Learning Structures of Negations from Flat Annotations

Vinodkumar Prabhakaran
Department of Computer Science
Columbia University
New York, NY, USA
vinod@cs.columbia.edu

Branimir Boguraev
IBM Watson
Thomas J. Watson Research Center
Yorktown Heights, NY, USA
bran@us.ibm.com

Abstract

We propose a novel method to learn negation expressions in a specialized (medical) domain. In our corpus, negations are annotated as 'flat' text spans. This allows for some infelicities in the mark-up of the ground truth, making it less than perfectly aligned with the underlying syntactic structure. Nonetheless, the negations thus captured are correct in intent, and thus potentially valuable. We succeed in training a model for detecting the negated predicates corresponding to the annotated negations, by re-mapping the corpus to anchor its 'flat' annotation spans into the predicate argument structure. Our key idea—re-mapping the negation instance spans to more uniform syntactic nodes—makes it possible to re-frame the learning task as a simpler one, and to leverage an imperfect resource in a way which enables us to learn a high performance model. We achieve high accuracy for negation detection overall, 87%. Our re-mapping scheme can be constructively applied to existing flatly annotated resources for other tasks where syntactic context is vital.

1 Introduction

Accounting for extra-propositional aspects of meaning in text is a very active NLP research area in recent years, exploring different aspects of meaning such as *factivity* (Saurí and Pustejovsky, 2009), *uncertainty/hedging* (Farkas et al., 2010), *committed belief* (Prabhakaran et al., 2010), and *modalities* (Prabhakaran et al., 2012a). Among these, negation detection has generated special interest because of demonstrated needs for negation detection capability in practical applications such as information retrieval (Averbuch et al., 2004), information extraction (Meystre et al., 2008), sentiment analysis (Wiegand et al., 2010; Councill et al., 2010), and relation detection (Chowdhury and Lavelli, 2013).

Accurately detecting negations is especially important in systems processing medical/clinical text. Consider the segment *"Mild hyperinflation without focal pneumonia"*, taken from a patient's clinical record. It indicates the absence of *focal pneumonia* in the patient. Not capturing this extra-propositional aspect of negation concerning *focal pneumonia* will lead to wrong—and harmful—inferences in downstream processing, e.g. by a clinical decision support system. The need for sophisticated negation detection capabilities in clinical text is even more urgent given the broadening spectrum of applications in this domain: clinical question answering (Lee et al., 2006), clinical decision support (Demner-Fushman et al., 2009), medical information extraction (Uzuner et al., 2010), medical entity relation mining (Tymoshenko et al., 2012), patient history tracking (Raghavan et al., 2012), etc. Our motivation for detecting negations in medical texts also stems from practical concerns of an operational medical question answering (QA) system (Ferrucci et al., 2013).

Most recent approaches to negation detection adopt supervised machine learning techniques to learn the phraseology of negation-containing expressions. They often follow a two step process—detection of negation *cues* ("no", "without", . . .), followed by detection of their associated *scopes*. Cue detection is a relatively simple task, since the set of cue words is not large. Determining the scope of a negation cue, on the other hand, is more challenging. Negation constructs do not necessarily apply to entire sentences: in the earlier example, *Mild hyperinflation* is not negated. The scope detection task is to identify the part(s) of the sentence that come under the scope of a negation cue. Scope detection is crucial for interpreting negations, and to that end, the BioScope corpus (Vincze et al., 2008) was released, with annotations of both negation cues *and* their associated scopes.

The fact that these scopes are represented only as text-spans is a drawback of BioScope. Without being anchored to a syntactic analysis of the sentences in which they occur, BioScope's scope annotations suffer from a variety of inconsistencies of mark-up. They also may, and occasionally do, fail to align with the underlying syntactic structures (Vincze et al., 2011; Stenetorp et al., 2012). Such inconsistencies make it hard for a system to learn the actual syntactic patterns connecting negation cues and their scopes—which are, after all, the real object of negation interpretation.

The insight that we develop in this paper is that a scope span can be associated with one or more nodes in the syntactic analysis of a negated expression, and that these will be further connected—in a systematic way—to the negation cue node. Mapping loosely and/or inconsistently bounded spans to unique syntactic nodes (and configurations thereof) reduces the noise inherent in BioScope. The learning task for scope detection would now be the easier one of learning negation scoping patterns from syntactic representations.

To elaborate on this, we look at BioScope's issues in some detail (Section 3.1). Our intent here, however, is not to offer a review or criticism of the corpus, nor to suggest how to correct those issues. Given that we *do want* to use BioScope (we motivate our choice of BioScope separately in Section 2), we propose a new method for learning how to detect negated constructs which are rooted in syntactic structure elements, and therefore directly usable by downstream components, many of which typically assume awareness of syntax. Our method is to re-map BioScope's scope span annotations onto the syntactic space and then to use those annotations' corresponding node structure(s) to train a system to automatically detect negated syntactic nodes. As outlined earlier, due to the re-mapping, many syntactic inconsistencies would not be seen by the learner, which now is trained on cleaner data and consequently, faces a simpler learning problem.

