NEGATION’S NOT SOLVED: GENERALIZABILITY VERSUS OPTIMIZABILITY IN CLINICAL NATURAL LANGUAGE PROCESSING

Authors	Stephen Wu[a] Timothy Miller[b] James Masanz[a] Matt Coarr[c] Scott Halgrim[d] David Carrell[d] Cheryl Clark[c]
Affiliation	[a]Department of Health Sciences Research Mayo Clinic 200 First Street SW Rochester, MN 55905, USA
	[b] Children's Hospital Boston Informatics Program Harvard Medical School 300 Longwood Ave Boston, MA 02115, USA
	[c] The MITRE Corporation 202 Burlington Road Bedford, MA 01730, USA
	[d] Group Health Research Institute 1730 Minor Ave, Suite 1600 Seattle, WA 98101, USA
Corresponding author	Stephen Wu Mayo Clinic Department of Health Sciences Research 200 First Street SW Rochester, MN 55905, USA Email: wu.stephen@mayo.edu, stw4@alumni.duke.edu Phone: +1(507) 538-0167 Fax: +1(507) 284-0360

Word Count
3,936 words

ABSTRACT

A review of published work in clinical natural language processing (NLP) may suggest that the negation detection task has been "solved." This work contends that an *optimizable* solution does not equal a *generalizable* solution. Using four manually annotated corpora of clinical text, we show that negation detection can be optimized in relatively constrained settings, but performance is not reliably generalizable unless in-domain training data is available – in which case fully-supervised domain adaptation techniques may prove effective. Various factors (e.g., annotation guidelines, named entity characteristics, the amount of data, and lexical and syntactic context) play a role in making generalizability difficult, but none completely explains the phenomenon. This indicates the need for future work in domain-adaptive and task-adaptive methods for clinical NLP.

Keywords
Natural language processing
Information extraction
Electronic medical records
Negation detection
Domain adaptation

1 Introduction

Negation in unstructured clinical text is a well-known phenomenon. It is crucial for any practical interpretation of clinical text, since negation is common in clinical narrative. For example, the medical significance of “no wheezing” is quite different from that of “wheezing.” With the increasingly widespread use of electronic medical records (EMRs), computational methodologies for negation detection have also become well-known, most notably the early and strikingly straightforward NegEx algorithm [1]. In NegEx, simple regular expressions yield solid performance on detecting the negation of Findings, Diseases, and Mental or Behavioral Dysfunctions from the Unified Medical Language System (UMLS). The success of NegEx (and other techniques) is attributable to the constrained pragmatics of clinical text: because physicians are writing the text in order to convey the health status of a patient, there is a limit to the ways that medically pertinent concepts can be negated. Since existing algorithms have performed well in many published studies [2-8], many clinical natural language processing (NLP) practitioners consider negation detection a solved problem (see Table 1) with a simple, generalizable solution.

However, our present work will show that this “solved” designation is premature because current solutions are easily *optimizable* but not necessarily *generalizable*. Negation detection is still a challenge when considered from a practical, multi-corpus perspective, i.e., one in which an algorithm is deployed in many clinical institutions and on many sources of text. As the NLP Attribute Discovery team for the Strategic Health IT Advanced Research Project on the Secondary use of the EHR (SHARPn), we attempted to detect negation in four corpora, using machine learning, rules, domain adaptation, and various evaluation scenarios. These corpora include the new SHARPn NLP Seed Corpus of clinical text with multiple layers of syntactic and semantic information, including named entities (NEs) and polarity (i.e., negation). We also used the 2010 i2b2/VA NLP Challenge corpus, the MiPACQ corpus, and the NegEx Test Set. The SHARPn Attribute Discovery negation detection system used in our evaluation is currently available in Apache cTAKES (clinical Text Analysis and Knowledge Extraction System; ctakes.apache.org) as part of the ctakes-assertion project, including an integrated domain adaptation algorithm [9]. A thorough methodological treatment is described in a forthcoming publication.

We conclude that practical negation detection is not reliable without in-domain training data and/or development. “Benchmark” gold standard data sets differed sufficiently to have a profound effect on the viability of negation detection algorithms. Furthermore, it is difficult to determine an optimal mix of training data, or to standardize a definitive “benchmark” metric, since both are influenced by corpus-specific annotation guidelines and data sources. The results we report here should remind users of negation detection algorithms to be vigilant in tuning systems to their data, whether by training with local data or modifying rules. We also call for future work in domain-adaptive and task-adaptive methods.

After a discussion of the extensive related work in negation detection, the remainder of this article will introduce the data and methods for corpus and system comparisons of negation detection, present the resulting performance of systems on the different corpora, and discuss implications for negation detection and annotation schema in the larger picture of clinical informatics.

2 Related Work

Negation detection was a very practical early motivation for NLP adoption among the informatics community, and thus significant effort has gone into this task. While there have been many systems implementing negation detection, publicly available corpora for testing them are limited by patient privacy concerns, as is typical in clinical NLP.

