NEGATION'S NOT SOLVED: RECONSIDERING NEGATION ANNOTATION AND EVALUATION

Authors	Stephen Wu[1] Timothy Miller[2] James Masanz[1] Matt Coarr[3] Scott Halgrim[4] David Carrell[4] Cheryl Clark[3]
Affiliation	[1]Department of Health Sciences Research Mayo Clinic Rochester, MN, USA
	[2] Children's Hospital Boston Informatics Program Harvard Medical School Boston, MA, USA
	[3] The MITRE Corporation Bedford, MA, USA
	[4] Group Health Research Institute Seattle, WA, 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 Phone: +1(507) 538-0167 Fax: +1(507) 284-0360

Keywords

Natural language processing
Information extraction
Electronic medical records
Negation detection

Word Count

3,901 words

ABSTRACT

Objective

To characterize and ameliorate the weaknesses of clinical negation detection techniques across corpora that have different annotation schema.

Materials and Methods

Named entities and negation attributes have been manually annotated in several corpora, including the new SHARPn NLP Seed Corpus, the 2010 i2b2/VA NLP Challenge assertion annotations, the MiPACQ Corpus, and the NegEx Test Set. We analyze each of these corpora through the lens of two contrasting systems: a rule-based NegEx baseline, and the SHARP machine-learning-based attribute discovery tool.

Results

The machine learning system performed relatively well when trained and tested on data from a single corpus (e.g., F_1=93.5%, 93.6%, and 73.6% for SHARP, i2b2, and MiPACQ, respectively). When training and testing on different corpora, performance dropped significantly in most cases. Co-training tests typically performed better than cross-training, but co-training was not uniformly better than the best single model.

Discussion

We suggest that the weak cross-training and inconsistent co-training performance arise partially from differences in annotation guidelines for the corpora, most importantly in the way that named entities are defined and annotated. Both the training and evaluation of negation detection systems are affected by these differences in guidelines.

Conclusion

Though negation detection is a straightforward task in relatively constrained settings, when evaluated in heterogeneous corpora and annotation schema, performance may fall well below that of published benchmarks for both machine learning-based and rule-based approaches. Furthermore, it is difficult to determine the optimal mix of training data, or a standardized way to constrain evaluation metrics, since both are influenced by the corpus and annotation characteristics.

BACKGROUND AND SIGNIFICANCE

Negation in unstructured clinical text is a well-known phenomenon. It is crucial for any practical interaction with clinical text, since the medical significance of "no wheezing" is quite different from just "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, the medically pertinent concepts (and what can be said about them) are constrained. The sublanguage around these concepts that expresses negation (and other modality markers) is therefore constrained as well. Since existing algorithms have performed well at capturing negation,[2-8] many clinical natural language processing (NLP) practitioners consider negation detection a solved problem (see Table 1).

However, the present work uncovers some surprising potential pitfalls in negation annotation and detection. In the course of executing what would appear to be a standard negation task as part of the Strategic Health IT Advanced Research Project on the Secondary use of the EHR (SHARPn) Attribute Discovery team, we found that "benchmark" gold standard data sets (and their respective annotation guidelines) differed sufficiently to have a profound effect on the viability of negation detection algorithms. What follows is an exploration of the differences between four corpora, analyzed through the lenses of rule-based negation detection and machine learning-based negation detection. We conclude that practical negation detection in unrestricted clinical corpora is still a challenging task for both machine learning-based and rule-based approaches. Furthermore, it is difficult to determine an optimal mix of training data orto standardize evaluation metrics, since both are influenced by corpus-specific annotation guidelines. The results we report here pave the way for future work in domain-adaptive and task-adaptive methods, and illustrate the benefit of more extensive and consistently-annotated corpora.

As the Attribute Discovery team for SHARPn NLP, our overarching goal is to discover clinically relevant attributes of named entities (NEs). Aside from negation, these include uncertainty (i.e., the NE is possible but not confirmed), conditional usage (i.e., the NE depends on circumstances or is in the future), the subject under discussion (i.e., the NE is related to the patient, a family member, or someone else), and generic usage (i.e., the NE is not asserting something about a subject). In illustrating the influence of data sources on the negation detection task, we make use of two outputs of the SHARPn NLP team; first, the new SHARPn NLP Seed Corpus of clinical text with multiple layers of syntactic and semantic information, including NEs and polarity (i.e., negation). Comparisons are made between this corpus, the 2010 i2b2/VA NLP Challenge corpus, the MiPACQ corpus, and the NegEx Test Set. Second, the SHARPn Attribute Discovery tool has a new Polarity module currently available in the Apache cTAKES project (clinical Text Analysis and Knowledge Extraction System; ctakes.apache.org); a thorough methodological treatment is described in a forthcoming publication.

