Published in final edited form as:
J Biomed Inform. 2010 December ; 43(6): 953–961. doi:10.1016/j.jbi.2010.08.003.

Detecting Hedge Cues and their Scope in Biomedical Literature with Conditional Random Fields

Shashank Agarwal, MS[1] and **Hong Yu, PhD**[2,3,*]

[1]Medical Informatics, University of Wisconsin-Milwaukee, Milwaukee, WI, USA

[2]Department of Health Sciences, University of Wisconsin-Milwaukee, Milwaukee, WI, USA

[3]Department of Computer Science, University of Wisconsin-Milwaukee, Milwaukee, WI, USA

Abstract

Objective—Hedging is frequently used in both the biological literature and clinical notes to denote uncertainty or speculation. It is important for text-mining applications to detect hedge cues and their scope; otherwise, uncertain events are incorrectly identified as factual events. However, due to the complexity of language, identifying hedge cues and their scope in a sentence is not a trivial task. Our objective was to develop an algorithm that would automatically detect hedge cues and their scope in biomedical literature.

Methodology—We used conditional random fields (CRF), a supervised machine-learning algorithm, to train models to detect hedge cue phrases and their scope in biomedical literature. The models were trained on the publicly available BioScope corpus. We evaluated the performance of the CRF models in identifying hedge cue phrases and their scope by calculating recall, precision and F1-score. We compared our models with three competitive baseline systems.

Results—Our best CRF-based model performed statistically better than the baseline systems, achieving an F1-score of 88% and 86% in detecting hedge cue phrases and their scope in biological literature and an F1-score of 93% and 90% in detecting hedge cue phrases and their scope in clinical notes.

Conclusions—Our approach is robust, as it can identify hedge cues and their scope in both biological and clinical text. To benefit text-mining applications, our system is publicly available as a Java API and as an online application at http://hedgescope.askhermes.org. To our knowledge, this is the first publicly available system to detect hedge cues and their scope in biomedical literature.

Keywords

uncertainty detection; hedge cue detection; text mining; natural language processing

* To whom correspondence should be addressed: Name: Hong Yu, Mailing address: 2400 E Hartford Ave, Room 939, Milwaukee WI 53211, hongyu@uwm.edu, Phone: 414-229-3344, Fax: 414-229-2619.

1. Introduction

Biomedical literature often includes sentences that express uncertainty or speculation, as seen in the following two examples:

(1) *These findings are discussed in relation to possible therapeutic approaches to the immunotherapy of APL*.

(2) *No focal consolidation to suggest pneumonia*.

In sentence (1), the authors indicate that therapeutic approaches for APL (Acute Promyelocytic Leukemia) may be possible using immunotherapy, and the outcomes of their study are presented in relation to this possibility. Similarly, in sentence (2), the report indicates that the patient might be suffering from pneumonia because of the observation that focal consolidation is absent. In our examples, “possible” in sentence (1) and “suggest” in sentence (2) indicate uncertainty and speculation, a linguistic phenomenon known as hedging [1]. Such cue words or phrases are therefore referred to as **hedge cues**.

In addition to the work of Lakoff, Palmer [2] and Chafe [3] studied phenomena related to hedging in the open domain; Palmer focused mainly on the use of modal verbs in hedging, while Chafe looked at the use of such words as “about” and “kind of” to express an imperfect match between knowledge and categories. In the domain of scientific literature, Hyland conducted a comprehensive study on the presence and use of hedge cues [4] and suggested that hedging serves the purpose of weakening the force of statement and signaling uncertainty. Based on exhaustive analysis of a corpus of molecular biology articles, he proposed a pragmatic classification of hedge cues comprising modal auxiliaries, epistemic lexical verbs, epistemic adjectives, adverbs, nouns and other non-lexical cues.

To help researchers discover information from literature, many text-mining applications have been developed, and it is essential for such applications to identify the presence of uncertainty and speculation in text [5]. Since hedged statements are often hypothetical and may lack the proof needed to verify them as factual information, text-mining applications should present the information extracted from such sentences separately from factual information. For instance, guidelines for coding radiology reports state that uncertain diagnoses should never be coded [6].

Despite the importance of this issue, the task of hedge detection is frequently ignored by most of the current biomedical text-mining approaches. Such approaches can be generally classified into three main categories – co-occurrence-based approaches (e.g., [7],[8]), rule-based approaches (e.g., [9],[10]), and statistical and machine learning-based approaches (e.g., [11–17]). None of these approaches detects hedging in text.