Negation detection systems have shown excellent performance in clinical text, beginning with the rule-based NegEx algorithm [1]. NegEx was originally evaluated on spans of text that matched UMLS Findings, Diseases, and Mental or Behavioral Dysfunctions among 1000 test sentences sampled from discharge summaries at the University of Pittsburgh Medical Center; a regression test set was released later with de-identified notes of 6 different types. NegEx has produced numerous updated and customized systems, including the updated version released with ConText[10] which performed well on a benchmark NegEx Test Set (available at https://code.google.com/p/negex/wiki/TestSet). Our tests used the ytex version[11] of NegEx as a baseline and included the NegEx Test Set as a benchmark.

Similar to NegEx, many other negation algorithms take a rule-based approach, with a variety of techniques: lexical scan with context free grammar [2], negation ontology [3], or dependency parse rules [4]. Some negation algorithms treat the problem as a machine learning classification task[5] or as some hybrid between rules and machine learning [6 7]. The performance of these systems and their data sources is summarized in Table 1 below.

Table 1: Extensive successful previous work on negation detection in clinical text

Algorithm	**Data source**	**Prec.**	**Rec.**	**F1**
Negfinder [2]	10 surgery notes & discharge summaries; UMLS concepts	91.84	95.74	92.96
NegEx[1]	UPMC ICU discharge summaries; clinical conditions	84.49	77.84	80.35
Neg assignment grammar[3]	Hopkins HNP notes; SNOMED concepts	91.17	**97.19**	93.90
Negation Detection Module[7]	Stanford radiology reports; unmapped text phrases	**98.63**	92.58	**94.91**
ConText[10]	UPMC 6 note types; clinical conditions	92	94	93
MITRE assertion[6]	2010 i2b2/VA; unmapped “problem” text phrases	92	95	94
DepNeg[4]	Mayo clinical notes; symptoms & diseases	96.65	73.93	83.78

All these general approaches were represented in the 2010 i2b2/VA NLP Challenge task on assertions [8]. In addition to catalyzing innovation from multiple systems, this shared task produced a benchmark data set that is available for research with a simple data use agreement; it interprets negation on medical problem NEs as an assertion that the problem is absent.

The four corpora used in our study all annotate *named entities* explicitly (though they differ on whether they are mapped to an ontology), but only include the *scope of negation indicators* implicitly (through the pertinent NEs). Some efforts have reversed this, giving an implicit notion of named entities but an explicit notion of negation scope: notably the BioScope Corpus[12] that was used as part of the CoNLL 2010 Shared Task [13]. Bioscope annotates negation, uncertainty, and their scopes on de-identified clinical free text (1,954 radiology reports), biological full articles (9 articles from FlyBase and BMC Bioinformatics), and scientific abstracts (1,273 abstracts also in the GENIA corpus). Here, the scope of negation is specified as the maximum span within which the negation cue word could be applicable, and the scope cannot be disjoint from the cue word. This is in contrast to the work we present here, which focuses on named entities. We ignore scope for two reasons: First, the lack of gold standard named entity mentions is an additional source of error that no other corpus would have, making the comparison unfair. Second, while negation scope annotations overcome some recall issues for non-standard terminology (e.g., “patient is not feeling as much like a pariah today” would represent negation correctly despite finding no NE), they do not overcome issues in fine-grained annotation guideline distinctions (see Section 3.2 on Annotation Guidelines).

3 Methods

Here, we first describe the annotated NLP corpora used in training and testing, with salient information about the gold standard entity and negation annotation guidelines. We then briefly discuss the new SHARPn Polarity Module and a rule-based baseline.

3.1 NLP corpora with negation annotations

Our work used four clinical NLP annotation efforts; the SHARPn NLP Seed Corpus, the 2010 i2b2/VA NLP Challenge Corpus; the MiPACQ corpus; and the NegEx Test Set. Statistics in Table 2 show their overall relative sizes, train/test splits, and proportion of negated concepts.

Table 2: Characteristics of four corpora with negation annotations

	sharp		i2b2		mipacq		negex
	Train	Test	Train	Test	Train	Test	Test
Num. Documents	140	22	477	349	2,443	324	120
Num. Sentences	5,014	569	48,482°	33,022°	19,672	2,236	2,376*
Num. Named Entities	10,575	1,154	18,550	11,968	23,249	1,721	2,371
Num. Negated NEs	918	48	3,609	2,535	1,681	158	491
% Negated NEs	8.7%	4.2%	19.5%	21.2%	7.2%	9.2%	20.7%
Data Source(s)	Mayo, Group Health		Partners, BIDMC, UPMC		Medpedia, NLM ClinQ, Mayo		UPMC

*subset selected manually; °automatic sentence detection on pre-whitespace-tokenized text

First, the SHARPn NLP Seed Corpus consists of de-identified radiology notes related to Peripheral Arterial Disease (PAD) from Mayo Clinic, and de-identified breast oncology progress notes regarding incident breast cancer patients from Group Health Cooperative. This multi-layered annotated corpus follows community adopted standards and conventions for the majority of annotation layers, which include syntactic trees, predicate-argument structure, coreference, UMLS named entities, UMLS relations, and Clinical Element Models (CEM) templates [14]. Negation is included in the CEM templates as an attribute of UMLS concepts.

