NegAIT: A new parser for medical text simplification using morphological, sentential and double negation

Partha Mukherjee [a,*], Gondy Leroy [a], David Kauchak [b], Srinidhi Rajanarayanan [a], Damian Y. Romero Diaz [a], Nicole P. Yuan [a], T. Gail Pritchard [a], Sonia Colina [a]

[a] University of Arizona, Tucson, AZ, United States
[b] Pomona College, Claremont, CA, United States

ABSTRACT

Many different text features influence text readability and content comprehension. Negation is commonly suggested as one such feature, but few general-purpose tools exist to discover negation and studies of the impact of negation on text readability are rare. In this paper, we introduce a new negation parser (NegAIT) for detecting morphological, sentential, and double negation. We evaluated the parser using a human annotated gold standard containing 500 Wikipedia sentences and achieved 95%, 89% and 67% precision with 100%, 80%, and 67% recall, respectively. We also investigate two applications of this new negation parser. First, we performed a corpus statistics study to demonstrate different negation usage in easy and difficult text. Negation usage was compared in six corpora: patient blogs (4 K sentences), Cochrane reviews (91 K sentences), PubMed abstracts (20 K sentences), clinical trial texts (48 K sentences), and English and Simple English Wikipedia articles for different medical topics (60 K and 6 K sentences). The most difficult text contained the least negation. However, when comparing negation types, difficult texts (i.e., Cochrane, PubMed, English Wikipedia and clinical trials) contained significantly (p < 0.01) more morphological negations. Second, we conducted a predictive analytics study to show the importance of negation in distinguishing between easy and difficulty text. Five binary classifiers (Naïve Bayes, SVM, decision tree, logistic regression and linear regression) were trained using only negation information. All classifiers achieved better performance than the majority baseline. The Naïve Bayes' classifier achieved the highest accuracy at 77% (9% higher than the majority baseline).

© 2017 Elsevier Inc. All rights reserved.

1. Introduction

Nearly nine out of ten adults lack the knowledge to manage their health effectively, often resulting in several negative consequences, for example, low health literacy is associated with a higher rate of hospitalization and decreased use of preventive measures [1]. Additionally, patients without reasonable health literacy struggle to properly manage those diseases such as high blood pressure, asthma, diabetes and HIV/AIDS that require continuous care [2–6]. Other critical tasks such as communication between patients and healthcare providers are hindered by a lack of health literacy.

There are many possible approaches to improve patient health literary [7], however, providing text-based patient materials is one of the most efficient solutions to educate people. It is important for effective preventive care, treatment and recovery [8–11]

and is a major component for adherence [10,12–14]. It is critically important that the text is written to optimize reader comprehension because clear and understandable language has been shown to help patients remember medical information [12], motivate them to read and understand text [15], help medication adherence and disease management [16] and affect the patients' perception of the medical staff [14].

Many different text features affect text difficulty [17] and many different approaches have been suggested to increase comprehension, for example, replacing technical jargon with more common terms or improving the organization of text [18]. The use of information technology tools has also been advocated to assist with this process and improve understanding of clinical information and data [19,20].

In this paper, we examine the role of negation on text difficulty. Negation has been thought to affect communication and writing guidelines often include suggestions to avoid negation and use a positive form [12,21]. Several studies have shown how doctor-patient communication can benefit from avoiding the use

---

of negation both in terms of better interpretation of risk information and medication adherence intentions [12,14]. The use of negation has been shown to delay comprehension [22–24] and increase confusion [14].

We introduce a new negation parser for health and medical text that identifies three different types of negation: sentential, morphological and double negation. Morphological negations are individual negative words that are the result of either appending or prepending *negating* word-particles to a stem, for example: 'im-' in 'impossible', 'ir-' in 'irrelevant', and '-less' in 'limitless'. Sentential negations are formed by independent particles such as 'no', 'neither', 'nor', 'not' or the contraction '-t'/'-nt' as in "can't" [12,25]. Double negations can consist of the combination of two sentential negations as in "The patient *didn't* want *no* one to visit", but they are rare, especially in writing. A more common form of double negation is the combination of sentential and morphological negation, for example, "It is *not uncommon* for a woman to release more than one secondary oocyte per cycle in the first months after she quits taking oral contraceptives." [26].

We first describe the development and evaluation of the parser. We validate its performance against a human annotated gold standard that contains 500 medical sentences (study 1). We then apply the parser to show different negation usage in texts with different difficulty levels: patient blogs, Cochrane reviews, PubMed articles, clinical trial documents, and medical Wikipedia articles (both English and Simple English Wikipedia articles) (study 2). Finally, to understand the role that negation plays in text difficulty, we use predictive analytics to evaluate its usefulness as a feature for automated classification into easy and difficult text. We evaluate all three types of negations as input features using five machine learning classifiers (study 3). We conclude by discussing the usefulness of the parser along with possible future studies.

2. The effect of negation on readability

Linguistics recognizes three main types of negation across languages: morphological, sentential and verbal negation [27]. In this paper, we focus on the first two, morphological, sentential negation, and also double negation, with double negation representing multiple, interacting negation occurrences in a single sentence. Verbal negation does not occur in English [27] with the possible exception of the particle "ain't", which is rare in writing, and hence is not addressed.