Hedge cue detection is not an easy task. Although certain cue phrases, such as ‘could’, ‘appears’, ‘possible’, ‘can’, ‘potential’ and ‘indicate’, are commonly used in hedged statements, identifying hedged statements based merely on the presence of cue terms may lead to false results. Two examples are shown below:

(3) *We can now study regulatory regions and functional domains of the protein in the context of a true erythroid environment, experiments that have not been* ***possible*** *heretofore.*

(4) *If symptoms persist further evaluation would be* ***indicated****.*

In addition to frequent cue phrases, certain cue phrases appear infrequently to indicate uncertainty or speculation. Two examples are shown below:

(5) *The new conserved motifs are H-x3-L-x3-C-R-x-C-G and D-x3-I-h-0050-x2-F-C-x2-C, and their function* ***remains to be determined****.*

(6) *Based on these results we* ***estimate*** *a 5–10% difference in virus production of the LTR variants when compared to that of wild-type.*

In sentence (5), the phrase "remains to be determined" is applied to the function of the two motifs, indicating that their function is unknown. In sentence (6), the authors are uncertain about the actual difference in virus production of the LTR type when compared to the wild-type virus, but they hypothesize the difference to be in the range of 5–10%. A dictionary of cue phrases might not include such infrequent cue phrases, which would affect the recall of the system.

Although detection of hedge cues in a sentence is an important and challenging task in and of itself, it is equally important to determine the scope of the hedge cue, since all observations or reported events in the sentence may not be hedged. This can be seen in the following example sentences where the hedge cue is in boldface and its scope is marked in square brackets:

(7) Thus, the novel enhancer element identified in this study is [**Probably** a target site for both positive and negative factors].

(8) Right middle and [**probable** right lower] lobe pneumonia.

In sentence (7), the authors do not express uncertainty regarding the discovery or novelty of the enhancer element, but they are speculative with respect to its role as a target site for positive and negative factors. Similarly, in sentence (8), the clinician does not hedge on the presence of pneumonia in the right middle lobe but is uncertain about the presence of pneumonia in the right lower lobe. Hence, a system that identifies hedge cues must identify their scope as well; otherwise factual information will also be reported as uncertain information.

Detecting hedge cues and their scope is, therefore, a challenging research task, and we propose that the task of information extraction should address it in addition to relation identification. We report here on the development of a supervised machine-learning system called HedgeScope that detects hedge cues and their scope in biomedical sentences. The next section describes related work, followed by the methods and evaluation.

2. Related work

Most of the studies in the area of detecting hedging in biomedical literature have focused on determining the presence or absence of hedge cues in sentences; the scope of such cues is ignored in most studies. Unlike our study, some studies assign different levels of certainty to the sentence based on the hedge cue.

For example, Friedman and co-workers developed a natural language processing application to identify clinical information in narrative reports and mapped the information into a structured representation containing clinical terms; this system factored the use of hedging in clinical notes [18]. Their system assigned one of five certainty categories to each extracted finding. The five categories were no certainty, low certainty, moderate certainty, high certainty and cannot evaluate. The findings and certainty modifiers were extracted using rules based on semantic grammar.

Light and co-workers manually annotated speculative sentences in Medline abstracts and found that the annotation could be done reliably by humans [19]. In their annotation, the sentences were classified as one of the three categories: definitive, low speculative and high

speculative. A Support Vector Machine (SVM) classifier and sub-string matching technique were used to automatically classify abstract sentences as speculative or definitive sentences. The sub-string matching technique achieved a slightly better accuracy (95%) than the SVM classifier (92%), with a precision and recall of 55% and 79%, respectively. Although the classifiers were able to reliably classify sentences as speculative or definitive sentences, they were unable to achieve a good performance on the task of distinguishing between high speculative and low speculative sentences.

Medlock and Briscoe extended the study of Light et al. [20]. To do so, they defined what comprises a 'hedge instance', annotated a corpus that was made publicly available and trained a weakly supervised machine-learning model using SVM. Light and co-workers' sub-string matching based classifier was used as the baseline system. Medlock and Briscoe's model achieved a recall/precision break-even point (BEP) of 76%, while the baseline system achieved a BEP of 60% on their test set. Medlock and Briscoe's work was subsequently expanded by Medlock [21] in which the use of additional features such as part of speech, lemmas and bigrams was explored to improve the performance of the classifier. The use of part of speech did not impact the performance of the classifier; however, using lemma improved performance to 80% BEP and the use of bigrams improved performance to 82% BEP. Szarvas also extended Medlock and Briscoe's study [22]. He found that radiology reports typically contained unambiguous lexical hedging cues, while multi-word hedge cues were commonly found in scientific articles. Szarvas then developed a maxent-based classifier to classify hedge sentences in both radiology free-text reports and scientific articles. Feature selection for the classifier was done automatically and manually. Keywords from external dictionaries were also added to improve the performance of the classifier. The system was evaluated on Medlock and Briscoe's dataset and obtained a BEP of 85%. Kilicoglu and Bergler developed a classifier that was based on a dictionary of hedge cues which was developed from existing linguistic studies and lexical resources and incorporated syntactic patterns [23]. Their system was tested on two test sets: a test set of 1,537 sentences released by Medlock and Briscoe [20] on which the system achieved a BEP of 85%, and a test set of 1,087 sentences released by Szarvas [22] on which the system achieved a BEP of 82%.

To recognize modal information in biomedical text, Thompson and co-workers collected a list of words and phrases that express modal information [24]. They also proposed a categorization scheme based on the type of information conveyed, and using this scheme, they annotated 202 Medline abstracts. The collected list of modal words and phrases was validated through the annotations. In a study exploring the relationship between sentences that contain citations and hedge sentences, DiMarco and Mercer found that hedging occurs more frequently in the context of citations [25]. Their study also deduced that hedging could be used to classify citations.