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.

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 negation detection module released with ConText[9] 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[10] 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[9]	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[11] that was used as part of the CoNLL 2010 Shared Task.[12] 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 negation annotations we explore; we do not explore scope annotations 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 such scope annotations overcome some recall issues for negation of non-standard terminology (e.g., “patient is not feeling as much like a pariah today”), they do not overcome issues in fine-grained annotation guideline distinctions (see Discussion section on Annotation Guidelines).

MATERIALS AND METHODS

This study is designed as an evaluation of negation detection systems on differenttest corpora. 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 discuss the new Polarity component of the SHARPn Attribute Discovery tool and briefly mention the other systems used for comparison.

Corpora and Guidelines

SHARPn NLP Seed Corpus

The SHARPn NLP Seed Corpus consists of 97 de-identified radiology notes related to Peripheral Arterial Disease (PAD) from Mayo Clinic, and 86 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. Negation is included in the CEM templates as an attribute of UMLS concepts.

The SHARPn NLP named entity (NE) annotations are Diseases and Disorders, Signs and Symptoms, Procedures and Methods, Devices, Medications and Drugs, and Labs. Spans of text of these types are mapped to concept unique identifiers (CUIs) in the UMLS, though an allowance is made for concepts not represented in the UMLS (i.e., “CUI-less” concepts). If subconcepts were of the same semantic group, the most specific concept within a span was mapped. For example, in the statement "no small bowel obstruction," the more specific "small bowel obstruction" would be annotated, as opposed to the more general "bowel obstruction." However, overlapping spans of different semantic groups were annotated (e.g., annotate "small bowel" in the above, since it is an anatomical site). Named entities, attributes, and relations were annotated in the SHARPn NLP corpus with a single pass. The guidelines call for an explicit negation indicator, the presence of which sets the polarity of a NE to -1 (negated); it is +1 (not negated) by default. Within its training set, the SHARPn Seed Corpus has 10,574 NEs labeled with a negation attribute (whether positive or negative).

2010 i2b2/VA NLP Challenge Corpus

The 2010 i2b2/VA NLP Challenge Corpus contained a total of 871 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 reports.

In the 2010 i2b2/VA NLP challenge corpus, annotators were instructed to mark "only complete noun phrases or adjective phrases" as concepts, including most modifiers (e.g., chest x-ray), articles and possessives (e.g., her chest x-ray), and up to one prepositional phrase per concept (e.g., pain in the chest). Concepts were constrained to be phrases of UMLS-like semantic groups titled "problems," "treatments," or "tests." Attribute annotations were only carried out on "problems" and consisted of 6 possible categories: "present," "absent," "possible," "conditional," "hypothetical," and "not associated with patient." The "absent" category matches most closely with negation in other annotation schema, but it includes inherently negated terms (e.g., afebrile) as well. There is only one assertion status possible per concept. The i2b2/VA “absent” annotations provided 11,968 training NEs labeled with a (asserted or negated) negation attribute.

MiPACQ Corpus

MiPACQ corpus[13 14] 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;353 *clinical questions* from the National Library of Medicine's Clinical Questions corpus (http://clinques.nlm.nih.gov), collected by interviews with physicians; and 13,091 sentences from Mayo Clinic clinical notes and pathology notes related to colon cancer.

Per the MiPACQ annotation guidelines, each of these sources is annotated with named entities of a full standard set of UMLS semantic groups,[15] as opposed to subsets used in other corpora. These are generally full noun phrases, but multiple annotations on the same string are permitted, especially when a relation can be identified (e.g., one annotation specifying a procedure, another specifying the anatomical site where it took place). Guidelines on negation are not detailed, but it is implied that the negation of a condition amounts to the absence of that condition. The MiPACQ Corpus provided 22,544 NEs labeled with a negation attribute (positive or negative) for training.

NegEx Test Set

The NegEx Test Set is a set 2,376 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,[9] 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).