Negation is an important feature in text comprehension. Many studies have shown that avoiding negation and using positive counterparts results in text that is more readable and understandable. Burgers et al. [12] substituted sentential negations with positive variants in medical leaflets. The substitutions resulted in an increase in actual comprehension, perceived comprehension and medication-adherence intentions of patients. The authors did not address morphological negations since they were scarce in their corpus. Similarly, Beukeboom et al. [14] found that avoiding negations resulted in improved comprehension in oral communication, and caused patients to view medical staff more positively, which may be an important factor in contexts such as delivering bad news [10].

Most research on the effects of negation has focused on sentential negation [22–24,28,29]. To our knowledge, the impact of morphological and double negation on readability has yet to be explored [12,30,31]. Evidence from psycholinguistics, neuroscience and natural language processing has shown that negation presents difficulties in information retrieval and processing [23,24], comprehension [14], and memory [12]. These results are consistent across languages, for example in English [14], German [23], Danish [22] and Dutch [12], which suggests that the processing difficulties

have their origins in the underlying cognitive processes involved in negation. However, sometimes negation cannot be avoided, though different variants may be more accessible. For example, "It is *unclear* from the scan if …." And "It is *not clear* from the scan if …" convey the same information using different types of negation. More research is required to determine which of these variants is better for user comprehension.

There are two key characteristics of negation that have been suggested to explain the effect of negation on readability [32,33]. First, negation make uses of working memory resources that are not otherwise used in positive sentences. It establish a point of comparison with a specific reality; a reader first has to build an interpretation of the message conveyed by the particular sentence and then a negated version of that interpretation [24]. This is a process that does not occur in positive sentences. While positive sentences do exclude other possible realities (e.g., 'the sky is blue' excludes the possibility of it being maroon or pink), they do not actively address those realities. Cognitive science calls this *the principle of truth*, where the mind only represents falsity when an expression refers directly to it in order to reduce the load on working memory [24]. The second characteristic of negation involves the use of more than one 'mental space' [25], which could also result in spending cognitive resources on negation. Regardless of the underlying cognitive mechanism, negation affects human comprehension and therefore text readability.

3. Identification of negation using natural language processing

Different tools exist that can detect some types of negation in text, such as NegEx [34], DEEPEN [35], MedLEE [36], ConText [37] and ontology-based-approaches [38]. However, these are not general-purpose parsers and mostly focus on negation of specific medical terms and conditions and often in very limited domains. For example, NegEx [34] requires that the user provides the list of phrases that might be negated: a user enters "skin pigmentation" to see if there are any occurrences of "skin pigmentation" in the input texts in negated form. For searching for information about a particular disease, this can be helpful. However, NegEx only identifies a very restricted set of negations and is not intended for all types of text. DEEPEN [35] was developed to decrease NegEx's false positive rate with the help of Stanford's dependency parser. The system was tested on EHR data from Indiana University and further evaluated using 159 clinical notes from the Mayo Clinic dataset for generalizability. MedLEE [36] system identifies negations that are followed by words or phrases which denotes specific semantic classes such as temporal change or clinical finding in radiology domain. The system also identifies the negation of only the following verb [39]. ConText [37] is also an extension of NegEx to determine whether clinical conditions are negated, hypothetical, historical or experienced by someone other than patients. It achieved good performance to infer the status of the clinical condition. Elkin et al. [38] describe an automated mechanism that assigns negations to clinical concepts. Concepts are identified using the SNOMED-CT tool. The study compares the accuracy with that of the human assigned negations. The challenge with these systems is that they are problem specific and do not offer general negation detection in medical texts.

4. Negation parser

4.1. Overview

We developed a general-purpose parser, NegAIT (Negation Assessment and Inspection Tool), to annotate negation in text. The parser takes text as input and identifies all occurrences of mor-

phological, sentential and double negation, not just those associated with a particular set of input phrases. The parser is written in a Java-Scala integrated environment using a rule-based approach utilizing the Open Domain Informer (ODIN) event extraction framework [40], a tool designed for information extraction. To detect negation, we first tokenize the text, split it into sentences, stem the words using the Porter stemmer [41] and parse it using the Stanford parser [42]. Then, we apply a set of hand-written rules that combine regular expressions and lexicons to identify the different negation types.

Morphological negations are identified based on a predefined set of morphemes. However, since relying on morphemes alone results in many false positives, candidate negative words are filtered using additional lexicons. Sentential negations are identified using the Stanford parser's dependencies [42] and a set of hand-written rules. Identification of double negations is based on the previously identified morphological and sentential negations. Below we describe each step in more detail.

4.2. Lexicons

Our parser relies on lexicons to annotate negations. We created two lexicons to filter candidate morphological negations (see Table 1). The first lexicon is the accept list which includes all words with prefixes that are valid morphological negations (top part of Table 1). To generate this lexicon we extracted all words from English Wiktionary [43] with the prefixes 'ab/dis/dys/il/im/in/ir/un' [44]. We accepted all words that can be decomposed into 'not + bare word' (e.g. 'abnormal', which can be decomposed into 'not normal') as morphological negation. This prevents our lexicon from including words that do not imply a strict case of negation ('disappear' cannot be decomposed as 'not appear' since the meaning is different) and words that contain patterns similar to a morpheme but that are part of a complete word, such as 'il-' as in 'illogical' vs. 'il-' as in 'Illinois'. English Wiktionary [43], Collins Online [45] and Merriam Webster Online [46] were consulted when the definition and usage of the word were not known.