We verify that our re-mapping process identifies the correct negated syntactic node with high accuracy (93%); this validates the approach we propose here. Our supervised learning system, trained using re-mapped scope nodes to detect them automatically, obtains an overall accuracy of 87%, using automatically tagged cues. In the light of state-of-the-art performance figures, ours is a novel, constructive and pragmatic approach which allows us to leverage effectively an important resource, despite its representational imperfections, and to utilize the essential 'nuggets' it captures and exposes—namely the expressions of negated predicates. This strategy can also be applied to other tasks where syntactic context is important but resources are annotated by text spans only (e.g. hedge detection (Farkas et al., 2010)).

The rest of the paper is grounded in discussion of related work, and of BioScope and its annotations (Section 2), highlighting some relevant details of the issues with these (Section 3). We then outline the syntactic framework we use in Section 4. Section 5 presents our re-mapping of BioScope, and Section 6 offers experiments and results. In Section 7, we compare our performance with previously published studies. Section 8 concludes the paper.

2 Background

Early approaches in negation detection were limited in the nature of negation they were concerned with. The prime example here, NegEx (Chapman et al., 2001), took a view of negation interpretation to be "determining whether a finding or disease ... is present or absent". From such a standpoint, the notion of scope is limited, since the scope is always the finding or disease that follows a negation cue. While this works well for simpler expressions of negations, it tends to fail for more complex negation constructs. More recent approaches attempt to tackle the variability in scopes encountered in broader data by using statistical learning methods grounded in publicly available corpora with cue and scope annotations.

The first such corpus was BioScope (Vincze et al., 2008), which annotates negation cues and associated scopes in 3 genres—medical abstracts, scientific papers and clinical records. The BioNLP Event Extraction (EE) shared task corpus (Kim et al., 2009) also marks negation in the event annotations on sentences from molecular biology literature. Most recently, the *SEM 2012 shared task corpus (Morante and Blanco, 2012) marks negations, their foci, and scopes in sentences from Conan Doyle stories in an attempt to extend the research on negation to the general domain. Both the BioNLP-EE and *SEM corpora capture negations within—and therefore aligned with—syntactic analyses. Thus they deploy annotation schemes which assume downstream consumers of some granular negation representation, learnable from the annotated resource(s). However, the language in both of them differs greatly from the language encountered in clinical text, making them unsuitable for our QA system requirements. In contrast, BioScope matches our genre of clinical text. As an additional plus, it captures negation in a task-independent, linguistically motivated framework, which enables the building of systems applicable to a wider range of domains.

BioScope's negation-scope-as-span annotation framework, however, limits th corpus utility. Various approaches have used it to train negation scope span detection systems, and many have shown the importance of deep syntactic features in that task (e.g., (Ballesteros et al., 2012; Velldal et al., 2012; Zou et al., 2013)). They share a drawback: they are optimized for predicting the spans *as they are annotated* in BioScope—despite its various syntactic inconsistencies. For example, Ballesteros et al. (2012) use manual rules to detect the voice (passive or active) of a verb phrase; this is motivated by an annotation guideline for whether to include verb subjects in the span or not. In reality, what matters in the end is whether a detection system can capture the underlying phenomenon of negation that the annotations stand to represent, and not whether it can accurately replicate the representational choices the annotations follow. In light of this, our approach differs from the conventional ones, in that it mitigates the effects of inconsistencies in BioScope's original annotations by re-mapping it, as we explain in Section 5 below.

3 BioScope Corpus

The BioScope corpus (Vincze et al., 2008) is annotated for hedges and negations in sentences from biomedical domain; in this work, we use only the negation annotations. A negation (or hedge) annotation comprises a cue and a corresponding scope. The scope (hereafter BioScopeScopeSpan) is marked as a contiguous text-span including the associated cue annotation (BioScopeCue). BioScope contains sentences from three sub-genres—abstracts, full papers, and clinical records. We use all three sub-corpora. We divide each sub-corpus into 'Train' (70%), 'Dev' (15%) and 'Test' (15%) sets through random sampling. We use sentences in the Train and Dev sets to build and select best models and report the results obtained by our best models on Dev and Test sets.

3.1 Issues Challenging the Use of BioScope

BioScope is an important resource that has helped deeper understanding of various linguistic aspects of negation in a task independent manner. But, as we saw in the preceding sections, while demonstrating the importance of syntactic context for negation detection, recent efforts share the frustration arising from the fact that BioScopeScopeSpan annotations do not align with underlying syntactic structure. This problem is further exacerbated by inconsistencies in the corpus annotation. From a performance-driven point of view alone, negation detection systems trained over BioScope annotations are optimized to match the annotated spans in the corpus (as discussed in Section 2). However, for a negation detection system followed by downstream components implementing negation-driven inference, spans alone are not sufficient—especially spans which do not align with syntax. Negated expressions need to be captured within their syntactic context, and for this, we need the uniformity of syntax structures.