Signs, symptoms, diseases, and findings with qualitative values were included as manually annotated named entities, but demographics, risk factors, and findings with quantitative values were excluded. Each NE was then annotated for negation, temporality (past, present, or future), and experiencer (patient vs. other). The NegEx Test Set provided 2,371 NEs labeled with a positive or negative negation attribute, which were used for both training and testing.

NegEx Baseline System (YTEX)

Evaluations used the NegEx algorithm, as implemented in the Yale cTAKES Extensions (YTEX),[10] as a baseline. 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 settings other than the one in which it was initially developed. The YTEX negation module was used alongside the standard cTAKES pipeline.

SHARPn Polarity Module

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/). 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 whenreuse training data invalidates the test performance measures.

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.

RESULTS

Single Corpus Cross-training

Table 2 shows the SHARPn Polarity module trained on each of the four training corpora (the rows), evaluated on each of the test corpora (columns). For simplicity in this section, we will refer to each corpus as a "domain," though we recognize that each corpus bridges multiple medical domains. "In-domain," then refers to training and testing a model in the same corpus; "cross-domain" refers to training a model on one corpus and testing that model on a different corpus.

Table 2: Machine learning models trained (rows) and tested (columns) on different corpora, and their comparison with a rule-based method (bottom row).

	sharp			i2b2			Mipacq			negexts		
Train	**P**	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**
sharp	**97.7**	**89.6**	**93.5**	93.3	71.1	80.7	86.6	47.3	61.2	97.3	79.2	87.3
i2b2	79.1	70.8	74.7	**96.2**	**91.2**	**93.6**	78.1	66.7	71.9	95.9	94.9	95.4
mipacq	83.8	64.6	72.9	91.8	75.0	82.6	**86.5**	**64.0**	**73.6**	98.6	42.4	59.3
negexts	56.9	60.4	58.6	88.4	74.9	81.1	69.9	71.3	70.6	100	99.8	99.9
none(ytex)	56.9	68.8	62.3	84.9	79.5	82.1	66.7	76.5	71.3	94.9	95.7	95.3

The diagonal cells in Table 2 show relatively strong results for in-domain F-measures (SHARP 93.5%, i2b2 93.6%, MiPACQ 73.6%; note that NegEx Test Set numbers are not meaningful). Off-diagonal cells report cross-domain results.

The widely used rule-based NegEx algorithm (bottom row) performed quite well on the NegEx Test Set (F_1=95.3%), but when used without modification on other corpora, performance fell to unacceptable levels (e.g., F_1=62.3% on SHARP data). In each data set, a machine learning model automatically trained on the in-domain corpus outperformed NegEx, and at least one cross-trained model also outperformed NegEx (e.g., on the SHARP test set, an i2b2-trained model gets F_1=74.7% vs. YTEX NegEx's F_1=62.3%).

The four corpora appear to differ in their *usefulness* as training sets. For training sets, this was measured by the macro-averages across rows (excluding in-domain tests): 80.7% (i2b2), 76.4% (SHARP), 71.6% (MiPACQ), and 70.1% (NegEx Test Set). Additionally, the corpora differ with respect to their *difficulty* as test sets, as we calculated by macro-averaging down columns (excluding in-domain tests). Without in-domain training data, evaluating on SHARP and MiPACQ corpora seemed to be particularly difficult (68.7% and 67.9%, respectively); i2b2 and NegEx test sets were significantly easier (81.5% and 80.7%, respectively).

The usefulness and difficulty of corpora is more nuanced than these averages, and they did not correlate directly with corpus size, as one might expect if the corpora were generated by the same (hypothetical) source distribution. For example, on NegEx test data, training with the largest corpus, MiPACQ, yielded significantly worse performance (F_1=59.3%) than training with the i2b2 corpus (F_1=95.4%), which is half its size. Furthermore, the poor performance was not uniform or symmetric; for example, there are two cross-training tests with recall in the 40-percent range: training on SHARP but testing on MiPACQ, and training on MiPACQ but testing on NegEx Test Set.

Out-of-domain Co-training

One practical question a user might ask is: "What corpora should I use to train a negation detection system for my data?" Table 3 below illustrates the difficulty of answering this question. The first four rows are reproduced from the F-measure columns of Table 2, where the first three rows are cross-domain tests ranked by score, the fourth row is in-domain tests.