For the prefix 'non/non-', we inspected a random selection of 500 words from the 6000 available on the Wiktionary list. The inspection showed that all of these words could be converted to 'not + bare word' (e.g. nonfatal = not fatal). As such, we included all words with a 'non'-prefix as negation. The final accept-list contained 5702 words with eight different prefixes. "un" was the most frequent prefix representing 5195 words and "ab" the least frequent with only 4 words.

The second lexicon is the discard list which is used to avoid false positives for words with the 'less' prefix or suffix. These words are considered a morphological negation *unless* they occur on this list. To create the discard list, we reviewed all words from Wiktionary that started or ended with 'less' (such as lessen and lesser). We consider a word to be a negation if it can be decomposed into 'without + bare word' (such as 'endless'). Words that did not fit this criterion were added to the discard list. Since most words that start or end with 'less' are negations, the discard list is relatively small and contains only 32 words.

4.3. Rules for negation detection

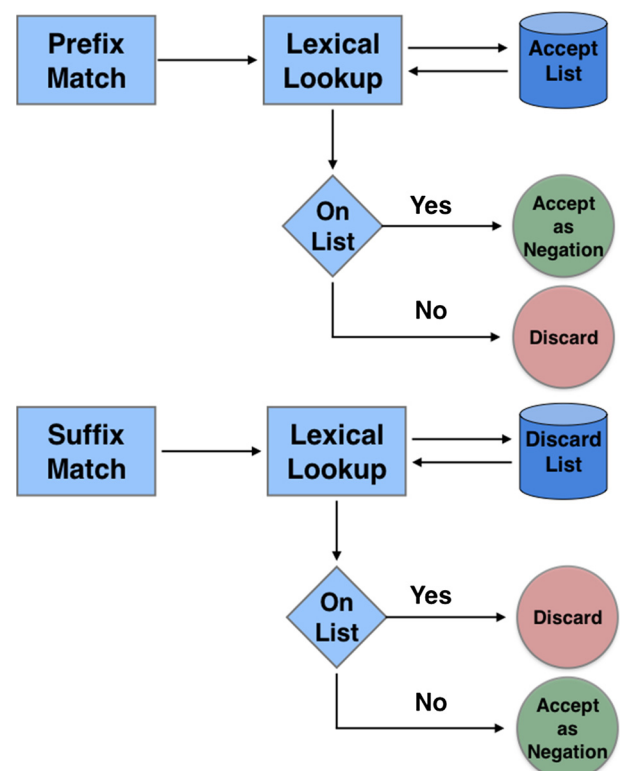NegAIT is a rule-based system using YAML (Yet Another Markup Language) and executed by ODIN [47].

4.3.1. Morphological negation

Morphological negations are identified in two-steps (Fig. 1). First, we identify all words that include one of the known negation prefixes or suffixes [44]. For prefixes, each identified word is checked against the accept list (Table 1) and accepted as a negation if it is present and discarded otherwise, e.g. the sentence "The doctor disagreed with the test report" contains a morphological negation since "disagree" starts with the prefix 'dis' and is on the accept list. For suffixes, each identified word is checked against the discard list and accepted as a negation when *not* on the list and discarded otherwise. For example the sentence "The ruthlessness of the doctor is represented by means of his attitude towards his patients" does not contain any morphological negation. Although "ruthless" does have the suffix 'less' it is present on the discard list and were therefore not be identified (ruthlessness does not signify 'not ruthless'; 'ruthless' is the adverb while 'ruthlessness' is the noun variant of this lemma).

Fig. 1. Process for the identification of morphological negations.

Table 1
Lexicons and examples as they appear on the English Wiktionary [43].

Lexicon	Morpheme	Number of items in list	Examples
1. Accept-list	"ab + word"	4	Abactinal, abnormal
	"dis + word"	33	Disagree, dissatisfied
	"dys + word"	55	Dysbalanced. dysfluent
	"il + word"	16	Illimited, illogic
	"im + word"	69	Immaterial, imperfect
	"in + word"	313	Inarticulate, incoherent
	"ir + word"	17	Irrational Irregular
	"un + word"	5195	Unaging, unbalanced, unfed
Total:		5702	
2. Discard-list	"word + less"	22	Helpless, ruthless, unless
	"less + en"	10	Lessen
	"less + er"		Lesser
Total:		32	

4.3.2. Sentential negation

Sentential negation is identified using the Stanford parser's dependencies and a short accept list (Fig. 2). The Stanford parser dependency annotations provide descriptions of grammatical relationships in a sentence. The negation modifier is a relation between a negation word and the word it modifies. If the Stanford parser's dependencies indicate negation (e.g. "n't" in: 'The medicine didn't stop the fever'), the negation is accepted as a sentential negation. We also conduct a lexical lookup to check for any occurrences of the following six words which are not detected by these dependencies: 'no', 'neither', 'nor', 'stop', 'none' and 'not'. If these are detected, we also identify them as a sentential negation. For example the sentence "The doctor could not diagnose the disease" contains the sentential negation "not" identified by the lexical lookup. These six words are complementary to the parser and were chosen by examining resources online as well as text examples. Anecdotally (and experimentally below), they do appear to capture most of the cases not already identified by the parser.

