well by numerous studies in the past. In this paper, we demonstrate the use of kernel methods to extend the performance of NegEx. A kernel leveraging the rules of NegEx and its output as features, performs as well as the rule-based system. An improvement in performance is achieved if this kernel is coupled with a bag of words kernel. Our experiments show that kernel methods outperform the rule-based system, when evaluated within and across two different open datasets. We also present the results of a semi-supervised approach to the problem, which improves performance on the data.

1 Introduction

Clinical narratives consisting of free-text documents are an important part of the electronic medical record (EMR). Medical professionals often need to search the EMR for notes corresponding to specific medical events for a particular patient. Recruitment of subjects in research studies such as clinical trials involves searching through the EMR of multiple patients to find a cohort of relevant candidates. Most information retrieval approaches determine a document to be relevant to a concept based on the presence of that concept in the document. However, these approaches fall short if these concepts are negated, leading to a number of false positives. This is an important problem especially in the clinical domain. For example, the sentence: "The scan showed no signs of malignancy" has the concept 'malignancy' which was looked for in the patient, but was not observed to be present. The task of negation detection is to identify whether a given concept is negated or affirmed in a sentence. NegEx (Chapman et al., 2001) is a rule-based system developed to detect negated concepts in the clinical domain and has been extensively used in the literature.

In this paper, we show that a kernel-based approach can map this rule-based system into a machine learning system and extends its performance. We validate the generalization capabilities of our approach by evaluating it across datasets. Finally, we demonstrate that a semi-supervised approach can also achieve an improvement over the baseline rule-based system, a valuable finding in the clinical domain where annotated data is expensive to generate.

2 Related Work

Negation has been a popular research topic in the medical domain in recent years. NegEx (Chapman et al., 2001) along with its extensions (South et al., 2006; Chapman et al., 2013) is one of the oldest and most widely used negation detection system because of its simplicity and speed. An updated version - ConText (Harkema et al., 2009) was also released to incorporate features such as temporality and experiencer identification, in addition to negation. These algorithms are designed using simple rules that fire based on the presence of particular cues, before and after the concept. However, as with all rule-based systems, they lack generalization. Shortage of training data discouraged the use of machine learning techniques in clinical natural language processing

Proceedings of the Second Workshop on Extra-Propositional Aspects of Meaning in Computational Semantics (ExProM 2015), pages 41–46,
Denver, Colorado, June 5, 2015.

(NLP) in the past. However, shared tasks (Uzuner et al., 2011) and other recent initiatives (Albright et al., 2013) are making more clinical data available. This should be leveraged to harness the benefits offered by machine learning solutions. Recently, Wu et al. (2014) argued that negation detection is not of practical value without in-domain training and/or development, and described an SVM-based approach using hand-crafted features.

3 Kernel Methods

Our approach uses kernel methods to extend the abilities of the NegEx system. A kernel is a similarity function K, that maps two inputs x and y from a given domain into a similarity score that is a real number (Hofmann et al., 2008). Formally, it is a function $K(x, y) = \langle \phi(x), \phi(y) \rangle \rightarrow R$, where $\phi(x)$ is some feature function over instance x. For a function K to be a valid kernel, it should be symmetric and positive-semidefinite. In this section, we describe the different kernels we implemented for the task of negation detection.

3.1 NegEx Features Kernel

The source code of NegEx[1] reveals rules using three sets of negation cues. These are termed as pseudo negation phrases, negation phrases and post negation phrases. Apart from these cues, the system also looks for a set of conjunctions in a sentence. We used the source code of the rule-based system and constructed a binary feature corresponding to each cue and conjunction, and thus generated a feature vector for every sentence in the dataset. Using the LibSVM (Chang and Lin, 2011) implementation, we constructed a linear kernel which we refer to as the NegEx Features Kernel (NF).

3.2 Augmented Bag of Words Kernel

We also designed a kernel that augmented with bag of words the decision by NegEx. For each dataset, we constructed a binary feature vector for every sentence. This vector is comprised of two parts, a vector indicating presence or absence of every word in that dataset and augment it with a single feature indicating the output of the NegEx rule-based system. We did not filter stop-words since many stop-words serve as cues for negated assertions. The idea behind constructing such a kernel was to allow the model to learn relative weighting of the NegEx output and the bag of words in the dataset. Again, a linear kernel using LibSVM was constructed: the Augmented Bag of Words Kernel (ABoW).