Second, the 2010 i2b2/VA NLP Challenge Corpus contained a total of 477 manually annotated, de-identified reports from Partners Healthcare, Beth Israel Deaconess Medical Center, and the University of Pittsburgh Medical Center. The majority of notes were discharge summaries, but the University of Pittsburgh Medical Center also contributed progress notes.

Third, the MiPACQ corpus[15 16] annotates multiple syntactic and semantic layers, similar to the SHARPn NLP corpus. There are three major divisions to the sources of data: a snapshot of *Medpedia articles* on medical topics, written by clinicians, retrieved on April 26, 2010; *clinical questions* from the National Library of Medicine’s Clinical Questions corpus (http://clinques.nlm.nih.gov), collected by interviews with physicians; and sentences from Mayo Clinic clinical notes and pathology notes related to colon cancer.

Finally, the NegEx Test Set is a set of manually-selected sentences from 120 de-identified University of Pittsburgh Medical Center reports (20 each of radiology, emergency department, surgical pathology, echocardiogram, operative procedures, and discharge summaries). This set was used to evaluate the ConText algorithm [10], while another 120 reports of similar distribution (not publically available) were used for the development of the negation portion of ConText (i.e., an updated NegEx).

Table 3: Named entities (NEs) and negated NEs in the MiPACQ and SHARP corpora

	mipacq				sharp			
	Train	Test	Train	Test	Train	Test	Train	Test
	%type#type	%type#type	%neg#neg	%neg#neg	%type#type	%type#type	%neg#neg	%neg#neg
AnatomicalSite	20.36%(4591)	25.24%(428)	4.01%(184)	7.48%(32)	39.87%(4216)	50.69%(585)	0.43%(18)	0%()
DiseaseDisorder	26.53%(5981)	23.29%(395)	7.82%(468)	11.65%(46)	27.54%(2912)	29.29%(338)	17.07%(497)	13.31%(45)
Lab	-	-	-	-	1.91%(202)	0.69%(8)	2.97%(6)	25.00%(2)
Medication	14.74%(3324)	13.50%(229)	4.60%(153)	8.30%(19)	2.98%(315)	-	6.35%(20)	0%()
Procedure	19.62%(4424)	22.52%(382)	3.28%(145)	1.05%(4)	16.64%(1759)	11.01%(127)	3.01%(53)	0%()
SignSymptom	16.28%(3671)	12.68%(215)	19.83%(728)	26.51%(57)	5.70%(603)	2.17%(25)	52.57%(317)	4.00%(1)
Entity	0.35%(79)	0.06%(1)	1.27%(1)	0%()	2.96%(313)	3.73%(43)	0.64%(2)	0%()
Event	2.10%(474)	2.71%(46)	0.42%(2)	0%()	2.40%(254)	2.43%(28)	4.42%(5)	0%()

3.2 Comparison of annotation guidelines

Manually annotated negation in one of these corpora is not strictly equivalent to that in other corpora. We cannot directly compare annotation guidelines because we do not have corpora that are *multiply-annotated* with different guidelines. However, we should note that all annotation projects reported high inter-annotator agreement within their respective projects. Here, we qualitatively analyze the annotation guidelines concerning the annotation of both NEs (concepts) and attributes (assertion status), hypothesizing that some differences in annotation guidelines may negatively affect the performance of negation algorithms across corpora.

The primary difference between the annotation guidelines of the corpora appears to be in the definition of NEs, rather than direct indications of how negation should be handled. First, NE annotation guidelines differ in the *semantic types that are allowed*. The broadest is the MiPACQ corpus, which annotates 17 UMLS Semantic Groups. (However, in practice, some semantic groups have zero or negligible frequencies, and we have grouped them together in our analysis.) SHARP only annotates the 6 most clinically relevant groups, namely, Diseases and Disorders, Signs and Symptoms, Labs, Medications, Procedures, and Anatomical Sites. The NegEx Test Set is much more narrow, including only Signs, Symptoms, Diseases, and Findings with qualitative values. The i2b2 corpus is similarly restrictive, only annotating "problems," i.e., Diseases, Signs and Symptoms.

The corpora also differ in the *span to consider* when identifying NEs. NegEx Test Set is the most permissive, annotating whole clinically-relevant phrases as NEs regardless of their syntactic type (e.g., the statement "Right ventricular function is normal" is treated as a single entity as shown by the underlining). i2b2/VA guidelines only consider whole noun and adjective phrases as possible NEs (e.g., “her shortness of breath and coughing resolved” includes the modifier “her” in the NE). Similar to i2b2/VA, MiPACQ also indicates that whole noun phrases should be candidate NEs, but smaller units are typically used in practice (e.g., “her chest x-ray” leaves out the modifier “her”). SHARP predominantly annotates maximal strings that match UMLS terms as NEs, which often excludes long paraphrases and closed-class modifying adjectives (similar to MiPACQ), although there are some cases of CUI-less NEs and multi-span NEs.