The misalignment issues of BioScopeScopeSpan annotations with respect to the underlying syntactic structures have already been extensively studied (Vincze et al., 2011; Stenetorp et al., 2012). Vincze et al. (2011) point out infelicities and mismatches, comparing BioScope annotations with the more syntactically oriented negated event annotations in the BioNLP-EE corpus (Kim et al., 2009). Inconsistencies are largely due to 'loose' annotation guidelines for BioScope, which are not rigorous enough in ensuring that annotation spans align with syntactic analyses. Given our position in this work—utilize BioScope, despite its shortcomings, in an alternative framework of analysis and training (see Section 2)—we explain some of the commonly occurring inconsistencies in this section. For this purpose, we use example annotations *e1-e5* from BioScope. (**Boldface** denotes BioScopeCue annotations and *italics* denotes corresponding BioScopeScopeSpan annotations as present in the BioScope corpus.)

One of the main source of inconsistencies within the syntactic space is with regard to the inclusion or exclusion of subjects of propositions. For example, in *e1*, the annotations identify the negation span to be the entire clause following the word *but*, including its subject and object. However, in *e2*, only the object of the predicate is marked as the negation scope (Figure 1). Vincze et al. (2011) state that "the treatment of subjects [in BioScope] remains problematic since in BioScope it is only the complements that are usually included within the scope of a keyword (that is, subjects are not with the exception of passive constructions and raising verbs)". Leaving aside the rationale for such a guideline, we note that such an inconsistency is harmful: proper interpretation of negated propositions does require a subject, and making annotations consistent by ignoring subjects, if present, does not help downstream components. Additionally, it makes the learning of contexts of negated propositions difficult.

e1: The cDNA hybridized to multiple transcripts in pre-B and B-cell lines, but *transcripts were* ***not*** *detected at significant levels in plasmacytoma, T-cell, and nonlymphoid cell lines.*

e2: Moreover, cAMP activators did ***not*** *activate NF-kappa B in Jurkat cells.*

Another problem with BioScopeScopeSpan annotations stems from the requirement that such annotations should have contiguous spans. For example, since sentence *e3* is a passive construction, the corresponding BioScopeScopeSpan annotation captures the subject (*mechanism*) as well. The contiguity requirement then forces the proposition *IFNs mediate this inhibition*—which modifies the subject but is itself not negated (Figure 2)—to be included within the BioScopeScopeSpan and therefore to be interpretable as negated. Clearly, there may be arbitrary intervening text in such, and similar, constructions, again making the learning task difficult.

e3: However, *the mechanism by which IFNs mediate this inhibition has* ***not*** *been defined.*

Sometimes, the BioScopeScopeSpan annotation boundaries do not align with syntactic constituents. For example, in *e4*, the BioScopeScopeSpan annotation excludes the determiner *the* from the scope while in *e5*, the determiner *the* is part of the scope. This might be due to the guideline that the scope should include the cue as well, causing to extend the scope annotation leftward until it covers the cue word (*absence*). Still, we are left with a span boundary which crosses, partially, a noun phrase boundary.

e4: Tal-1 transcription was shown to be monoallelic in Jurkat, a T-cell line that expresses tal-1 in the ***absence*** *of apparent genomic alteration of the locus.*

e5: The effects of selenium were specific for NF-kappa B, since *the activity of the transcription factor AP-1 was* ***not*** *suppressed.*

A system trained and optimized on how well it predicts the BioScopeScopeSpan boundaries suffers from also being forced to learn such syntactic inconsistencies along with the syntactic patterns that truly capture negation. In addition to learning the actual negation patterns, such a system is also forced to learn artifacts of annotation guidelines like: when to include or exclude subjects and when to include or exclude determiners. In order to circumvent this, we propose an approach in which we first re-map BioScope annotations onto nodes in the syntactic tree, and then train a system using features derived from the nodes, and node configurations, providing the context for the negation cue and scope nodes. We next describe the syntactic framework we use and then explain our approach in detail.

4 Syntactic Framework

Negation, as a language device, is naturally conceptualized as applying to fully instantiated predicate-argument clusters. We therefore use predicate argument graphs as structural abstractions of syntax trees. Additional advantages of these abstractions include their affinity for having extra-propositional

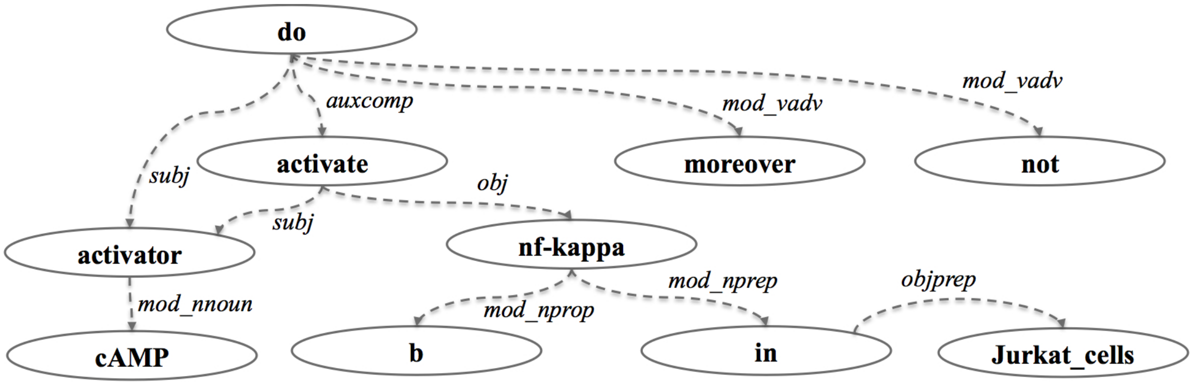

Figure 1: PAS for *e2*: "Moreover, cAMP activators did not activate NF-kappa B in Jurkat cells"