Shatkay and co-workers developed a classifier for biomedical text to classify text along five dimensions [13]. One of the dimensions was degree of certainty, according to which the statement could be assigned a value between 0 and 3, with 0 indicating no certainty and 3 indicating absolute certainty. They annotated a corpus of 10,000 sentences and sentence fragments selected from full-text articles from a variety of biomedical journals. An SVM classifier was trained on the annotated sentences to classify the certainty of a statement. To evaluate the performance of the classifier, a five-fold cross validation on the annotated data was performed, and a recall of 99% and precision of 99% was reported.

Uzuner and co-workers [26] developed two systems, ENegEx (Extended NegEx) and StAC (Statistical Assertion Classifier), to determine if medical problems mentioned in clinical narratives are present (positive assertion), absent (negative assertion), uncertain (uncertainty

assertion) or associated with someone other than the patient (alter-association assertion). ENegEx extended NegEx to apply rules to capture whether a medical problem mentioned in clinical narratives is present or absent [27]. NegEx's rule-base has been extended by other applications as well; for example, ConText [28] extended the rule-base to identify features such as temporality and experiencer of a disease in clinical narratives. StAC is a statistical system that uses supervised machine learning algorithm Support Vector Machines (SVM) to determine the assertion class. StAC makes use of lexical and syntactical features for training. It was reported that ENegEx's performance at identifying uncertainty assertions ranged from 1% to 16% F1-score, whereas StAC's performance ranged from 38% to 89% F1-score [26]. Neither the ENegEx system, nor the corpus used for training and evaluating StAC, was publicly available.

Morante and Daelemans [29] developed a two-phase approach to detect the scope of hedge cues in biomedical literature. In the first phase, hedge cues were identified by a set of classifiers, and in the second phase, another set of classifiers was used to detect the scope of the hedge cue. The system performed better than the baseline in identifying hedge cues and their scope. The percentage of correct scopes for abstracts, full-text and clinical articles was 65.55%, 35.92% and 26.21%, respectively.

Most of the systems reported above were developed to detect hedging in either clinical notes or the biomedical literature. In contrast, our system was trained on annotations from a large corpus of both clinical and biomedical texts, and therefore its ability to detect hedging in both the medical and genomics domain is robust. Such a cross-domain hedging detection system will also assist text-mining systems that require the analysis of both clinical data and primary literature, an application example being the clinical question answering system AskHERMES [30],[31] that we are now developing. Furthermore, while the previous systems detect hedging in a sentence, most of them do not detect the scope of hedge cue; as we have found that results detecting hedging with no regard for scope to be misleading, we report on the detection of both phenomena here.

Finally, none of the previous systems is available for general use. To our knowledge, HedgeScope is currently the only implemented system that is publicly available and detects hedge cues and their scope in both the biological literature and clinical notes.

3. Methods

Our systems were built by training the supervised machine-learning algorithms known as conditional random fields (CRF). The systems were trained on a variety of features. We trained our systems on a corpus of hedges, as described below.

3.1. Hedge Corpus

We used the publicly available BioScope corpus [5] for training and for evaluation. The development of the annotation guideline and the annotation process is described in [5]. The BioScope corpus consists of three sub-corpora: abstracts from 1,273 articles used in the GENIA corpus, full-text of nine articles and 1,954 medical free texts. Together, these sub-corpora consist of more than 20,000 sentences, which correspond to approximately 435,000 word tokens.

We first selected all hedge sentences from the three sub-corpora. A hedge sentence is a sentence that contains at least one hedge cue annotation. We counted the number of hedge sentences and then randomly selected an equal number of non-hedge sentences from a pool of all non-hedge sentences. We thus obtained 6,950 sentences with 3,475 hedge sentences and 3,475 non-hedge sentences. We pooled these sentences and randomly divided them into

two groups, one being the training set and the other being the testing set. Hence, both the testing and training sets for hedge sentences contained 3,475 sentences.

We also built training and testing sets specific to biological and clinical sentences. Sentences from the abstract sub-corpus and the full-text sub-corpus were considered to be biological sentences, while medical free-text sentences were considered to be clinical sentences. Hence, there were 2,620 biological hedge sentences and 855 clinical hedge sentences. As in the case of all sentences, we selected an equal number of positive biological and clinical sentences and divided them evenly. Hence both the training and testing sets for biological hedge sentences contained 2,620 sentences and both the training and testing sets for clinical hedge sentences contained 855 sentences.

Besides the test set generated from the BioScope corpus, we also used the test set made publicly available by Medlock and Briscoe [20]. In this corpus, neither hedge cues nor their scope are marked; rather, the sentences are labeled to indicate if they are hedge sentences are not. This test set contains a total of 1,537 sentences with 380 hedge sentences and 1,157 non-hedge sentences.

3.2. Pre-Processing

Before training the models, we preprocessed all sentences in the BioScope training and testing sets by separating punctuation from the word tokens. This was done because a punctuation mark, such as a comma, could indicate the boundary of a clause, and hence could aid in determining the limits of the scope of a particular instance of hedge cue.