Another difference in NE annotation guidelines is the *amount of overlap allowed* between NEs. The NegEx Test Set has only one phrase annotated per sentence, hence no overlap in NEs; i2b2/VA only annotates full noun and adjective phrases, so fully subsumed NEs are not allowed. In contrast, SHARP annotates subspans as long as they are mapped from the UMLS and of a different semantic type (e.g., both “chest” (anatomical site) and “chest x-ray” (procedure) in “her chest x-ray”). MiPACQ removes this restriction of different semantic types, but stipulates that some relationship must be shared between the subspan and the full span – this is in practice very similar to SHARP (e.g., there is a locationOf relationship between “chest” and “chest x-ray”).

Overall, the four guidelines are not as precise with negation annotation definitions as they are with NEs. The SHARP, MiPACQ, and NegEx Test Set representations imply a relation between an explicit negation marker and the negated term (e.g., a cue word like “no” would be marked, and the following term “shortness of breath” would then set a negation_indicator=present). The i2b2/VA guideline assumes a pragmatic inference about the intent of the author in describing his/her observations (e.g., “no shortness of breath” would mark assertion=absent without marking the cue word). This difference does lead to some minor morphology-related annotation differences. For example, “afebrile” is marked as “absent” for i2b2, but not in SHARP, MiPACQ, or NegEx Test Set since there is no external negation indicator.

3.3 SHARPn Polarity Module and YTEX NegEx

As with many existing approaches, the SHARPn Polarity module treats negation detection as a classification problem for NEs. This module is implemented within the cTAKES system, leveraging feature extraction and machine learning programming interfaces available in the ClearTK suite of tools (available at https://code.google.com/p/cleartk/). Features such as co-occurring bags-of-words, cue words, dependency regular expressions, and tree kernels served as input to a binary support vector machine (SVM) classifier. The polarity module used in our tests is currently available as a tagged branch of the Apache cTAKES source code repository, and will be part of a future cTAKES release.

We trained the SHARPn Polarity module on each of the four corpora; train/test splits were provided for the SHARPn, i2b2/VA, and MiPACQ corpora; for these three corpora, training and testing in our evaluations uniformly respected these training and testing splits (e.g., even in cases like training on SHARP data but testing on i2b2 data). Because the development set corresponding to the NegEx Test Set was not available, we used the Test Set as both training data and testing data; the tables presenting our results use hash shading to show when reuse training data invalidates the test performance measures.

Additionally, we used frustratingly easy domain adaptation (FEDA)[17] to build some of our multi-corpus models. This simple domain adaptation technique requires in-domain training data. Treating the four corpora as domains, the feature space is five times as large – each feature repeated per corpus, plus one “general” feature. At test time, the domain of the test sample is supplied to the classifier, and instances are classified with a weighting of the domain-specific model against the "general" model.

For both training and testing, we used gold standard NEs and negation annotations as defined in each of the corpora; we also used the default cTAKES pipeline and models (in the tagged version) to produce all other portions (e.g., sentence annotations, tokens, POS tags, dependency parses, constituency parses, semantic role labels). While there is some risk for error propagation from these other components into negation detection, we believe this risk is minimized and can be "ignored" for the main precision, recall, and F-measure metrics, because systemic errors would appear in both training and testing data, and any impact on negation performance would be mediated through their representation in a machine learning feature vector.

Our evaluations used the NegEx algorithm as a baseline, as implemented in the Yale cTAKES Extensions (YTEX) [11]. Because NegEx is a rule-based method, we would expect it to be immune to performance improvement or degradation based on training data. However, it is well-known that customization of rules is likely necessary when applying NegEx in settings other than the one in which it was initially developed. The YTEX negation module was used alongside the standard cTAKES pipeline.

4 Results

For simplicity in this section, we will consider each corpus as its own "domain," though we recognize that each corpus bridges multiple medical domains.

4.1 Single test corpus performance

The practical question a user might ask is: "How can I maximize negation detection performance for my data?" Table 4 below illustrates the difficulty of answering this question by showing performance on four corpora (columns) by various systems (rows). We have grouped these systems to be representative of three strategies for negation detection that are used in the community: 1) the unedited, rule-based algorithms; 2) machine learning classifiers when only out-of-domain data (OOD) is available; 3) machine learning classifiers when some in-domain data is available. Table 4 also includes significance bands down each column; pair-wise approximate randomization significance tests for F_1 score, aggregated by document, are reported for $p<0.05$. Values in a column labeled with different successive superscripted letters (e.g., 93.9^{a} and 92.6^{b}) indicate that there is a significant difference between two systems.

Table 4: Performance (F_1 score) in practical negation detection situations

	Test	**sharp**	**i2b2**	**mipacq**	**negexts**
Rule-based	**ytex (rules)**	62.3^{c}	82.1^{d}	$71.3^{a,b}$	95.3^{a}
ML with out-of-domain (OOD) training	**sharp**		80.7^{e}	61.2^{b}	87.3^{b}
	i2b2	$74.7^{b,c}$		$\mathbf{71.9^{a,b}}$	$\mathbf{95.4^{a}}$
	mipacq	$72.9^{b,c}$	82.6^{d}		59.3^{d}
	negexts	58.6^{c}	81.1^{e}	$70.6^{a,b}$	
	All 3 OOD	$\mathbf{79.0^{b}}$	$\mathbf{83.9^{c}}$	$69.1^{a,b}$	69.9^{c}
ML with in-domain	**In-Domain**	93.5^{a}	93.6^{a}	$73.6^{a,b}$	99.9
	All	89.7^{a}	92.6^{b}	$\mathbf{75.3^{a}}$	69.9^{c}

training	**All + FEDA**	**97.9**[a]	**93.9**[a]	73.9[a]	58.0[d]

First, the widely used rule-based NegEx algorithm (top row) performed quite well on the NegEx Test Set (F_1=95.3%). When used without modification on other corpora, performance fell to unacceptable levels (e.g., F_1=62.3% on SHARP data). As might be expected, we may conclude that widely-used rule-based algorithms need to be modified according to their target data.