4 Datasets

A test set of de-identified sentences, extracted from clinical notes at the University of Pittsburgh Medical Center, is also available with the NegEx source code. In each sentence, a concept of interest has been annotated by physicians with respect to being negated or affirmed in the sentence. The concepts are non numeric clinical conditions (such as symptoms, findings and diseases) extracted from six types of clinical notes (e.g., discharge summaries, operative notes, echo-cardiograms).

The 2010 i2b2 challenge (Uzuner et al., 2011) on relation extraction had assertion classification as a subtask. The corpus for this task along with the annotations is freely available for download.[2] Based on a given target concept, participants had to classify assertions as either present, absent, or possible in the patient, conditionally present in the patient under certain circumstances, hypothetically present in the patient at some future point, and mentioned in the patient report but associated with someone other than the patient. Since we focus on negation detection, we selected only assertions corresponding to the positive and negative classes from the five assertion classes in the corpus, which simulates the type of data found in the NegEx Corpus. The i2b2 corpus has training data, partitioned into discharge summaries from Partners Healthcare (PH) and the Beth Israel Deaconess (BID) Medical Center. This gave us datasets from two more medical institutions. The corpus also has a test set, but does not have a split corresponding to these institutions.

Using the above corpora we constructed five datasets: 1) The dataset available with the NegEx rule-based system, henceforth referred to as the NegCorp dataset; 2) We adapted the training set of the i2b2 assertion classification task for negation detection, the i2b2Train$_{mod}$ dataset; 3) The training subset of i2b2Train$_{mod}$ from Partners Health-

[1] From https://code.google.com/p/NegEx/

[2] From https://www.i2b2.org/NLP/DataSets/

Dataset	Affirmed	Negated	Total
NegCorp	1885	491	2376
i2b2Train$_{mod}$	4476	1533	6009
PH subset	(1862)	(635)	(2497)
BID subset	(2614)	(898)	(3512)
i2b2Test$_{mod}$	8618	2594	11212

Table 1: Number of affirmed and negated concepts in each dataset.

care, henceforth referred to as the PH dataset; 4) The training subset of i2b2Train$_{mod}$ from the Beth Israel Deaconess Medical Center, henceforth referred to as the BID dataset; and 5) The adapted test set of the 2010 i2b2 challenge, henceforth referred to as the i2b2Test$_{mod}$ dataset. Table 1 summarizes the distribution for number of affirmed and negated assertions in each dataset.

5 Experiments

We implemented the kernels outlined in Section 3 and evaluated them within different datasets using precision, recall and F1 on ten-fold cross validation. We compared the performance of each model against the NegEx rule-based system as baseline.

5.1 Within dataset evaluation

As can be seen in Table 2, the NegEx Features Kernel performed similarly to the baseline (the improvement is not significant). However, the ABoW kernel significantly outperformed the baseline ($p<0.05$, McNemar's test). Joachims et al. (2001) showed that given two kernels K1 and K2, the composite kernel $K(x, y) = K1(x, y) + K2(x, y)$ is also a kernel. We constructed a composite kernel adding the kernel matrices for the ABoW and NF kernels, which resulted in a further (but not significant) improvement.

5.2 Cross dataset evaluation

In order to test the generalizability of our approach, we evaluated the performance of the ABoW kernel against the baseline. We trained the ABoW kernel on different datasets and tested them on the i2b2Test$_{mod}$ dataset. Table 3 summarizes the results of these experiments.

	Datasets		
System	NegCorp	BID	PH
NegEx (baseline)	94.6	84.2	87.3
NF Kernel	95.6	87.3	87.5
ABoW Kernel	**97.0**	**90.6**	**89.9**
ABoW+ NF Kernel	**97.3**	**92.4**	**91.3**

Table 2: Within dataset performance of kernels based on F1-score using 10-fold cross validation. Bold results indicate significant improvements over the baseline ($p<0.05$, McNemar's test).