3.3. Conditional Random Fields

Conditional random fields are probabilistic models that offer an advantage over the hidden Markov Model (HMM) for sequential data because the independence assumption in HMM can be relaxed in CRF [32]. Studies have shown that CRF models outperformed HMM in NLP tasks including POS tagging [32], information extraction [33] and has shown to be the best ML model for named entity recognition in the biomedical domain [34],[11]. We therefore explored CRFs on hedge scope detection.

We used the open source CRF algorithm implementation provided by the ABNER library to train test models [11]. ABNER was originally developed using the Mallet CRF framework to identify biomedical named entities (e.g., proteins and cell lines) from biological literature. ABNER's library implementation allows users to train their own models as well and hence can be viewed as a library implementing the CRF framework, which was used in the current work.

3.4. Detecting hedge cues

3.4.1. Hedge cue detection using CRF—We first trained a CRF model to identify hedge cues. ABNER, the CRF algorithm implementation that we used to train the model, required that the data be input in a specific manner. To this end, we marked each word in the BioScope corpus to indicate whether it was a part of the hedge cue or not. The first word in the hedge cue was marked with 'B-CUE' to indicate the beginning of a cue, the remaining words in the hedge cue were marked with 'I-CUE' to indicate that they were inside the cue and words that were not a part of the cue were marked with 'O' to indicate that they were outside the cue. If a hedge cue consisted of only one word, then only the beginning marker (B-CUE) was used to mark it. A separate marker was not used to mark the end of the cue phrase. The trained model was used to automatically identify hedge cues in the test sentences by marking the first word with the beginning tag and the remaining words with the

intermediate tag. We call the trained system HedgeCue. We experimented with different strategies and baseline systems, as shown in Table 1.

3.4.2 Baseline system to detect hedge cue—For comparison, we developed a regular expression-based baseline system (BaselineCue, as shown in Table 1) that detects hedge cues. In the training phase, the system automatically extracts hedge cues from the training set. In the testing phase, the system marks a test sentence as a hedge sentence if any of the cue phrases appear in the sentence.

3.5. Detecting scope of a hedge cue

3.5.1. Detection scope of a hedge cue using CRF—We applied CRF models to detect the scope of a hedge cue and marked scope in the same way as the hedge cue was marked. The first word in the scope of the hedge cue was tagged with a beginning tag, while the remaining words within the scope were tagged with an intermediate tag. The words of the cue phrase within the scope were not given any special consideration, and they were treated as any other word within the scope. The trained models were used to identify the scope of hedge cues in the test set.

We observed that the scope of a hedge cue was often a clause containing a hedge cue phrase. We speculate that linguistic features can be useful for hedge scope identification. To this end, we explored POS as learning features for the CRF model. Specifically, we replaced all words except the words of the cue phrase with their corresponding part of speech tags in the training data (Figure 1). We experimented with either replacing the hedge-word with a custom tag ‘CUE’ or retaining the word. In the case of the test set, since the cue phrases were not marked, we used HedgeCue or BaselineCue to identify the hedge cues.

Morante and Daelemans (2009) [29] also developed a supervised machine-learning (ML) model for hedge cue detection. Although they made use of the Bioscope corpus, they limited the data to abstracts only, a small portion of the Bioscope corpus. They first trained on three independent ML classifiers; subsequently, a fourth classifier was built upon the output of the three independent classifiers. They however, did not report the results of each classifier, nor did they report how such a two-tiered model of four ML classifier improved the performance. In contrast, single-classifier-based CRF models have shown success in biomedical named entity recognition [11],[34]. We therefore trained such a single-classifier-based CRF model for hedging cue and scope detection.

3.5.2. Baseline systems for detecting scope of hedge cues—We developed two baseline systems to detect the scope of hedge cues. BaselineScope-1 first applies BaselineCue to mark a hedge cue in a sentence and then marks the scope as the text from the beginning of the identified cue phrase to the first occurrence of a comma or period (Figure 2). BaselineScope-2 marks the scope as the text from the beginning of the identified cue phrase to the first occurrence of a period (Figure 2).

3.6. Evaluation

To evaluate the performance of the systems on the BioScope testing data, we calculate and report the system’s recall, precision and f-1 score. The recall and precision of the systems were calculated as follows:

Recall = True positive count / (True positive count + False negative count)

Precision = True positive count / (True positive count + False positive count)

The system’s F1 score was calculated as the harmonic mean of the recall and precision. We also calculated the system’s accuracy, which is the number of correctly predicted words

divided by the total number of words. For every word in the test sentence, if both the original annotation and tested system marked the word as a part of a cue phrase or scope, then the word was counted as a true positive; if the original annotation only marked the word as a part of the cue phrase, then the word was counted as a false negative; if only the tested system marked the word as a part of the cue phrase, then the word was counted as a false positive; and if neither the original annotation nor the tested system marked the word as a part of the cue phrase, then the word was counted as a true negative. We report the performance of HedgeCue, HedgeScope and the baseline systems.