For situations in which only OOD data is available (common in clinical text), one strategy is to use a single OOD corpus as training data (rows 2-5). Using a single OOD corpus has widely varying results, with models ranging from 59.3% to 95.4% F-score on the NegEx Test Set. Another strategy is to “use all the (OOD) data you have” (row 6), but again the results are mixed. With the highest OOD models in bold, it is not clear which strategy is optimal, and it is difficult to tell what pairs of corpora yield good performance (see below on “difficulty” and “usefulness”).

The situation is much improved when in-domain data is available (rows 7-9). Note that the performance of any OOD models are uniformly lower than training with in-domain data alone (row 7). We still face the same problem of whether to use a single in-domain corpus or to “use all the data you have” – the choice differs by corpus (row 7 vs row 8). However, we note that using domain adaptation (row 9) improves results over the in-domain data alone (row 7), though the results are not statistically significant at the p<0.05 level.

Thus, whether there is in-domain data available or not, we cannot conclude a uniform policy such as “use all available data to train your model” or “train a model on a single most similar corpus.” However, we can conclude that, if in-domain data is available, adding additional corpora via fully-supervised domain adaptation techniques will not hurt performance.

4.2 Corpus difficulty and usefulness

The “difficulty” of a corpus and the “usefulness” of a corpus seem to vary, as evidenced by the second portion of Table 4. Testing on MiPACQ data has an average F_1 score of 70.9% down the column of trained systems, indicating it is probably the most difficult to test on. Training on i2b2 data (row 3) achieved a macro-averaged F_1 score of 80.7% across the row of test sets, indicating its training set is perhaps the single most useful for training.

Difficulty and usefulness are not symmetric: i2b2 data is clearly the best out-of-domain training data for the NegEx Test Set (F_1=95.4%); but the converse is not true (F_1=81.1% for a NegEx-trained model on the i2b2 test set, significantly outperformed by MiPACQ with F_1=82.6%). Difficulty and usefulness also do not correlate directly with corpus size, number of NEs, or number of negated NEs (results not shown), confirming that the different domains have fundamentally different characteristics that are not overcome with more samples from a different domain.

4.3 Average performance and NE characteristics

We considered average performance of several models on multiple corpora. In Table 5 we include averages with and without FEDA (i.e., for rows 8-9 of Table 4), labeling pairwise statistical significance at p<0.05 between the domain adapted and non-domain adapted versions with an asterisk. The NegEx Test Set is used for training rather than testing.

Table 5: Average F-score with and without frustratingly easy domain adaptation (FEDA)

Test \Train	**All**	**+ FEDA**
sharp	89.66	97.87
i2b2	92.57	93.93*
mipacq	75.29	73.93
negex	▨	▨
macro-avg	85.84	88.58
micro-avg	91.91	93.28*

Here, we report both macro-averages (arithmetic mean of the three test sets) and micro-averages (weighted by the number of instances in each test set). The micro-averaged scores are heavily weighted towards the i2b2 numbers because the i2b2 test set is the largest; macro-averages, on the other hand, are much lower than has been previously reported in literature, in large part due to the difficulty of the MiPACQ corpus. Overall, except for MiPACQ data, domain-adapted models outperform un-adapted models.

Negation predictions were further analyzed to see if the differences in NE annotation guidelines influenced performance. Figure 1 shows that longer Named Entities are more difficult to negate correctly in all of the corpora; in the i2b2 corpus, single-word terms were easy to negate, whereas in other corpora single-word terms were substantially harder. This could be due to i2b2's different accounting of inherently negated terms such as "afebrile," yet as a whole there are insufficient examples of these terms to affect performance to the degree observed.