4.3.3. Double negation

The parser identifies double negations when a sentence contains two instances of sentential negation or an instance of sentential negation combined with an instance of morphological negation within a window of six words in a sentence. The double negation annotation is added in addition to sentential and morphological negation. For example:

(1) "The hospital won't allow no more visitors" contains double negation since "won't" and "no" are two sentential negations,
(2) "It is not illogical to carry out the experiment" contains double negation contains sentential negation "not" and morphological negation "illogical" (from lexical lookup).

The six-word window was decided empirically during the development process; it is large enough to capture most double

negations, but small enough to ignore negations that span unrelated phrases in a sentence.

5. NegAIT parser evaluation

We created a gold standard data set of negations to evaluate the negation parser's performance.

5.1. Gold standard data set and metrics

From both English Wikipedia [48] and Simple English Wikipedia [49] we randomly sampled 60 articles on diseases (30 from each) and then selected 250 sentences from each by selecting the first 10 sentences of each article, or as many sentences as possible in cases where the article was too short. Three independent annotators, who were not part of the development team, each identified sentential, morphological and double negations in each sentence. All annotations were found to be highly consistent with Cohen's kappas of 0.97, 0.95 and 1.0 for sentential, morphological and double negation respectively. For all negations combined, the Cohen's kappa was 0.88. To get a single gold standard data set, the results of the three annotators were combined with the third annotator making the final decision on the few differences in annotations that existed.

The gold standard included 59 morphological negations, 63 sentential negations and 4 double negations.

Examples are:

(1) Sentential Negation: Although vaccines have been developed, **none** are currently available in the United States.
(2) Morphological Negation: It is related to Typhoid fever, but such as Typhoid, it is **unrelated** to Typhus.
(3) Double negation: Since the organism hides from the body's immune system in red blood cells, it is difficult if **not impossible** for an infection to be totally cleared.

As double negatives are a rare event, we created an additional small set of sentences containing two or more negations for a follow-up study testing double negation. We extracted sentences from 625 medical texts from normal Wikipedia (see Table 3) that contained either one of negative words from: "not", "no", "neither", "none", "nobody", "n't", "never", "nowhere", "nothing", and "negative". We identified 1093 such sentences and tested our parser on 250 sentences that were chosen randomly. An independent expert identified 9 double negations in these 250 sentences.

To measure the performance of the parser we measure precision and recall for each negation type:

$$Precision = \frac{true\ positives}{true\ positives + false\ positives}$$

$$Recall = \frac{true\ positives}{true\ positives + false\ negatives}$$

5.2. Results

Overall, the parser achieved 93% precision and 89% recall (Table 2). Both precision and recall are highest for sentential negation. All sentential negations were identified by the parser resulting in a recall of 100%, though there were some false positives resulting in a 95% precision. For morphological negation the explicit accept list resulted in a fairly high precision of 88%, but not all words were on the list, resulting in a recall of 81%. As the tool evolves we will continue to update the lexicons. The performance for double negations is lower since double negations are infrequent
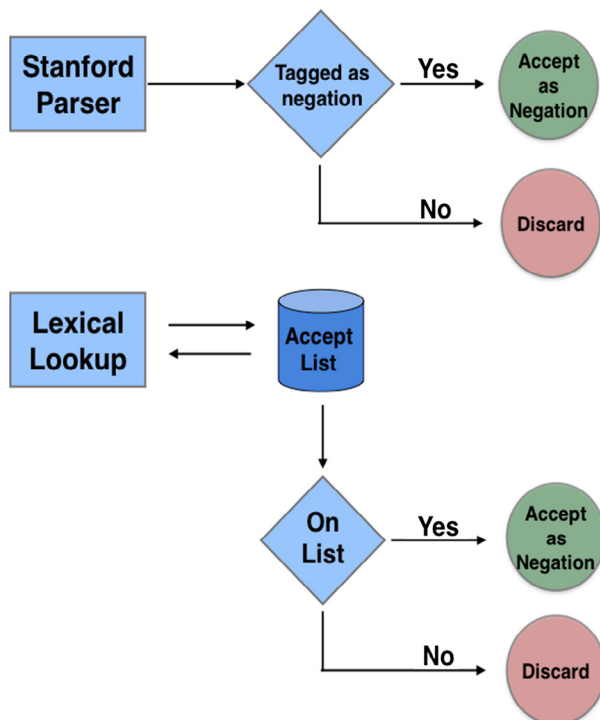

Fig. 2. Process for the identification of sentential negations.

Table 2
Parser performance on gold standard dataset.

(N = 500)	Precision (%)	Recall (%)
Morphological neg.	88.46	80.70
Sentential neg.	95.00	100.00
Double neg.	66.67	66.67
All negations	93.27	88.99

and identifying double negations relies on correctly identifying two negation occurrences in the sentence.