We also calculated the percentage of the correct scope (PCS) to evaluate the performance of scope predicting systems. If for a sentence, none of the words were marked as false positive or false negative, then we considered that the system had correctly predicted the scope of the sentence. Note that for sentences with no hedging, the system correctly predicted the scope of the sentence only if it indicated that there were no hedge cues or their scope in the sentence.

We split all test sets into 10 equal parts to measure the variance in results. For all results, we report the standard deviation along with the average.

To evaluate the performance of our systems on the test set provided by Medlock and Briscoe [20], we used HedgeCue and BaselineCue to detect the presence of hedge cues in the sentences. The systems were trained on sentences from both the training set and the testing set derived from the BioScope corpus. If the system predicts that the sentence contains a hedge cue, the sentence is marked as a hedge sentence; otherwise it is marked as a non-hedge sentence. We report the recall, precision and F1-score of our systems at detecting the hedge status of sentences in this data set.

We were unable to test our system against other systems or datasets, such as ENegEx [26], StAC [26], Thompson and co-workers' system [24] and Shatkay and co-workers' system [13], as they were not publicly available.

4. Results

We found that the BaselineCue system extracted 197 cue phrases. The performance of HedgeCue and BaselineCue at predicting hedge cues in the clinical sub-corpus, the biological sub-corpus, and the combination of both clinical and biological sub-corpora in BioScope test set is shown in Table 2.

Table 3–Table 5 shows the performance of HedgeScope and BaselineScope systems in predicting the scope of a hedge cue in the BioScope testing set. In Table 3, both biological and clinical sentences were used for training and testing; in Table 4, only biological sentences were used for training and testing; and in Table 5, only clinical sentences were used for training and testing. As defined earlier, the PCS is calculated as the number of sentences for which the scope is correctly identified divided by the total number of sentences. The micro-average of the F1-score of HedgeScope and BaselineScope systems when trained and tested separately on biological or clinical data was 87.14% and 82.48%, respectively. Compared to this, the F1-score of HedgeScope and BaselineScope on all sentences was 86.97% and 80.12%, respectively. Hence, training a dedicated model for biological and clinical data increased the performance by 0.2–2.3% ($p < 0.0001$, two-tailed t-test).

The performance of HedgeCue and BaselineCue at detecting the hedge status of sentences in the test set provided by Medlock and Briscoe [20] is shown in Table 6. The classifiers were

trained on clinical sentences only, biological sentences only, and both clinical and biological sentences. Results for all three training data combinations are shown in Table 6.

5. Discussion

Here, we have developed CRF-based models to predict the hedge cues and their scope in biomedical sentences. We compare these models with baseline systems, which make use of regular expressions and rules to mark the hedge cues and their scope in a sentence. Our results indicate that models using CRF for detection of hedge cue and their scope in biomedical sentences perform better than models based on the use of regular expressions ($p < 0.0001$). Our system can be used to detect hedge cues and their scope in both biological and clinical text.

For the detection of hedge cues, we observed that in the case of biological sentences, the F1-score and accuracy of HedgeCue is better than BaselineCue ($p < 0.0001$); however, the recall of BaselineCue is better than that of HedgeCue ($p < 0.0001$). This is because BaselineCue collects all phrases that have been seen as hedge cues and marks any such phrase in the sentence as a hedge cue, without considering the context in which it appears. Hence, BaselineCue achieves a lower precision than the CRF system, which lowers its F1-score and accuracy. Interestingly, the performance of BaselineCue was better than that of HedgeCue at detecting hedge cues in clinical sentences, as the increase in recall was enough to overcome the decrease in precision. This suggests that the hedge cues in clinical sentences are rarely ambiguous, an observation made earlier by Szarvas [22].

With respect to the task of detecting the scope of hedge cues, we noticed that the micro-average of the F1-score of HedgeScope trained specifically for biological or clinical text was better than the F1-score of the CRF model trained on the combination of biological and clinical text. This is because there are several differences in biological and clinical text. For example, biological sentences from articles published in journals are generally grammatically well-formed, while many sentences from clinical notes are not (e.g. *"Left lower lobe air space disease, atelectasis vs pneumonia."*).

We found that the HedgeScope system (CRF-based) performed better than the BaselineScope system (regular expression based; F1-Score and PCS $p < 0.0001$). In case of biological sentences, a better performance was obtained when the cue phrases were identified using the HedgeCue system, whereas in clinical sentences, a better performance was obtained when the cue phrases were identified using the BaselineCue system. This is in line with the performance of HedgeCue and BaselineCue at detecting hedge cues in clinical and biological sentences.

In analyzing the cases in which HedgeScope did not identify the scope of hedge cues correctly, we found that the errors could be classified into three categories: 1) false positive errors: the model assigns scope where none exists (i.e. it is a non-hedge sentence); 2) false negative errors: the model assigns no scope when one does exist (i.e. it is a hedge sentence); and 3) boundary errors: the model correctly identifies the sentence as a hedge sentence, but it assigns a different scope than that assigned in the testing data. The first category of errors (false positive errors) was observed in 61 of the 3,475 test sentences. In most cases where the model assigned a scope and hedging did not exist, the hedge cue was a common hedge cue phrase, but it did not indicate hedging in the context of that sentence. For example, 'or' was incorrectly predicted to be a hedge cue in the sentence *'Site-directed mutagenesis demonstrated that the two NF-IL-6 motifs could be independently activated by LAM, LPS, or TNF-alpha and that they acted in an orientation-independent manner'*.