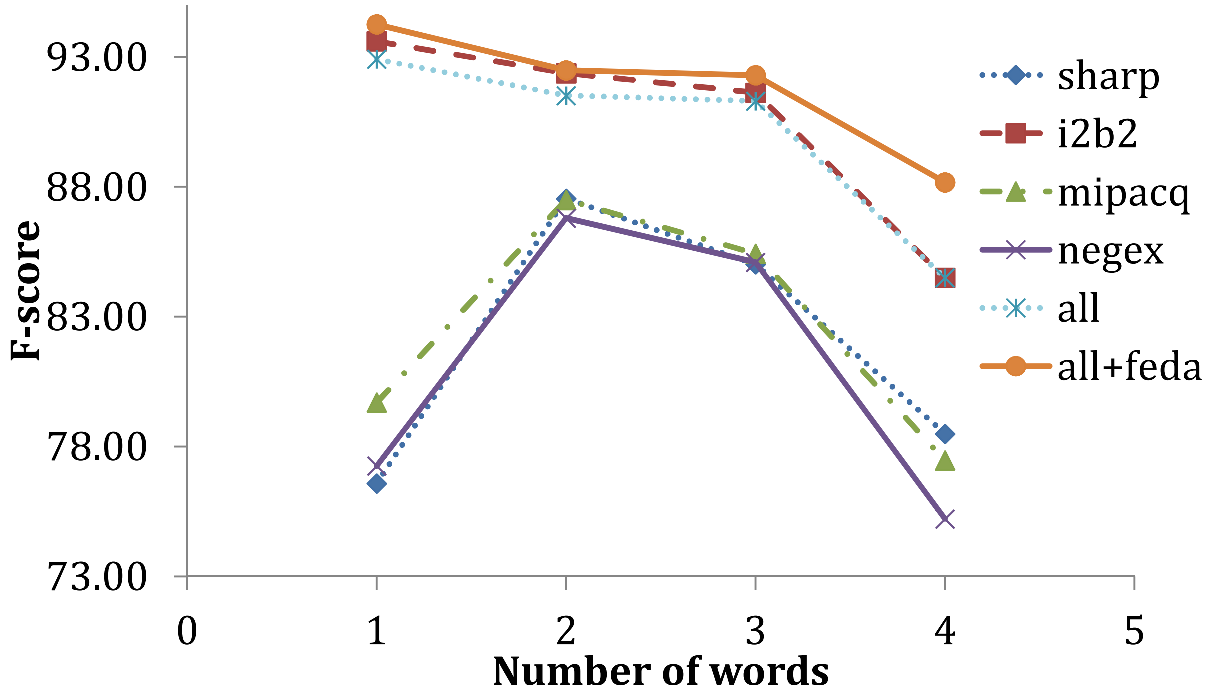

Figure 1: The effect of length on the average F-score of 6 models

As shown in Figure 2, this multi-corpus model (labeled "all" in the legend) also performed much more reliably on Labs, Symptoms, Events (including i2b2 "problems" and NegEx NEs), and Disorders than on other semantic groups. This was consistent regardless of which corpus was used to train a model. (Note that because of the differing annotation guidelines surrounding NEs, all i2b2 and NegEx named entities were considered Events. SHARP and MiPACQ semantic groups were used as labeled.)