aspects of meaning 'layered' onto the representation (precedents in prior studies can be found in e.g. (Saurí and Pustejovsky, 2009; Diab et al., 2009)), and their pervasive use in a state-of-the-art QA system—for question analysis, candidate generation, and analysis of passage evidence (Ferrucci et al., 2010; Ferrucci, 2012)—which is at the heart of our medical adaptation (Ferrucci et al., 2013).

We use predicate-argument structure (PAS) (McCord et al., 2012) derived from dependency parses produced by the English Slot Grammar parser (McCord, 1990). In addition to normalizing across different tree structures expressing essentially the same meaning, PAS provides a simplified view over 'raw' syntactic trees, gathering all arguments to a predicate from local, and distant, parse tree nodes (see (McCord et al., 2012) for details). Figures 1 and 2 show the PASes for examples *e2* and *e3*. By localizing the logical arguments to a proposition, predicate-based representation provides direct access to *all* arguments of e.g. a verb frame: an important requirement for extracting context-denoting syntactic features.

PAS-based view into sentences offers unambiguously uniform treatment of some of the issues highlighted in the previous section. For example, going back to *e2*, and the rationale for including or excluding subjects in the scope of a negation, we observe that verb nodes in the PAS always have fully instantiated frames, with subject arguments bound to the

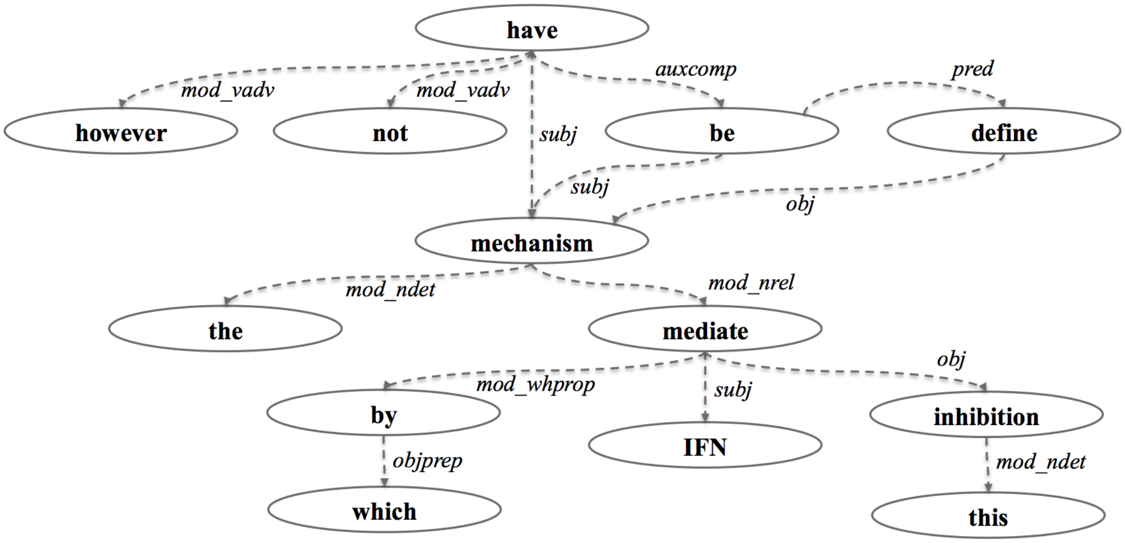

Figure 2: PAS for *e3*: "However, the mechanism by which IFNs mediate this inhibition has not been defined."

predicate nodes corresponding to the deep syntactic subject: observe how *activator* is 'subj' both to *do* and *activate*. Thus whether to include a subject into a verb scope (e.g. *not activate*) becomes largely irrelevant, and a PAS-based scope rendering can always include subjects. As another example, for the PAS for *e3*, the granular analysis of the arguments to the predicate for *define* can be leveraged to designate the predicate node for *mechanism* as the scope of the negation *not (defined)*, while excluding the *IFN mediate inhibition* subtree from the same scope.

5 Learning Negations from Re-mapped BioScope

Our goal is a system for automatic identification of negations and their scopes within the PAS of a sentence. Our resource for this is BioScope, with its text-based span annotations. We propose a novel approach, realized as a two-step process:

(1) **BioScope-to-PAS mapping**: map BioScope's text-span cue and scope annotations to PAS nodes (CuePredicate and NegatedPredicate) by identifying the predicate nodes in the PAS of the sentence that best capture the annotations.

(2) **NegatedPredicate learning**: train a statistical model to automatically identify the scope predicate using features from the PAS context of cue and scope predicates.