The second category of errors (false negative errors) was observed in 135 of the 3,475 test sentences. We found that in most cases in which the model did not assign a scope when such scope existed, the sentence incorporated an infrequent hedge cue. For example, in the sentence *'Reevaluate for renal stones.', 'reevaluate'* was not detected as a hedge cue. Errors occurred in 196 sentences due to error categories 1) and 2). As there are 3,475 sentences in the test data, this indicates that our system achieved an accuracy of 94.36% (F1-score: 94.20%) at predicting the presence of hedging in a sentence.

In the third category of errors, the model correctly identifies the sentence as a hedge sentence, but it assigns a different boundary than that assigned in the testing data. This type of error occurred in 450 sentences. For example, in the following sentence, the correct scope is marked with square brackets and the scope detected by our model is marked with curly brackets: *'Since the IRF-1 gene is both virus and IFN inducible, an intriguing [issue is raised as to {whether the IRF-1 gene is functioning in IFN-mediated regulation of cell growth and differentiation}].'*. In this example, *'issue is raised'* is a hedge cue that our system failed to identify. We found that in most cases our system assigned a smaller scope than the scope assigned for the gold standard sentence.

Despite these errors, our system achieved a strong performance in scope detection, which makes it suitable to be used in conjunction with other text-mining applications in both the biological and clinical domains. We found that HedgeScope was able to identify the correct scope in cases in which the simpler BaselineScope approach failed. Consider the following sentences in which the correct scope is marked by square brackets in the first sentence, and in the second sentence, in which a scope does not exist even though the sentence includes the frequently used hedge cue 'predicted':

- *Interestingly, [Dronc appears to have a substrate specificity that is so far unique among caspases]: while all other known caspases have only been shown to cleave after aspartate residues, Dronc can also cleave after glutamate residues [11].*
- *29 asthma patients with forced expiratory volume in 1 s (FEV1) < 70% predicted were studied.*

For the first example, the BaselineScope system incorrectly marked the scope as "*appears to have a substrate specificity that is so far unique among caspases: while all other known caspases have only been shown to cleave after aspartate residues"* and *"can also cleave after glutamate residues [11]"*, but the entire scope was correctly identified by the HedgeScope model. In the second example, the HedgeScope system did not mark the sentence, as there is no hedging in the sentence, but the BaselineScope system marked the scope from "predicted" to the end of the sentence.

On evaluating the performance of our system on the test data made available by Medlock and Briscoe [20], we noticed that the best performance (F1-score 87.61 %) was obtained by HedgeCue when trained on biological sentences only. A better performance with models trained on biological sentences can be expected because the test set comprises biological sentences. This data set has been used to test other hedge status detection algorithms [20–23]. A BEP of 85% (and hence, an F1-score of 85%), achieved by Szarvas, and Kilicoglu and Berger, is the highest reported performance on this test data. In comparison, our system achieved an F1-score of 87.61%.

A CRF-based approach was used by Morante and Daelemans [29] to identify hedge cues and their scope in biomedical literature. Similar to our approach, Morante and Daelemans' system was also trained on the BioScope data. A comparison of their reported results with our own shows that our system had a better performance than theirs. This could be due to the difference in the training data used; Morante and Daelemans used only the abstract sub-

corpus for training. Surprisingly, our system's overall performance (PCS ~81%) was also better than the performance of Morante and Daelemans' on the abstract sub-corpus (PCS ~66%). This could be due to the difference in the size of the training data or the features used for selection.

Unfortunately, Morante and Daelemans' system is not publicly available, so we were unable to test the performance of their system on the same test sets as our system was tested on.

6. Conclusion and Future Work

We have created several CRF-based models that can automatically predict the hedge cues and their scope in biomedical literature. These models can also be used to predict the hedge status of a target entity in the sentence. The choice of which model to use depends on the task at hand. For predicting the scope of hedge cues in biological sentences, we recommend using a CRF-based model that identifies cue phrases using a CRF-based cue phrase identifier and replaces non-cue phrase words with their parts of speech. However, to predict the scope of hedge cues in clinical sentences, we recommend using the CRF-based model that identifies cue phrases using a regular expression-based cue phrase identifier and replaces non-cue phrase words with their part of speech. Although the recall of our trained system is lower than the recall of the baseline systems, the trained systems achieve a much higher precision than the baseline systems, resulting in a much higher F1-score. The models we have trained perform well in detecting hedge cues and their scope in both biomedical and clinical documents. To our knowledge, this is the first openly available system that predicts the scope of hedge cues in both the biological and clinical domain. An online version of the hedge scope detector is available at http://hedgescope.askhermes.org.