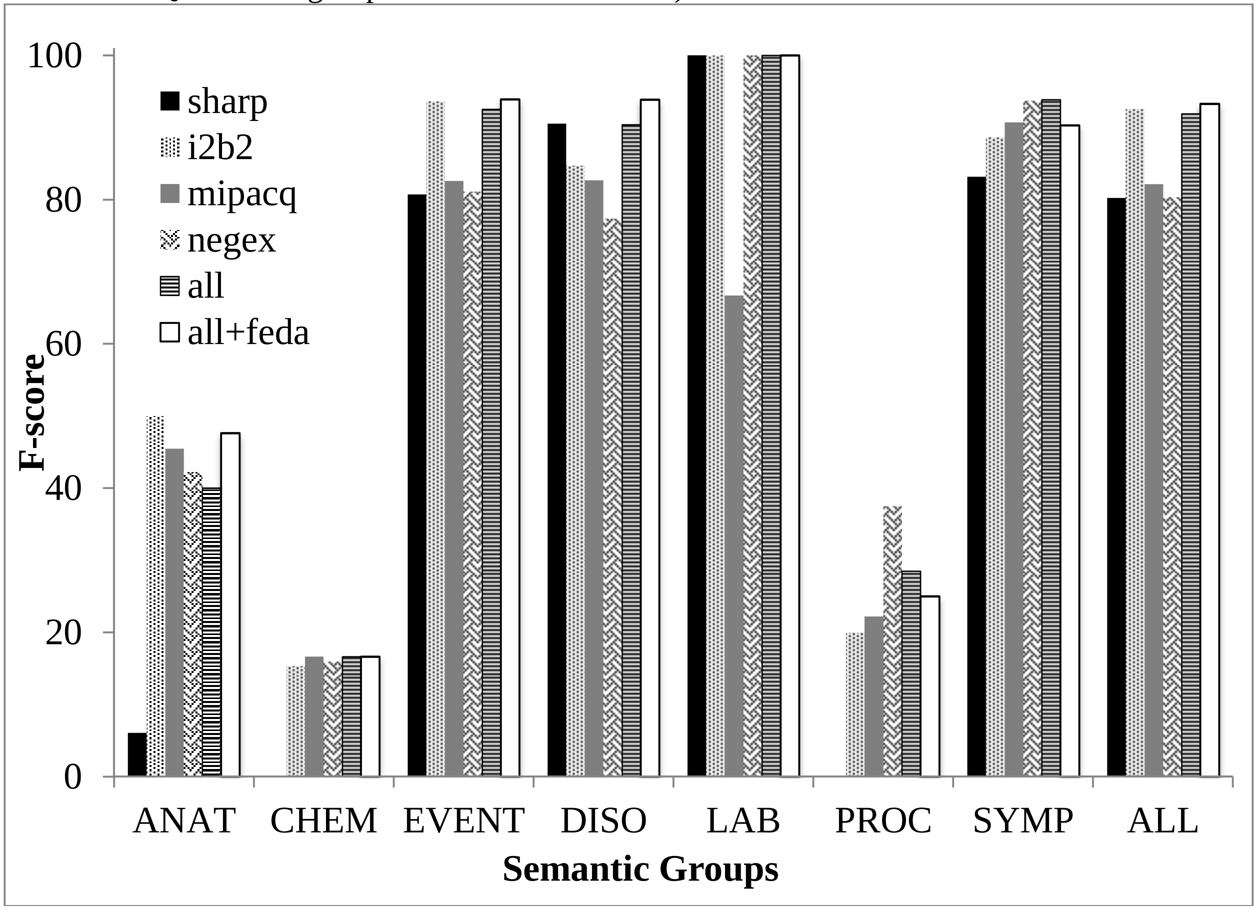

Figure 2: The effect of named entity (NE) semantic group on the F-score of 6 models

Training with the SHARP corpus had some of the worst performance, including near-zero performance on anatomical sites, chemicals and medications, and procedures semantic groups, despite having training data in those groups. Also, a MiPACQ-trained model did not outperform other models, despite that most of the test set NEs of minority semantic groups came from the MiPACQ corpus.

5 Discussion

5.1 Salient features

From the foregoing tests, NE properties like length and semantic group (and thus, annotation guidelines) did not fully explain the discrepancy in performance between different models. Thus, we qualitatively examined the broader differences between corpora by looking at negation contexts in each corpus. We defined negation contexts as the features of the SHARPn Polarity module.

Table 6 calculates and ranks the χ^2 statistic corresponding to each feature (i.e., on a 2x2 grid of whether the NE was negated vs. whether the feature was present) within all four sets of training data. Thus, the ranking in Table 6 corresponds to the model trained on “All” training sets, in row 8 of Table 4 and in the preceding section. Table 6 also compares the rank of features in the "all" model to salient features in each individual corpus.

Table 6: Top negation context features in a multi-corpus model, by chi-square value; and feature rank in domain-specific models

		Feature Rank in Training Data				
Feature Name	**Chi^2**	**all**	**i2b2**	**mipacq**	**negexts**	**sharp**
Deppath_dt_nmod_mod	16713.1	1	5	1	2	1
Bag_Preceding_0_5:no	15601.1	2	1	3	3	2
TreeFrag_AL_Polarity:(DT no)	15263.2	3	3	2	4	3
Bag_Preceding_0_10:no	14928.9	4	2	4	1	5
Bag_Preceding_0_3:no	14207.4	5	4	5	5	4
Preceding_0_5_0:no	10683.5	6	6	9	6	10
ClosestCue_PhraseCategory:no	9848.3	7	7	6	12	6
ClosestCue_Word:no	8866.9	8	9	7	7	15
ClosestCue_PhraseFamily:negation	8110.7	9	10	8	8	9
TreeFrag_AL_Polarity:(NP (DT no) (CONCEPT))	8038.3	10	8	18	13	23
Deppath_negverb->dobj_mod	3817.3	11	12	13	16	285
TreeFrag_AL_Polarity:(VBZ semclass_deny)	3809.4	12	13	10	15	851
Bag_Preceding_0_10:denies	3081.2	13	22	12	21	1195
TreeFrag_AL_Polarity:(DT any)	2721.2	14	43	11	24	486
Preceding_0_5_2:no	2672.9	15	15	28	22	53
ClosestCue_PhraseCategory:deny	2479.0	16	16	19	38	327
Bag_Preceding_0_5:denies	2380.3	17	28	16	26	2070
Bag_Following_0_5:or	2350.9	18	25	30	9	46
TreeFrag_AL_Polarity:(NP (DT no) (NML))	2247.9	19	27	44	34	19
Bag_Following_0_10:or	2242.1	20	26	29	10	39

It is evident that the most important features were consistent across all the corpora, representing the "easy cases" of negation: namely, when the word "no" is related to a concept by proximity or by syntax. The SHARP corpus differs somewhat, likely due to the sources of data for the SHARPn Seed Corpus: Mayo Clinic radiology reports (do not directly report a patient interaction) and Seattle Group Health breast cancer-related notes (only one example of a patient "denying" smoking). This distinction does not explain why MiPACQ, rather than SHARP, is a more "difficult" corpus.

5.2 The Big Picture for Negation Detection

Because of the relatively constrained pragmatic uses of negation in clinical text, negation detection algorithms are easy to optimize for specific corpora, as illustrated in Table 1. However, we believe the research community has at times conflated this with being immediately effective off-the-shelf. Evaluation of systems is artificially inflated by the ad hoc development of training and testing corpora and their differing annotation guidelines. When in-domain, consistently-annotated training data is scarce or nonexistent, negation detection performance remains unimpressive (middle portion of Table 4), just as in other NLP problems like parsing or named entity recognition. Furthermore, it is difficult to simply characterize the differences between domains, e.g., by NE length (Figure 1), semantic group (Figure 2) or lexical and syntactic context (Table 6).