The parser performance for the study focusing explicitly on double negation is shown in Table 2a with two different settings for window size. Precision is highest (70%) with a small window size with reasonable recall (78%). With a larger window size, recall is much higher (89%) but precision is low (53%).

6. Negation corpus statistics analysis

Our ultimate goal is to develop a text simplification tool that includes simplification of negation. In previous work, we discovered through corpus statistics that text difficulty is related to the frequency of phenomena, i.e., frequency of terms and of grammatical structures. Follow-up user studies showed better user comprehension for text with more frequent patterns [50]. We follow a similar approach for negation and conduct first a corpus statistics analysis for the different types of negation. The result of this analysis will inform further studies and our negation tool.

6.1. Datasets

We applied NegAIT to six types of medical texts with varying degrees of difficulty: (a) patient blogs, which are typically regarded as easier texts, (b) Cochrane abstracts [51], (c) PubMed abstracts [52], (d) clinical trial text [53], (e) English Wikipedia articles on illnesses [48], and (f) Simple English Wikipedia articles [49] on illnesses. The blogs and Simple English Wikipedia articles represent easy text and the other four represent difficult text. To ensure a variety of topics where possible, we gathered texts from Cochrane, PubMed and Clinical trials by searching for 15 illnesses via their respective search interfaces. We chose conditions that are the leading causes of death according to the CDC (heart disease, cancer, respiratory disease, stroke, Alzheimer's disease, diabetes, influenza, pneumonia, nephritis, nephrosis and suicide), as well as four additional, common conditions (obesity, arthritis, asthma and autism). For the other three corpora we collected general texts on medical and health-related topics since they do not have a search interface. Table 3 shows the details of the corpora. Blogs and Simple English Wikipedia proved the most difficult to gather a large number of articles.

- Blogs: We identified ten different websites containing patient blogs and downloaded all posts from these websites for a total of 139 blog posts.
- Simple English Wikipedia articles: We downloaded 289 Simple English Wikipedia articles [49]. We collected the texts from each and every article in the list.

Table 2a
Parser performance for double negation on additional study.

(N = 250)	Precision (%)	Recall (%)
Window size = 6	53.33	88.89
Window size = 2	70.00	77.78

Table 3
Corpora for analysis.

Corpus	# of exts	# of sentences	Avg. sentence length per text
Easy text			
Blogs	139	4473	32.18
Simple English Wikipedia	289	6591	22.81
Difficult text			
Cochrane	2641	91,184	34.53
PubMed	1226	19,698	16.06
Clinical Trials	12,763	48,304	3.79
English Wikipedia	625	60,108	96.17

- Cochrane: The Cochrane database contains review articles of PubMed studies. We queried the database for the same 15 diseases and downloaded all abstracts for a total of 2641 articles: arthritis (163), Alzheimer's (71), cancer (664) to name a few.
- PubMed: We queried PubMed for the same 15 diseases and collected all abstracts from the articles. This resulted in 1226 different abstracts.
- Clinical Trials: We queried for each of the 15 diseases and downloaded the maximum number of trials available: nephritis (354), nephrosis (74), Suicide (701) and autism (634). We cleaned the XML tags and use the text from "Brief Summary", which results in 12,763 texts corresponding to clinical trials.
- Regular English Wikipedia articles: We downloaded 625 regular Wikipedia articles on diseases in English [48]. We accessed every disease in the list [48] and collected the texts from its corresponding Wikipedia page.

We are interested in the differences between known easy texts compared to known difficult texts, i.e., the differences in negation usage in blogs and Simple English Wikipedia versus the other corpora.

6.2. Results

Table 4 shows the distribution of negation (all types combined) in the different corpora. Overall, English and Simple English Wikipedia articles along with Cochrane contain the most negation (20%, 19% and 19%, respectively) followed by Blogs (18%), clinical trials (15%) and PubMed (12.5%).

Fig. 3 shows the distribution of different types of negation within different medical corpora. Blogs and Simple English Wikipedia articles have less morphological negation than the other four (5% and 4.5%, respectively), while sentential negation is higher (15% and 14%). Double negations are scarce with less than 1% of sentences containing them.

To understand how different types of negation correspond to text difficulty, we compare the easy texts (Blogs and Simple English Wikipedia) to the difficult texts (the other four) and found that *easier texts contain fewer morphological negations, but more sentential negations.*

Table 4
Distribution of negation.

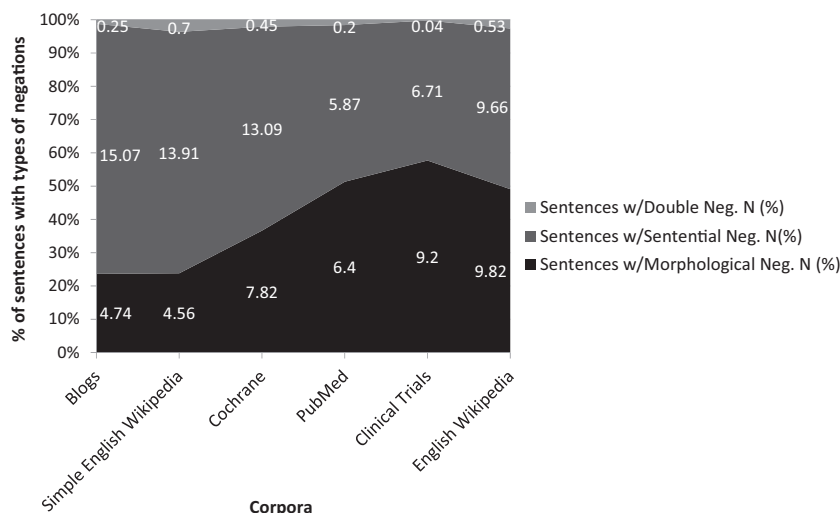| Corpus	Sentences in corpus	Sentences with negation (%)
Blogs	4473	18.91%
Simple English Wikipedia	6491	19.15%
Cochrane	91,184	19.39%
PubMed	19,698	12.48%
Clinical trials	48,304	15.14%
English Wikipedia	60,108	20.00%