Any annotated corpus has size limitations, and unseen data encountered by a system trained on such a corpus will hurt the system's performance. In future work we may explore methods for automatically identifying hedge cues from a large corpus, including contextual similarity, which is commonly used for identifying semantically related words or synonyms [35],[36]. We may also explore bootstrapping [37] or co-training approaches [38] that partially overcome the limitations of training size.

Acknowledgments

The authors thank Dr. Lamont Antieau for proofreading this manuscript. The authors acknowledge the support from the National Library of Medicine, grant number 1R01LM009836-01A1. Any opinions, findings, or recommendations are those of the authors and do not necessarily reflect the views of the NIH.

References

1. Lakoff G. Hedges: A study in meaning criteria and the logic of fuzzy concepts. Journal of Philosophical Logic 1973 Oct;vol. 2:458–508.
2. Palmer, FR. Mood and Modality. Cambridge, UK: Cambridge University Press; 2001.
3. Chafe, W. Evidentiality in English conversation and academic writing. In: Chafe, W.; Nichols, J., editors. Evidentiality: The Linguistic Coding of Epistemology (Advances in Discourse Processes). Norwood, NJ: Ablex Publishing; 1986. p. 261-272.
4. Hyland, K. Hedging in Scientific Research Articles. Amsterdam, Netherlands: John Benjamins Pub Co; 1998.
5. Szarvas, G., et al. Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing. Columbus, Ohio: Association for Computational Linguistics; 2008. The BioScope corpus: annotation for negation, uncertainty and their scope in biomedical texts; p. 38-45.
6. Moisio, MA. A Guide to Health Insurance Billing. Delmar Cengage Learning: 2006.

7. Bunescu, R.; Mooney, R.; Ramani, A.; Marcotte, E. Proceedings of the Workshop on Linking Natural Language Processing and Biology: Towards Deeper Biological Literature Analysis. New York City, New York: Association for Computational Linguistics; 2006. Integrating co-occurrence statistics with information extraction for robust retrieval of protein interactions from Medline; p. 49-56.
8. Matsuo Y, Ishizuka M. Keyword Extraction from a Single Document using Word Cooccurrence Statistical Information. International Journal on Artificial Intelligence Tools 2004;vol. 13:157–169.
9. Fukuda K, Tamura A, Tsunoda T, Takagi T. Toward information extraction: identifying protein names from biological papers. Pacific Symposium on Biocomputing 1998:707–718. [PubMed: 9697224]
10. Friedman C, Shagina L, Lussier Y, Hripcsak G. Automated encoding of clinical documents based on natural language processing. Journal of the American Medical Informatics Association Sep;vol. 11:392–402.
11. Settles B. ABNER: an open source tool for automatically tagging genes, proteins and other entity names in text. Bioinformatics 2005 Jul;vol. 21:3191–3192. [PubMed: 15860559]
12. Agarwal S, Yu H. Automatically classifying sentences in full-text biomedical articles into Introduction, Methods, Results and Discussion. Bioinformatics 2009 Dec;vol. 25:3174–3180. [PubMed: 19783830]
13. Shatkay H, Pan F, Rzhetsky A, Wilbur WJ. Multi-dimensional classification of biomedical text: Toward automated, practical provision of high-utility text to diverse users. Bioinformatics 2008 Sep;vol. 24:2086–2093. [PubMed: 18718948]
14. Rafkind, B.; Lee, M.; Chang, S.; Yu, H. Proceedings of the Workshop on Linking Natural Language Processing and Biology: Towards Deeper Biological Literature Analysis. New York City, New York: Association for Computational Linguistics; 2006. Exploring text and image features to classify images in bioscience literature; p. 73-80.
15. Müller H, Kenny EE, Sternberg PW. Textpresso: An Ontology-Based Information Retrieval and Extraction System for Biological Literature. PLoS Biol 2004;vol. 2:e309.
16. Donaldson I, et al. PreBIND and Textomy - mining the biomedical literature for protein-protein interactions using a support vector machine. BMC Bioinformatics 2003;4:11. [PubMed: 12689350]
17. Van Auken K, et al. Semi-automated curation of protein subcellular localization: a text mining-based approach to Gene Ontology (GO) Cellular Component curation. BMC Bioinformatics 2009;10:228. [PubMed: 19622167]
18. Friedman C, Alderson PO, Austin JHM, Cimino JJ, Johnson SB. A General Natural-language Text Processor for Clinical Radiology. Journal of the American Medical Informatics Association 1994 Mar;vol. 1:161–174. [PubMed: 7719797]
19. Light, M.; Qiu, XY.; Srinivasan, P. BioLINK 2004, Linking Biological Literature, Ontologies and Databases. Boston, MA, USA: Association for Computational Linguistics; 2004. The Language of Bioscience: Facts, Speculations, and Statements In Between; p. 17-24.
20. Medlock, B.; Briscoe, T. Weakly Supervised Learning for Hedge Classification in Scientific Literature. PROCEEDINGS OF THE 45TH ANNUAL MEETING OF THE ASSOCIATION OF COMPUTATIONAL LINGUISTICS; 2007. p. 992-999.
21. Medlock B. Exploring hedge identification in biomedical literature. J. of Biomedical Informatics 2008;vol. 41:636–654.
22. Szarvas, G. Hedge classification in biomedical texts with a weakly supervised selection of keywords. Proc 46th Meeting of the Association for Computational Linguistics; Columbus, Ohio: Association for Computational Linguistics; 2008. p. 281-289.
23. Kilicoglu H, Bergler S. Recognizing speculative language in biomedical research articles: a linguistically motivated perspective. BMC Bioinformatics 2008;vol. 9 Suppl 11:S10. [PubMed: 19025686]
24. Thompson, P.; Venturi, G.; McNaught, J.; Montemagni, S.; Ananiadou, S. LREC 2008 Workshop on Building and Evaluating Resources for Biomedical Text Mining. Marrakech, Morocco: 2008. Categorising Modality in Biomedical Texts; p. 27-34.