System	Precision	Recall	F1
NegEx (baseline)	89.6	79.9	84.5
ABoW trained on			
NegCorp	89.9	79.3	84.2
PH	89.4	**88.1**	**88.7**
BID	89.2	**89.9**	**89.7**
i2b2Train$_{mod}$	89.9	**90.0**	**90.0**

Table 3: Cross dataset performance on the i2b2Test$_{mod}$ dataset given different training datasets.

We found that the ABoW kernel significantly outperformed the baseline when trained on datasets that were generated from the same corpus, namely PH and BID. A kernel trained on i2b2Train$_{mod}$, i.e., combining the PH and BID datasets performs better than the individually trained datasets. These experiments also tested the effect of training data size (PH $<$ BID $<$ i2b2Train$_{mod}$) on the kernel performance. We observed that the performance of the kernel increases as the size of the training data increases, though not significantly. The kernel trained on a dataset from a different corpus (NegCorp) performs as well as the baseline.

5.3 Semi-supervised approach

We tried a semi-supervised approach to train the ABoW kernel, which we tested on the i2b2Test$_{mod}$ dataset. We trained a kernel on the NegCorp dataset and recorded the predictions. We refer to these labels as "pseudo labels" in contrast to the gold labels of the i2b2Train$_{mod}$ dataset. We then trained a semi-supervised ABoW kernel, ABoW$_{ss}$ on the i2b2Train$_{mod}$ dataset to learn the pseudo labels for

this predicted dataset. Finally, we tested ABoW$_{ss}$ on the i2b2Test$_{mod}$ dataset. Table 4 summarizes the results of these experiments. For ease of comparison, we restate the results of the ABoW kernel, ABoW$_{gold}$ trained on the gold labels of the i2b2Train$_{mod}$ dataset.

System	Precision	Recall	F1
NegEx	89.6	79.9	84.5
ABoW$_{ss}$	89.7	**82.1**	**85.7**
ABoW$_{gold}$	89.9	**90.0**	**90.0**

Table 4: Semi-supervised models on the i2b2Test$_{mod}$ dataset.

These results demonstrate that the kernel trained using a semi-supervised approach performs better than the baseline ($p<0.05$, McNemar's test), but performs worse than a kernel trained using supervised training. However, supervised training is dependent on gold annotations. Thus, the semi-supervised approach achieves good results without the need for annotated data. This is an important result especially in the clinical domain where available annotated data is sparse and extremely costly to generate.

6 Dependency Tree Kernels

Dependency tree kernels have been showed to be effective for NLP tasks in the past. Culotta et al. (2004) showed that although tree kernels by themselves may not be effective for relation extraction, combining a tree kernel with a bag of words kernel showed promising results. Dependency tree kernels have also been explored in the context of negation extraction in the medical domain. Recently, Bowei et al. (2013) demonstrated the use of tree kernel based approaches in detecting the scope of negations and speculative sentences using the BioScope corpus (Szarvas et al., 2008). However, the task of negation scope detection task is different than that of negation detection. Among other differences, an important one being the presence of annotations for negation cues in the Bioscope corpus. Sohn et al. (2012) developed hand crafted rules representing subtrees of dependency parses of negated sentences and showed that they were effective on a dataset from their institution.

Therefore, we implemented a dependency tree kernel similar to the approach described in Culotta and Sorensen (2004) to automatically capture the structural patterns in negated assertions. We used the Stanford dependencies parser (version 2.0.4) (de Marneffe et al., 2006) to get the dependency parse for every assertion. As per their representation (de Marneffe and Manning, 2008) every dependency is a triple, consisting of a governor, a dependent and a dependency relation. In this triple, the governor and dependent are words from the input sentence. Thus, the tree kernel comprised of nodes corresponding to every word and every dependency relation in the parse. Node similarity was computed based on features such as lemma, generalized part-of-speech, WordNet (Fellbaum, 1998) synonymy and the UMLS (Bodenreider, 2004) semantic type obtained using MetaMap (Aronson, 2001) for word nodes.