To ensure excellent negation performance for a machine learning model, it appears that we still need to annotate examples of negation on the target corpus for fully supervised training (or domain adaptation). Similarly, rule-based methods need a development set and experts who can develop domain-specific rules. Thus, we conjecture that negation is not "solved" until negation is tailored to specific applications and use cases, or until the more general problem of semi-supervised domain adaptation is solved.

6 Conclusion

While a review of published work may suggest that the negation detection task in clinical NLP has been "solved," our multi-corpus analysis of negation detection indicates that it is easy to *optimize* for a single corpus but not to *generalize* to arbitrary clinical text. Though negation detection can be straightforward in constrained settings, both rule-based and machine-learning approaches have mixed results in heterogeneous corpora. Furthermore, more training data was not necessarily better for the common case in which no in-domain data is available. However, training on all available data was a good strategy when some in-domain data was available and domain adaptation techniques were used. Future work includes task-adaptive negation detection algorithms and semi-supervised domain adaptation.

Acknowledgements

The authors would like to acknowledge David Carrell and David Harris for their work with the corpora and associated guidelines. Guergana Savova and Martha Palmer for the SHARPn NLP Seed Corpus and guidelines, Hongfang Liu for help with quantitative framing of the work, Chen Lin for his help with exploring the feature space, and Wendy Chapman for her help with NegEx.

Funding

This work was supported in part by the SHARPn (Strategic Health IT Advanced Research Projects) Area 4: Secondary Use of EHR Data Cooperative Agreement from the HHS Office of the National Coordinator, Washington, DC. DHHS 90TR000201.

Author Contributions

SW led the study design and analysis and drafted the manuscript. Using initial algorithms by CC, MC, and SW, the system infrastructure was set up by MC and SW. TM led the development of the current algorithm, with help from SW, JM, SH, DC, MC, and CC. All authors helped with manuscript editing.

Competing Interests

There are no competing interests for this work.

References

1. Chapman W, Bridewell W, Hanbury P, Cooper G, Buchanan B. A simple algorithm for identifying negated findings and diseases in discharge summaries. Journal of biomedical informatics. 2001;**34**(5):301-10

2. 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

3. Elkin PL, Brown SH, Bauer BA, et al. A controlled trial of automated classification of negation from clinical notes. BMC Med Inform Decis Mak 2005;**5**:13 doi: 10.1186/1472-6947-5-13[published Online First: Epub Date]|.

4. Sohn S, Wu S, Chute CG. Dependency Parser-based Negation Detection in Clinical Narratives. AMIA Summits on Translational Science proceedings AMIA Summit on Translational Science 2012;**2012**:1-8

5. Learning to detect negation with 'not'in medical texts. Proc Workshop on Text Analysis and Search for Bioinformatics, ACM SIGIR; 2003.

6. Clark C, Aberdeen J, Coarr M, et al. MITRE system for clinical assertion status classification. J Am Med Inform Assoc 2011;**18**(5):563-7 doi: 10.1136/amiajnl-2011-000164[published Online First: Epub Date]|.

7. 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 doi: 10.1197/jamia.M2284[published Online First: Epub Date]|.

8. 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 doi: amiajnl-2011-000203 [pii]

10.1136/amiajnl-2011-000203[published Online First: Epub Date]|.

9. Daume H, Langford J, Marcu D. Search-based structured prediction. Machine Learning 2009;**75**(3):297-325

10. Harkema H, Dowling JN, Thornblade T, Chapman WW. ConText: An algorithm for determining negation, experiencer, and temporal status from clinical reports. Journal of biomedical informatics 2009;**42**(5):839-51

11. Garla V, Re VL, Dorey-Stein Z, et al. The Yale cTAKES extensions for document classification: architecture and application. Journal of the American Medical Informatics Association 2011;**18**(5):614-20

12. Vincze V, Szarvas G, Farkas R, Mora G, Csirik J. The BioScope corpus: biomedical texts annotated for uncertainty, negation and their scopes. BMC Bioinformatics 2008;**9 Suppl 11**:S9 doi: 10.1186/1471-2105-9-S11-S9[published Online First: Epub Date]|.

13. The CoNLL-2010 shared task: learning to detect hedges and their scope in natural language text. Proceedings of the Fourteenth Conference on Computational Natural Language Learning---Shared Task; 2010. Association for Computational Linguistics.

14. Tao C, Jiang G, Oniki TA, et al. A semantic-web oriented representation of the clinical element model for secondary use of electronic health records data. J Am Med Inform Assoc 2013;**20**(3):554-62 doi: 10.1136/amiajnl-2012-001326[published Online First: Epub Date]|.

15. Albright D, Lanfranchi A, Fredriksen A, et al. Towards comprehensive syntactic and semantic annotations of the clinical narrative. Journal of the American Medical Informatics Association 2013

16. Cairns BL, Nielsen RD, Masanz JJ, et al. The MiPACQ clinical question answering system. AMIA Annu Symp Proc 2011;**2011**:171-80

17. Daume III H. Frustratingly Easy Domain Adaptation. Proc. ACL. Prague, Czech Republic: Association for Computational Linguistics, 2007.

Figure legends