Fig. 3. Distribution of negation types for different corpuses.

To verify that these trends were statistically significant, we performed a Welch's *t*-test, with Bonferroni correction to compensate for repeated testing, for each of the comparisons between the easy and difficult texts for each relationship. All differences were found to be statistically significant at p < 0.01.

7. Negation and text difficulty: A classification study

The corpus study showed that different types of text utilize different types of negation and with differing frequencies. To further understand the role that negation has on text difficulty, we examined the task of predicting the difficulty of a document using only the frequency of the different types of negations. If negation does play a role, then we will be able to distinguish between easy and difficult texts using it as a feature.

7.1. Datasets and classifiers

We used the 625 documents from English Wikipedia to represent difficult documents and the 289 documents from Simple English Wikipedia to represent easy documents. We ran NegAIT on each document and generated three features for each: (1) proportion of sentences with sentential negation, (2) proportion of sentences with morphological negation and (3) proportion of sentences with double negation.

We used five standard binary classifiers to compare the accuracy of identifying easy vs difficult texts: Naïve Bayes, logistic regression, decision trees, support vector machines (SVM) and linear regression. We used the R 3.3.1 libraries run the classifiers on the data.

The dataset contains about twice as many difficult documents (i.e. English Wikipedia) as easy documents (i.e. Simple English Wikipedia) resulting in a baseline majority prediction of ~68% by always predicting difficult. Most classifiers still perform well in this setting (as we see below) as long as there is not an extreme distribution bias in the labels [54–56]. However, as an additional experiment, we also created a balanced data set containing 50% difficult and 50% easy document by randomly choosing a subsample of 289 documents from the normal Wikipedia set (which represents difficult documents).

7.2. Results

We randomly selected 80% of the articles for training and 20% for testing. The accuracy of each of the classifiers on the test set is shown in Table 5 as well as the majority baseline (68%). The majority class baseline is the accuracy achieved in the test set by assigning the most common label from the training set, which is "difficult" in this study. The Naïve Bayes' classifier achieves the highest accuracy with 77% accuracy and linear regression performs with worst with 71% accuracy, only slightly above the majority baseline. For the balanced data set (majority baseline 50%), the Naïve Bayes' classifier remains the best performer with 78% accuracy followed closely by SVM (77.6%). Performance of linear regression remained low at 67% while performance for classification trees remained the same at 75.1%. These results show the importance of negation in text difficulty. Our parser can also be leveraged to help determining the text difficulty; using only negation features allows for a ~9% increase in accuracy with the best classifier.

8. Discussion

We developed and evaluated a negation parser, NegAIT, for use in corpus analysis and as part of a text simplification tool. The parser detects sentential, morphological and double negation with good precision and recall for both sentential negation (95% and 100%, respectively) and morphological negation (88% and 80%) and somewhat lower precision and recall for double negation (67% and 67%).

We used NegAIT to explore how negation is used in different types of medical text. The results were counter to common conjecturing with easy texts containing a higher frequency of negations.

Table 5
The accuracy and misclassification of classifiers for Wikipedia articles.

Classifier	Accuracy (%)	Accuracy(%), balanced set
Naïve Bayes	77.00	78.23
Logistic regression	74.84	72.01
Linear regression	71.03	67.03
Classification tree	75.32	75.12
SVM	75.41	77.63
Majority baseline (predict all as "difficult")	68.30	50.00

However, the types of negation differed and morphological negations were significantly less frequent in easy text (blogs and Simple English Wikipedia) than in difficult text (PubMed, Cochrane, clinical trials and English Wikipedia). Finally, we showed the importance of negation in predicting text difficulty through a classification study. Using only the proportion of the three types of negations in an article (morphological, sentential and double), classification algorithms are able to better predict whether text is difficult than a majority baseline: negation does play a role in the difficulty of medical text.

The finding that morphological negations are less frequent in blogs and Simple English Wikipedia may be related to lexical frequency. Lexical frequency of negated words (e.g., 'instability') may be lower than their positive counterparts (in this case 'stability'). Since low lexical frequency is correlated with more difficult texts that are often directed to a more specialized audience [57,58], it is possible that the smaller number of morphological negations in blogs and Simple English Wikipedia (the easier texts) reflects the tendency to use more familiar (i.e., high frequency) terms, i.e. the frequency of 'not clear' is expected to be higher than that of 'unclear'. This also explains the counter-intuitive finding in our study with easier texts containing more negation.