5.1 BioScopeScopeSpan-to-NegatedPredicate Mapping

Having obtained PASes for sentences in the corpus, we mark the PAS node with the minimal span that contains the entire BioScopeScopeSpan annotation as the NegatedPredicate. We define the 'span' of a PAS node to be the span of text covered by the subtree rooted at that node, which includes the spans of all of its descendants. Similarly, we mark the PAS node with the minimal span that contains the BioScopeCue annotation as the CuePredicate.

For example, in Figure 1, the predicate labeled *not* was marked as the CuePredicate and the predicate labeled *do* was marked as the corresponding NegatedPredicate. In order to perform a sanity check on our re-mapping, we judged whether the predicate nodes that we mark as NegatedPredicate in sentences from our Dev set are in fact the ones being negated. Of the 470 sentences containing negations, 13 (2.8%) failed to parse, breaking the mapping. In other words, our mapping strategy has coverage of about 97.2%. Of the sentences where a NegatedPredicate was obtained, our mapping achieved an accuracy of 92.8% in finding the correct negated predicate.

5.2 NegatedPredicate Learning

We now build a supervised learning system which, given a CuePredicate in a sentence, will identify its corresponding NegatedPredicate. For every predicate p in a sentence PAS with a CuePredicate, we create an instance <CuePredicate, p>. The instance <CuePredicate, p> is assigned *true* if p is the corresponding NegatedPredicate. For all other p in the PAS, <CuePredicate, p> is assigned *false*.

We extract three types of features for each instance <CuePredicate, p>: 1) *token features* (word lemma and POS tag) of CuePredicate and p, 2) *syntactic context features* (token features of parent predicates and all argument predicates) of CuePredicate and p, and 3) *predicate pair features* (is CuePredicate argument of p or vice versa?; distance between CuePredicate and p; relative position of CuePredicate and p).

We use the ClearTk (Ogren et al., 2008) framework to build our system and perform experiments. We use quadratic kernel SVMs in all our experiments. The ClearTK wrapper for SVM-Light (Joachims, 2006) internally shifts the prediction threshold using sigmoid fitting to deal with the highly skewed class imbalance (around 5% of positive cases) in our data. Prior studies (Prabhakaran et al., 2012b) have shown this approach to be effective in addressing the class imbalance problem.

During prediction, given an unseen sentence PAS and a CuePredicate (either GOLD or automatically predicted) in it, we need to find the corresponding NegatedPredicate. We iterate over all candidate predicates c in the sentence PAS and apply our trained model to assign a *true* or *false* value to <CuePredicate, c>. For any CuePredicate in a sentence there must be one and only one NegatedPredicate, since BioScope corpus marks a single BioScopeScopeSpan for every BioScopeCue. We choose the c for which <CuePredicate, c> is assigned a *true* value with the highest confidence as

	On Dev			On Test		
	Precision	Recall	F-measure	Precision	Recall	F-measure
Clinical	95.68	95.68	95.68	96.15	96.9	96.53
Abstracts	94.4	94.4	94.4	95.42	96.9	96.15
Papers	79.22	96.83	87.14	85.29	98.31	91.34
Overall	92.36	95.11	93.71	94.13	97.09	95.58

Table 1: Performance of our CuePredicate detection on Dev and Test sets

the NegatedPredicate. If <CuePredicate, c> is assigned a *false* value for all c, we choose the c with the least confident *false* assignment as the NegatedPredicate.

6 Experiments and Results

The most commonly used metric to evaluate negation scope span detection is Percentage of Correct Scopes (PCS). PCS measures the percentage of exact matches between predicted and actual scope spans. Since our task is different—negated predicate detection as opposed to negated span detection—we report the Percentage of Correct Scope Predicates (PCSP) obtained in our experiments. Models built from the composite training corpus comprising training corpora of all three genres (see Section 3) perform better than models built separately over each sub-corpus. We report results separately for each sub-corpus, as well as for the entire corpus, and compare them with a strong baseline.

6.1 Gold vs. Predicted CuePredicates

We report results for the NegatedPredicate detection task obtained using GOLD CuePredicates as well as predicted CuePredicates. In order to measure the performance on predicted CuePredicates, we built a CuePredicate detector using linear kernel SVM to detect whether a predicate is a negation cue or not. We use three types of features: 1) *token features* (lemma and POS) of the predicate, 2) *linear context* (token features of the token after the predicate in the sentence; features of tokens before the predicate turned out to be not useful), and 3) *syntactic context* (token features of parent and argument predicates). As shown in Table 1, our CuePredicate tagger obtained F-measures in the range of state-of-the-art results on negation cue detection using the BioScope (90-96% F-measure (Velldal et al., 2012)).

6.2 Baseline NegatedPredicate Predictor

Since this formulation of the task is new, we built a strong baseline system appropriate for it. In our baseline, we predict the NegatedPredicate to be the parent predicate of the CuePredicate, if the CuePredicate is a terminal node in the PAS (this will cover the most common cues such as *no* and *not*). If the CuePredicate is not a terminal node (which covers the cases of verbal negation cues such as *failed*), we choose the CuePredicate itself as the NegatedPredicate. Columns 1 and 3 of Table 2 show PCSP obtained by the baseline algorithm on our Dev and Test sets respectively using GOLD CuePredicates. Columns 5 and 7 show corresponding results using predicted CuePredicates.