Node similarity for dependency relation nodes was computed based on name of the dependency relation. A tree kernel then computed the similarity between two trees by recursively computing node similarity between two nodes as described in (Culotta and Sorensen, 2004). The only difference being, unlike our approach they have only word nodes in the tree. The kernel is hence a function $K(T1, T2)$ which computes similarity between two dependency trees $T1$ and $T2$. See (Culotta and Sorensen, 2004) for why K is a valid kernel function. However, we got poor results. In experiments involving within dataset evaluation, the tree kernel gave F1 scores of 77.0, 76.2 and 74.5 on NegCorp, BID and PH datasets respectively. We also tried constructing composite kernels, by adding kernel matrices of the tree kernel and the ABoW kernel or NF kernel, hoping that they captured complimentary similarities between assertions. Although performance of the composite kernel was better than the tree kernel itself, there was no significant gain in the performance as compared to those of the reported kernels.

7 Discussion

We observe that while the precision of all the classifiers is almost constant across all the set of experiments, it is the recall that changes the F1-score. This

implies that the kernel fetches more cases than the baseline. The bag of words contributes towards the increase in recall and thus raises performance.

It is instructive to look at sentences that were misclassified by NegEx but correctly classified by the ABoW$_{gold}$ system. The NegEx rule-based system looks for specific phrases, before or after the target concept, as negation cues. The scope of the negation is determined using these cues and the presence of conjunctions. False positives stem from instances where the scope is incorrectly calculated. For example, in "No masses, neck revealed lymphadenopathy", the concept 'lymphadenopathy' is taken to be negated. The issue of negation scope being a shortcoming of NegEx has been acknowledged by its authors in Chapman et al. (2001). There were certain instances where the NegEx negation cues and the target concept overlapped. For example, in "A CT revealed a large amount of free air", the target concept 'free air' was wrongly identified by NegEx as negated. This is because 'free' is a post negation cue, to cover cases such as "The patient is now symptom free". Similarly, with 'significant increase in tumor burden' as the target concept, the sentence "A staging CT scan suggested no significant increase in tumor burden" was wrongly identified as an affirmation. Since the closest negation cue was 'no significant,' NegEx would identify only concepts after the phrase 'no significant' as negated. We also found interesting cases such as, the "Ext: cool, 1 + predal pulses, - varicosities, - edema." where the concept 'varicosities' is negated using the minus sign.

We studied cases where NegEx made the right decision but which were incorrectly classified by our system. For example, in the assertion "extrm - trace edema at ankles, no cyanosis, warm/dry", the kernel incorrectly classified "trace edema" as negated. In "a bone scan was also obtained to rule out an occult hip fracture which was negative", the concept "occult hip fracture" was incorrectly classified as affirmed. We found no evident pattern in these examples.

The tree kernel was constructed to automatically capture subtree patterns similar to those handcrafted by Sohn et al. (2012). Although, it resulted in a poor performance, there are a number of possibilities to improve the current model of the kernel. Clinical data often consists of multi-word expressions (e.g., "congestive heart failure"). However, the word nodes in our dependency tree kernel are unigrams. Aggregating these unigrams (e.g., identification using MetaMap, followed by use of underscores to replace whitespaces) to ensure they appear as a single node in the tree could give dependency parses that are more accurate. Similarity for nodes involving dependency tree relations; similarity in our kernel is a binary function examining identical names for dependency relations. This could be relaxed by clustering of dependency relations and computing similarity based on these clusters. We followed Culotta and Sorensen (2004) and used WordNet synonymy for similarity of word nodes. However, open-domain terminologies such as WordNet are known to be insufficient for tasks specific to the biomedical domain (Bodenreider and Burgun, 2001). This could be coupled with domain specific resources such as UMLS::Similarity (McInnes et al., 2009) for a better estimate of similarity. Finally, since learning structural patterns is a complex task achieved by the tree kernel; training with a larger amount of data could result in improvements.

8 Conclusion

We demonstrate the use of kernel methods for the task of negation detection in clinical text. Using a simple bag of words kernel with the NegEx output as an additional feature yields significantly improved results as compared to the NegEx rule-based system. This kernel generalizes well and shows promising results when trained and tested on different datasets. The kernel outperforms the rule-based system primarily due to an increase in recall. We also find that for instances where we do not have additional labeled training data, we are able to leverage the NegEx Corpus as a bootstrap to perform semi-supervised learning using kernel methods.