Our future work will include investigating parse-tree structure for detecting double negation and increasing precision and recall and comparing lexical frequencies of words with and without morphological negations. Additionally, we will further explore the impact of morphological, sentential and double negations on the readability of medical texts through user studies and include the negation parser in our text simplification tool.

Resources

"The negation parser described in this work is available for download at http://nlp.lab.arizona.edu/content/resources. Those who use this parser should cite our paper in their research".

Conflict of interest

I hereby declare that I have no conflict of interest with the editors of Journal of Biomedical Informatics, and reviewers of this journal. None of my co-authors has any conflict of interest with the editors of JBI and the reviewers of this journal.

Acknowledgements

References

[1] U.D.O. Health, H. Services, Health Literacy and Health Outcomes, Quick Guide to Health Literacy: Fact Sheet, 2008. <https://health.gov/communication/literacy/quickguide/factsliteracy.htm>.

[2] M.V. Williams et al., Relationship of functional health literacy to patients' knowledge of their chronic disease: a study of patients with hypertension and diabetes, Arch. Inter. Med. 158 (2) (1998) 166–172.

[3] D. Schillinger et al., Association of health literacy with diabetes outcomes, Jama 288 (4) (2002) 475–482.

[4] D. Schillinger et al., Closing the loop: physician communication with diabetic patients who have low health literacy, Arch. Inter. Med. 163 (1) (2003) 83–90.

[5] M.V. Williams et al., Inadequate literacy is a barrier to asthma knowledge and self-care, Chest J. 114 (4) (1998) 1008–1015.

[6] N. Langebeek et al., Predictors and correlates of adherence to combination antiretroviral therapy (ART) for chronic HIV infection: a meta-analysis, BMC Med. 12 (1) (2014) 142.

[7] N. Egbert, K.M. Nanna, Health literacy challenges and strategies, Online J. Issues Nurs. 14 (3) (2009).

[8] O. Mouradi et al., Influence of text and participant characteristics on perceived and actual text difficulty, in: 2013 46th Hawaii International Conference on System Sciences (HICSS), IEEE, 2013.

[9] L. Nielsen-Bohlman, A.M. Panzer, D.A. Kindig, Committee on Health Literacy. Health Literacy: A Prescription to End Confusion, 2004, The National Academies Press, 2006.

[10] L.M. Ong et al., Doctor-patient communication: a review of the literature, Soc. Sci. Med. 40 (7) (1995) 903–918.

[11] K.B.H. Zolnierek, M.R. DiMatteo, Physician communication and patient adherence to treatment: a meta-analysis, Med. Care 47 (8) (2009) 826.

[12] C. Burgers et al., How (not) to inform patients about drug use: use and effects of negations in Dutch patient information leaflets, Pharmacoepidemiol. Drug Safety 24 (2) (2015) 137–143.

[13] G. Leroy, D. Kauchak, O. Mouradi, A user-study measuring the effects of lexical simplification and coherence enhancement on perceived and actual text difficulty, Int. J. Med. Inform. 82 (8) (2013) 717–730.

[14] C. Burgers, C.J. Beukeboom, L. Sparks, How the doc should (not) talk: when breaking bad news with negations influences patients' immediate responses and medical adherence intentions, Patient Educ. Counsel. 89 (2) (2012) 267–273.

[15] G. Leroy, D. Kauchak, The effect of word familiarity on actual and perceived text difficulty, J. Am. Med. Inform. Assoc. 21 (e1) (2014) e169–e172.

[16] A. Arbuthnott, D. Sharpe, The effect of physician–patient collaboration on patient adherence in non-psychiatric medicine, Patient Educ. Counsel. 77 (1) (2009) 60–67.

[17] D. Kauchak et al., Text simplification tools: using machine learning to discover features that identify difficult text, in: 2014 47th Hawaii International Conference on System Sciences, IEEE, 2014.

[18] K. Nagel et al., Using plain language skills to create an educational brochure about sperm banking for adolescent and young adult males with cancer, J. Pediatric Oncol. Nurs. 25 (4) (2008) 220–226.

[19] S.B. Johnson et al., Data management in clinical research: synthesizing stakeholder perspectives, J. Biomed. Inform. 60 (2016) 286–293.

[20] M.A. Morid et al., Classification of clinically useful sentences in clinical evidence resources, J. Biomed. Inform. 60 (2016) 14–22.

[21] Writing and Reading Success Center, L.B.C.C. Writing Clear Sentences: Avoiding Negative Sentence, 2014. <http://www.lbcc.edu/WRSC/SentenceRevision.cfm>.

[22] K.R. Christensen, Negative and affirmative sentences increase activation in different areas in the brain, J. Neurolinguist. 22 (1) (2009) 1–17.

[23] B. Kaup, Negation and its impact on the accessibility of text information, Memory Cogn. 29 (7) (2001) 960–967.

[24] B. Kaup, J. Lüdtke, R.A. Zwaan, Processing negated sentences with contradictory predicates: is a door that is not open mentally closed?, J Pragmat. 38 (7) (2006) 1033–1050.

[25] A. Verhagen, Constructions of Intersubjectivity Discourse Syntax and Cognition, Oxford University Press on Demand (2005).