6.3 Our NegatedPredicate Predictor

The results obtained by our NegatedPredicate detection system (Section 5.2) on Dev and Test sets using GOLD CuePredicates is shown in Columns 2 and 4 of Table 2. Our system outperforms the baseline by a large margin in all cases, with especially high performance in clinical records. We obtain an overall PCSP of 90.2% and 89.2% on Dev and Test sets respectively. The results we obtain in Test set are in the range of what we obtain using Dev set, which shows that our system does not overfit to our Dev set. On applying our system on predicted CuePredicates, the overall results (columns 6 and 8) decrease by around 3-5% from using GOLD CuePredicates. The overall PCSP value of 86.8% obtained on the Test set reflects the accuracy of our end-to-end system on a blind test. Note that this is a conservative estimate since we penalize our system for failed parses where the mapping step could not find a GOLD NegatedPredicate to compare against.

	Gold Cues (On Dev)		Gold Cues (On Test)		Predicted Cues (On Dev)		Predicted Cues (On Test)	
	Baseline	System	Baseline	System	Baseline	System	Baseline	System
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Clinical	83.45	97.12	88.37	100.00	82.01	93.53	87.60	96.90
Abstracts	81.34	89.18	79.07	84.50	76.49	83.96	77.52	82.56
Papers	73.02	79.37	81.36	86.44	66.67	71.43	77.97	83.05
Overall	80.85	90.21	82.06	89.23	76.81	85.11	80.49	86.77

Table 2: Percentage of Correct NegatedPredicate (PCSP) on Dev and Test sets

7 Comparison with Previous Approaches

Comparing our system with previously published approaches to negation scope detection is not straightforward, essentially because our and their tasks are different: negated predicate detection vs. negated scope span detection. The resulting difference in evaluation metrics makes PCS numbers reported elsewhere not directly comparable with our PCSP results presented in Table 2. To make such a comparison meaningful, we transform (reverse map) the NegatedPredicates we identify back into text spans and use those to derive PCS values better aligned with previously published ones. (Note that these PCS numbers are still not directly comparable, due to differences in experiment setup, e.g. cross validation vs. held out test set.

Transforming the NegatedPredicates back to BioScopeScopeSpan annotations is not trivial. As discussed in Section 5.1, we choose NegatedPredicate to be the predicate node that minimally covers BioScopeScopeSpan. Hence, the span of a NegatedPredicate may include text spans that were originally not part of the corresponding BioScopeScopeSpan annotation. Therefore, we built a statistically trained system to predict whether the span of a descendant node of a NegatedPredicate should, or should not, be included in reverse mapping that NegatedPredicate to the corresponding BioScopeScopeSpan.

We use the same set of features and learning configuration as we used for NegatedPredicate learning (Section 5.2). Our transformation obtained high accuracy (94.9%) for the clinical records. However, it was a harder task for abstracts (66.1%) and papers (73.1%) which contain more complex sentences.

We applied this transform on the predicate nodes identified by our end-to-end system (Section 6.3) in order to derive PCS values. In Table 3, we compare these PCS values against four previous studies above (due to lack of space, we do not discuss their techniques here), as well as with a baseline of our own where we use the covered text of the predicate node and all of its descendants as scopes used in the comparison. Our system (with transform) obtains higher PCS values than all other reported studies on the clinical records. The PCS values obtained for the abstracts and papers sub-corpora are lower, but still in comparable range to the other studies. It is important to note that the main source of error here is the NegatedPredicate-to-BioScopeScopeSpan trans-

	Morante09	Ballestros12	Velldal12	Ours (With Covered Text)		Ours (With Transform)		Zou13
				On Dev	On Test	On Dev	On Test	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Clinical	70.75	89.06	89.41	88.49	89.92	91.37	**92.25**	85.31
Abstracts	66.07	68.92	**72.89**	35.45	35.27	61.94	58.53	76.90
Papers	41.00	61.43	**68.09**	33.33	23.73	53.97	47.46	61.19
Overall	-	-	-	50.85	49.55	69.57	66.82	-

Table 3: PCS measures from previous BioScope span detection approaches and our end-to-end system.
Col. 1-3: end-to-end systems (Morante and Daelemans, 2009), (Ballesteros et al., 2012), and (Velldal et al., 2012);
Col. 4-7: our end-to-end system with different ways of obtaining the spans in our Dev and Test sets;
Col. 8: (Zou et al., 2013) system using GOLD cues (often 5-10% higher than using predicted cues)

form step, with its inherent lower accuracies for these two corpora, as reported above. We emphasize that for practical applications this transformation is of little use: what matters more, certainly for a negation detection system feeding downstream components, are the PCSP values presented in Section 6.