Table 3: Performance in practical negation detection situations on a held-out corpus

	sharp	i2b2	mipacq	negexts
Out-of-domain 1	58.6	80.7	61.2	59.3
Out-of-domain 2	72.9	81.1	70.6	87.3
Out-of-domain 3	74.7	82.6	71.9	95.4
In-Domain	93.5	93.6	73.6	99.9
All 3 Out-of-domain	**79.0**	**83.9**	**69.1**	**69.9**
All 4	89.7	92.6	75.3	

Table 3's fifth row simulates a typical downstream user's situation: no in-domain training data is available, and all available out-of-domain corpora are used to build a model (a different model in each column on this row). Note that the performance of these large, conglomerate out-of-domain models is uniformly lower than training with in-domain data. For the SHARP and i2b2 corpora, “use all the out-of-domain data you have” is the best strategy. However, for the MiPACQ and NegEx Test Set corpora, choosing a single out-of-domain model would have been better.

The sixth row of Table 1 represents a single model that is trained on all training sets available, simulating the case of a downstream user who is able to annotate a sizable amount of in-domain data. This conglomerate model obtains the best-performance on the MiPACQ test set of all tested models; it performs slightly below single-in-domain-corpus training for SHARP and i2b2. 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.”

DISCUSSION

Differences in Annotation Guidelines

We have hypothesized that some of the discrepancy in performance between corpora was caused by differences in annotation guidelines, stemming primarily from different accounts of named entities. We should note that all annotation projects reported high inter-annotator agreement, but we do not have corpora that are multiply-annotated with *different* guidelines. Here, we qualitatively analyze the annotation guidelines concerning the annotation of both NEs (concepts) and attributes (assertion status).

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 most permissive is the

MiPACQ corpus, which annotates 17 UMLS Semantic Groups. 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 how wide of a *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 a negation marker and the negated term, therefore they require a cue word (e.g., a cue word like “no” would be marked, and the following term “shortness of breath” would then set a negation_indicator=present accordingly). 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 morphological-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. However, these attribute annotation differences do not seem sufficient to explain the overall differences in cross-domain vs. in-domain performance.

The Big Picture for Negation Detection

Training or developing negation detection systems is hindered when the sources of data and the annotation guidelines do not align or are not large enough to overcome differences. Negation detection systems have demonstrated the utility of clinical NLP within controlled domains (Table 1), but when in-domain training data is scarce or nonexistent, negation detection performance remains challenging (Table 3). Note that to ensure excellent negation performance for a machine learning model, we still need to annotate examples of negation on the target corpus for fully supervised in-domain training. If negation detection is “solved” given a prerequisite in-domain annotation effort, it is only partially “solved.” Rule-based methods do not provide relief from this: negation is not fully “solved” if it is “solved” given an expert who can develop domain-specific rules.

Negation *evaluation* is also greatly affected by differing annotation guidelines. For example, the i2b2/VA corpus was annotated for only "problem" NEs. Thus, evaluation results are not necessarily applicable if a user expects a wider range of NEs to be properly negated, for example in an information retrieval or large-scale population/corpus analysis. Conversely, if a downstream user of a negation detection system is only concerned with the clinical problems, as may be the case in an information extraction or controlled classification setting, then the i2b2 corpus is quite appropriate; the performance on the MiPACQ and SHARP corpora might be misleading, since it will evaluate on NEs that are not of interest We have not observed any corpus that would serve as a fully generalizable benchmark for the negation task in clinical NLP.

The solution is not necessarily to standardize the task of negation detection, or even negation annotations. Negation annotations may be used toward many different ends. Tasks such as downstream information extraction, classification, summarization, question-answering, or patient/cohort retrieval differ with respect to the requirements that they impose on a negation detection algorithm. Robust solutions may require tailoring to specific tasks.

CONCLUSION

While a review of published work may suggest that the negation detection task in clinical NLP has been "solved," our analysis of negation detection performance in multiple corpora indicates substantial work remains to be done. Though negation detection can be straightforward in constrained settings, both rule-based and machine-learning approaches have mixed results in heterogeneous corpora. Furthermore, it is difficult to determine the optimal mix of training data, or a standardized way to constrain evaluation metrics, since both are influenced by the corpus-specific annotation guidelines. Future work should include the development of empirical model-selection algorithms to automate the selection of training and testing data, and task-adaptive negation detection algorithms.

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 theSHARPn (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. 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

10. 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

11. 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]|.

12. 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.

13. 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

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

15. Bodenreider O, McCray AT. Exploring semantic groups through visual approaches. Journal of biomedical informatics 2003;**36**(6):414-32

Figure legends