Acknowledgments

Research reported in this publication was supported by the National Library of Medicine of the National Institutes of Health under award number R01LM011116. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

Daniel Albright, Arrick Lanfranchi, Anwen Fredriksen, William F Styler, Colin Warner, Jena D Hwang, Jinho D Choi, Dmitriy Dligach, Rodney D Nielsen, James Martin, et al. 2013. Towards comprehensive syntactic and semantic annotations of the clinical narrative. *Journal of the American Medical Informatics Association*, 20(5):922–930.

Alan R Aronson. 2001. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. In *Proceedings of AMIA Annual Symposium.*, pages 17–21.

Olivier Bodenreider and Anita Burgun. 2001. Comparing terms, concepts and semantic classes in WordNet and the Unified Medical Language System. In *Proceedings of the NAACL 2001 Workshop: WordNet and other lexical resources: Applications, extensions and customizations.*, pages 77–82.

Olivier Bodenreider. 2004. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Research*, 32(Database issue):D267–70.

Zou Bowei, Zhou Guodong, and Zhu Qiaoming. 2013. Tree Kernel-based Negation and Speculation Scope Detection with Structured Syntactic Parse Features. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 968–976.

Chih-Chung Chang and Chih-Jen Lin. 2011. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27.

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, October.

Wendy W Chapman, Dieter Hilert, Sumithra Velupillai, Maria Kvist, Maria Skeppstedt, Brian E Chapman, Michael Conway, Melissa Tharp, Danielle L Mowery, and Louise Deleger. 2013. Extending the NegEx lexicon for multiple languages. *Studies in health technology and informatics*, 192:677.

Aron Culotta and Jeffrey Sorensen. 2004. Dependency Tree Kernels for Relation Extraction. In *Proceedings of the 42Nd Annual Meeting on Association for Computational Linguistics*, ACL '04.

Marie-Catherine de Marneffe and Christopher D. Manning. 2008. The Stanford typed dependencies representation. In *Coling 2008: Proceedings of the workshop on Cross-Framework and Cross-Domain Parser Evaluation*, pages 1–8.

Marie-Catherine de Marneffe, Bill MacCartney, and Christopher D Manning. 2006. Generating Typed Dependency Parses from Phrase Structure Parses. In *Proceedings of LREC*, pages 449–454.

Christiane Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. Bradford Books.

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):839–851.

Thomas Hofmann, Bernhard Schölkopf, and Alexander J. Smola. 2008. Kernel methods in machine learning. *The Annals of Statistics*, 36(3):1171–1220, June.

Thorsten Joachims, Nello Cristianini, and John Shawe-Taylor. 2001. Composite Kernels for Hypertext Categorisation. In *Proceedings of the Eighteenth International Conference on Machine Learning*, pages 250–257. Morgan Kaufmann Publishers Inc.

Bridget T McInnes, Ted Pedersen, and Serguei V S Pakhomov. 2009. UMLS-Interface and UMLS-Similarity : Open Source Software for Measuring Paths and Semantic Similarity. In *Proceedings of the Annual AMIA Symposium.*, volume 2009, pages 431–5.

Sunghwan Sohn, Stephen Wu, and Christopher G Chute. 2012. Dependency Parser-based Negation Detection in Clinical Narratives. In *Proceedings of AMIA Summits on Translational Science*.

Brett R South, Shobha Phansalkar, Ashwin D Swaminathan, Sylvain Delisle, Trish Perl, and Matthew H Samore. 2006. Adaptation of the NegEx algorithm to Veterans Affairs electronic text notes for detection of influenza-like illness (ILI). In *Proceedings of the AMIA Annual Symposium*, pages 1118–1118.

György Szarvas, Veronika Vincze, Richárd Farkas, and János Csirik. 2008. The BioScope Corpus: Annotation for Negation, Uncertainty and Their Scope in Biomedical Texts. In *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing*, BioNLP '08, pages 38–45.

Özlem Uzuner, Brett R South, Shuying Shen, and Scott L DuVall. 2011. 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association : JAMIA*, 18(5):552–6.

Stephen Wu, Timothy Miller, James Masanz, Matt Coarr, Scott Halgrim, David Carrell, and Cheryl Clark. 2014. Negations Not Solved: Generalizability Versus Optimizability in Clinical Natural Language Processing. *PloS one*, 9(11):e112774.