8 Discussion and Conclusion

The results for our system, with reverse mapping, offer indirect evidence for our observation in Section 3.1: training a system to predict BioScopeScopeSpan boundaries would require it also to learn inconsistencies in BioScope annotations. This is a hard learning task, given the noise discussed in Section 3.1. Indeed, our results for learning the reverse-mapping transformation show that it is harder to learn the specific annotation criteria in BioScope than to learn the structural patterns expressing negations (which, as we saw in Section 6, obtained close to 90% accuracy). While we had to build a system to transform nodes back to spans for the purposes of comparative analysis, such a system has no role in our quest for practical negation detection and representation.

This substantiates our strategy of using BioScope, *as is*, to learn not scope spans of negation expressions, but negated predicates within the predicate-argument structure (Section 5). The re-mapping route takes us where we want to be, from the point of view of a practical application of negation-based inference: with access to negated predicate nodes. The end-to-end accuracy (overall, across three different genres) of 87% on blind test validates the creative way we propose to make use of a valuable and unique resource—despite its imperfections—by extracting the real value in it, while mitigating the effects of its various inconsistencies.

The results in Tables 2 and 3 show that we have achieved our primary objective: using BioScope to train a system which detects structured negation expressions in clinical text. Our approach to negation scope learning in the syntactic space is a two-step one—first, re-mapping the text-span annotations for negation scopes in BioScope to the syntactic space and then training a scope predicate predictor. We show that our transformation introduces only a small percentage of error and also that our predicted nodes can be transformed back to original span annotations with performance comparable to other negation scope span prediction systems trained on the same dataset. Notably, in clinical records, our system outperforms reported state-of-the-art results (column 8 of Table 3).

In a broader context, the work we report here indirectly argues that the method we propose to circumvent certain limitations of a corpus like BioScope can be applied to similar tasks (such as hedging, sentiment analysis, and variety of modalities, cf. Section 1), for which current annotation resources offer flat, and possibly inconsistent, annotations. In addition, we chose PAS as our syntactic framework for the reasons listed in Section 4, but our approach is not limited to PAS. Indeed, the claims, and methods, are presented to be applicable, and workable, in a more general syntactic framework.

Acknowledgments

This work was conducted while the first author was doing an internship at IBM Research at the TJ Watson Research Center in Yorktown Heights, NY, USA. We thank several anonymous reviewers for their constructive feedback.

References

Mordechai Averbuch, Tom Karson, Benjamin Ben-Ami, Oded Maimon, and Lior Rokach. 2004. Context-sensitive medical information retrieval. In *Proc. of the 11th World Congress on Medical Informatics (MEDINFO-2004)*, pages 1–8. Citeseer.

Miguel Ballesteros, Virginia Francisco, Alberto Díaz, Jesús Herrera, and Pablo Gervás. 2012. Inferring the scope of negation in biomedical documents. In *13th International Conference on Intelligent Text Processing and Computational Linguistics (CICLING 2012)*, New Delhi, 2012. Springer, Springer.

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):301–310.

Md. Faisal Mahbub Chowdhury and Alberto Lavelli. 2013. Exploiting the scope of negations and heterogeneous features for relation extraction: A case study for drug-drug interaction extraction. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics:*

Human Language Technologies, pages 765–771, Atlanta, Georgia, June. Association for Computational Linguistics.

Isaac G Councill, Ryan McDonald, and Leonid Velikovich. 2010. What's great and what's not: learning to classify the scope of negation for improved sentiment analysis. In *Proceedings of the workshop on negation and speculation in natural language processing*, pages 51–59. Association for Computational Linguistics.

Dina Demner-Fushman, Wendy W Chapman, and Clement J McDonald. 2009. What can natural language processing do for clinical decision support? *Journal of biomedical informatics*, 42(5):760–772.

Mona Diab, Lori Levin, Teruko Mitamura, Owen Rambow, Vinodkumar Prabhakaran, and Weiwei Guo. 2009. Committed belief annotation and tagging. In *Proceedings of the Third Linguistic Annotation Workshop*, pages 68–73, Suntec, Singapore, August. Association for Computational Linguistics.

Richárd Farkas, Veronika Vincze, György Szarvas, György Móra, and János Csirik, editors. 2010. *Proceedings of the Fourteenth Conference on Computational Natural Language Learning*. Association for Computational Linguistics, Uppsala, Sweden, July.

David Ferrucci, Eric Brown, Jennifer Chu-Carroll, James Fan, David Gondek, Aditya A Kalyanpur, Adam Lally, J William Murdock, Eric Nyberg, John Prager, et al. 2010. Building watson: An overview of the deepqa project. *AI magazine*, 31(3):59–79.

David A Ferrucci, Anthony Levas, Sugato Bagchi, David Gondek, and Erik T Mueller. 2013. Watson: Beyond jeopardy! *Artif. Intell.*, 199:93–105.

David A Ferrucci. 2012. Introduction to "this is watson". *IBM Journal of Research and Development*, 56(3.4):1:1–1:15.

Thorsten Joachims. 2006. Training Linear SVMs in Linear Time. In *Proceedings of the 12th ACM SIGKDD international conference on knowledge discovery and data mining*, pages 217–226. ACM.