[26] A&P Anatomy, Reproductive System, Quizlet, 2016. <https://quizlet.com/42556025/ap-anatomy-reproductive-system-flash-cards/>.

[27] Ö. Dahl, Typology of negation, Exp. Neg. (2010) 9–38.

[28] P.B. Gough, Grammatical transformations and speed of understanding, J. Verbal Learn. Verbal Behav. 4 (2) (1965) 107–111.

[29] D.I. Slobin, Grammatical transformations and sentence comprehension in childhood and adulthood, J. Verbal Learn. Verbal Behav. 5 (3) (1966) 219–227.

[30] S. Sohn, S. Wu, C.G. Chute, Dependency parser-based negation detection in clinical narratives, in: AMIA Summits on Translational Science Proceedings AMIA Summit on Translational Science, 2012, pp. 1–8.

[31] S. Gindl, K. Kaiser, S. Miksch, Syntactical negation detection in clinical practice guidelines, Stud. Health Technol. Inform. 136 (2008) 187.

[32] M.C. MacDonald, M.A. Just, Changes in activation levels with negation, J. Exp. Psychol.: Learn. Memory Cogn. 15 (4) (1989) 633.

[33] B. Kaup, R.A. Zwaan, Effects of negation and situational presence on the accessibility of text information, J. Exp. Psychol.: Learn. Memory Cogn. 29 (3) (2003) 439.

[34] W.W. Chapman et al., A simple algorithm for identifying negated findings and diseases in discharge summaries, J. Biomed. Inform. 34 (5) (2001) 301–310.

[35] S. Mehrabi et al., DEEPEN: a negation detection system for clinical text incorporating dependency relation into NegEx, J. Biomed. Inform. 54 (2015) 213–219.

[36] C. Friedman et al., A general natural-language text processor for clinical radiology, J. Am. Med. Inform. Assoc. 1 (2) (1994) 161–174.

[37] H. Harkema et al., ConText: an algorithm for determining negation, experiencer, and temporal status from clinical reports, J. Biomed. Inform. 42 (5) (2009) 839–851.

[38] P.L. Elkin et al., A controlled trial of automated classification of negation from clinical notes, BMC Med. Inform. Decis. Making 5 (1) (2005) 13.

[39] R. Romano, L. Rokach, O. Maimon, Automatic discovery of regular expression patterns representing negated findings in medical narrative reports, in: International Workshop on Next Generation Information Technologies and Systems, Springer, 2006.

[40] M.A. Valenzuela-Escárcega, G. Hahn-Powell, M. Surdeanu, Description of the Odin Event Extraction Framework and Rule Language, 2015. Available from: arXiv:1509.07513.

[41] M.F. Porter, An algorithm for suffix stripping, Program 14 (3) (1980) 130–137.

[42] MC. De Marneffe, B. MacCartney, C.D. Manning, Generating typed dependency parses from phrase structure parses, in: Proceedings of LREC, 2006.

[43] Wiktionary, 2016. <https://en.wiktionary.org/wiki/Wiktionary:Main_Page> [cited 2015 November 11].

[44] V. Hulse, Productivity in Morphological Negation: A Corpus-Based Approach, 2011.

[45] Free Online Dictionary, Online Dictionary of American English, 2016. <http://www.collinsdictionary.com/>.

[46] M.-W. Dictionary, Merriam-Webster Online Dictionary, Retrieved November, 2010, 6.

[47] MA. Valenzuela-Escárcega, G. Hahn-Powell, M. Surdeanu, Odin's runes a rule language for information extraction, in: Proceedings of the 10th Edition of the Language Resources and Evaluation Conference (LREC), 2016.

[48] Wikipedia, 2016. <https://en.wikipedia.org/wiki/Lists_of_diseases> [cited 2016 July 1].

[49] Simple Wikipedia, 2016. <https://simple.wikipedia.org/wiki/List_of_diseases> [cited 2016 July 1].

[50] D. Kauchak, G. Leroy, A. Hogue, Measuring Text Difficulty Using Parse-Tree Frequency, *J. Assoc. Inf. Sci. Technol.*, Forthcoming, 2017.

[51] T.C. Collaboration, Cochrane, 2016.

[52] N.C.F.B. Information, PubMed Health, U.S. National Library of Medicine, 2016.

[53] Clinical Trials, ClinicalTrials.gov, 2016.

[54] P. Mukherjee, B.J. Jansen, Evaluating classification schemes for second screen interactions, in: 2015 International Conference on Computing, Networking and Communications (ICNC), IEEE, 2015.

[55] A. Dwyer, Managing unbalanced data for building machine learning models, in: Predictive Analytics and Data Mining, USA, 2015. <www.simafore.com>.

[56] K. Žubrinić, M. Miličević, I. Zakarija, Comparison of Naive Bayes and SVM classifiers in categorization of concept maps, Int. J. Comput. 7 (3) (2013) 109–116.

[57] G. Leroy, J.E. Endicott, Combining NLP with evidence-based methods to find text metrics related to perceived and actual text difficulty, in: Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium, ACM, 2012.

[58] G. Leroy et al., User evaluation of the effects of a text simplification algorithm using term familiarity on perception, understanding, learning, and information retention, J. Med. Inter. Res. 15 (7) (2013) e144.