Jin-Dong Kim, Tomoko Ohta, Sampo Pyysalo, Yoshinobu Kano, and Jun'ichi Tsujii. 2009. Overview of bionlp'09 shared task on event extraction. In *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing: Shared Task*, BioNLP '09, pages 1–9, Stroudsburg, PA, USA. Association for Computational Linguistics.

Minsuk Lee, James Cimino, Hai Ran Zhu, Carl Sable, Vijay Shanker, John Ely, and Hong Yu. 2006. Beyond information retrievalmedical question answering. In *AMIA Annual Symposium Proceedings*, volume 2006, page 469. American Medical Informatics Association.

Michael C. McCord, J. William Murdock, and Branimir Boguraev. 2012. Deep parsing in Watson. *IBM Journal of Research and Development*, 56(3):3.

Michael C. McCord. 1990. Slot grammar: A system for simpler construction of practical natural language grammars. In R. Studer, editor, *Natural Language and Logic: Proc. of the International Scientific Symposium, Hamburg, FRG*, pages 118–145. Springer, Berlin, Heidelberg.

Stéphane M Meystre, Guergana K Savova, Karin C Kipper-Schuler, and John F Hurdle. 2008. Extracting information from textual documents in the electronic health record: a review of recent research. *Yearbook of medical informatics*, pages 128–144.

Roser Morante and Eduardo Blanco. 2012. * SEM 2012 shared task: Resolving the scope and focus of negation. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation*, pages 265–274. Association for Computational Linguistics.

Roser Morante and Walter Daelemans. 2009. A metalearning approach to processing the scope of negation. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning*, pages 21–29. Association for Computational Linguistics.

Philip V. Ogren, Philipp G. Wetzler, and Steven Bethard. 2008. ClearTK: A UIMA toolkit for statistical natural language processing. In *Towards Enhanced Interoperability for Large HLT Systems: UIMA for NLP workshop at Language Resources and Evaluation Conference (LREC)*.

Vinodkumar Prabhakaran, Owen Rambow, and Mona Diab. 2010. Automatic committed belief tagging. In *COLING 2010: Posters*, pages 1014–1022, Beijing, China, August. COLING 2010 Organizing Committee.

Vinodkumar Prabhakaran, Michael Bloodgood, Mona Diab, Bonnie Dorr, Lori Levin, Christine D. Piatko, Owen Rambow, and Benjamin Van Durme. 2012a. Statistical Modality Tagging from Rule-based Annotations and Crowdsourcing. In *Proceedings of the Workshop on Extra-Propositional Aspects of Meaning in Computational Linguistics*, pages 57–64, Jeju, Republic of Korea, July. Association for Computational Linguistics.

Vinodkumar Prabhakaran, Owen Rambow, and Mona Diab. 2012b. Predicting Overt Display of Power in Written Dialogs. In *Human Language Technologies: The 2012 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, Montreal, Canada, June. Association for Computational Linguistics.

Preethi Raghavan, Albert Lai, and Eric Fosler-Lussier. 2012. Learning to temporally order medical events in clinical text. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 70–74, Jeju Island, Korea, July. Association for Computational Linguistics.

Roser Saurí and James Pustejovsky. 2009. Factbank: a corpus annotated with event factuality. *Language Resources and Evaluation*, 43:227–268. 10.1007/s10579-009-9089-9.

Pontus Stenetorp, Sampo Pyysalo, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. 2012. Bridging the gap between scope-based and event-based negation/speculation annotations: a bridge not too far. In *Proceedings of the Workshop on Extra-Propositional Aspects of Meaning in Computational Linguistics*, pages 47–56. Association for Computational Linguistics.

Kateryna Tymoshenko, Swapna Somasundaran, Vinodkumar Prabhakaran, and Vinay Shet. 2012. Relation mining in the biomedical domain using entity-level semantics. In *ECAI*, pages 780–785.

Özlem Uzuner, Imre Solti, and Eithon Cadag. 2010. Extracting medication information from clinical text. *Journal of the American Medical Informatics Association*, 17(5):514–518.

Erik Velldal, Lilja Øvrelid, Jonathon Read, and Stephan Oepen. 2012. Speculation and negation: Rules, rankers, and the role of syntax. *Computational linguistics*, 38(2):369–410.

Veronika Vincze, György Szarvas, Richárd Farkas, György Móra, and János Csirik. 2008. The BioScope corpus: biomedical texts annotated for uncertainty, negation and their scopes. *BMC bioinformatics*, 9(Suppl 11):S9.

Veronika Vincze, Gyorgy Szarvas, Gyorgy Mora, Tomoko Ohta, Richárd Farkas, et al. 2011. Linguistic scope-based and biological event-based speculation and negation annotations in the BioScope and Genia Event corpora. *Journal of Biomedical Semantics*, 2(Suppl 5):S8.

Michael Wiegand, Alexandra Balahur, Benjamin Roth, Dietrich Klakow, and Andrés Montoyo. 2010. A survey on the role of negation in sentiment analysis. In *Proceedings of the workshop on negation and speculation in natural language processing*, pages 60–68. Association for Computational Linguistics.

Bowei Zou, Guodong Zhou, and Qiaoming Zhu. 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, Seattle, Washington, USA, October. Association for Computational Linguistics.