25. DiMarco, C.; Mercer, RE. Computing attitude and affect in text: Theory and applications. Dordrecht: Springer-Verlag; 2005. Hedging in Scientific Articles as a Means of Classifying Citations.
26. Uzuner Ö, Zhang X, Sibanda T. Machine Learning and Rule-based Approaches to Assertion Classification. Journal of the American Medical Informatics Association 2009 Jan;vol. 16:109–115. [PubMed: 18952931]
27. Chapman WW, Bridewell W, Hanbury P, Cooper GF, Buchanan BG. A simple algorithm for identifying negated findings and diseases in discharge summaries. Journal of Biomedical Informatics 2001 Oct;vol. 34:301–310. [PubMed: 12123149]
28. Chapman, WW.; Chu, D.; Dowling, JN. Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing. Prague, Czech Republic: Association for Computational Linguistics; 2007. ConText: an algorithm for identifying contextual features from clinical text; p. 81-88.
29. Morante, R.; Daelemans, W. Proceedings of the Workshop on BioNLP. Boulder, Colorado: Association for Computational Linguistics; 2009. Learning the scope of hedge cues in biomedical texts; p. 28-36.
30. Yu H, Lee M, Kaufman D, Ely J, Osheroff JA, Hripcsak G, Cimino J. Development, implementation, and a cognitive evaluation of a definitional question answering system for physicians. Journal of Biomedical Informatics 2007 Jun;vol. 40:236–251. [PubMed: 17462961]
31. Yu, H.; Cao, Y. Automatically extracting information needs from Ad Hoc clinical questions. AMIA … Annual Symposium Proceedings / AMIA Symposium. AMIA Symposium; 2008. p. 96-100.
32. Lafferty, J.; McCallum, A.; Pereira, F. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. Proceedings of the Eighteenth International Conference on Machine Learning (ICML 2001); Williamstown, MA, USA: 2001. p. 282-289.
33. Pinto, D.; McCallum, A.; Wei, X.; Croft, WB. Table extraction using conditional random fields. Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval; Toronto, Canada: ACM; 2003. p. 235-242.
34. Leaman R, Gonzalez G. BANNER: an executable survey of advances in biomedical named entity recognition. Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing 2008:652–663. [PubMed: 18229723]
35. Dagan, I.; Marcus, S.; Markovitch, S. Contextual word similarity and estimation from sparse data. Proceedings of the 31st annual meeting on Association for Computational Linguistics; Columbus, Ohio: Association for Computational Linguistics; 1993. p. 164-171.
36. Yu H, Agichtein E. Extracting synonymous gene and protein terms from biological literature. Bioinformatics 2003 Jul;vol. 19:340–349.
37. Weiss SM, Kapouleas I. An empirical comparison of pattern recognition, neural nets, and machine learning classification methods. In proceedings of the eleventh international joint conference on artificial intelligence 1989:781–787.
38. Blum, A.; Mitchell, T. Combining labeled and unlabeled data with co-training. Proceedings of the eleventh annual conference on Computational learning theory; Madison, Wisconsin, United States: ACM; 1998. p. 92-100.

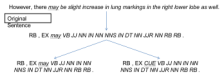

Figure 1.

Example of a sentence used for training after it was replaced with its part of speech tags. The underlined word is the hedge cue in the sentence, while the words in italics represent the scope of the hedge cue. In the first step, all words except the cue word (underlined) were replaced with their part of speech tags. The cue word was either not replaced (bottom left) or replaced with a custom tag “CUE” (bottom right).

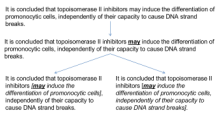

Figure 2.

An example showing the method in which BaselineScope marks the scope of a hedge cue in the sentence. The hedge cue is first identified using BaselineCue. BaselineScope then marks the scope of the hedge cue as the text from the hedge cue to the first comma or period (left), or the first period (right).

Table 1

Systems we explored for detecting hedge cues and their scope

System name	Detects	Features used	Training/Testing algorithm
HedgeCue	Hedge cues	Words	CRF
BaselineCue	Hedge cues	Words	Cue phrase lookup using Regular Expression
HedgeScope	Scope of a hedge cue	Words	CRF
		POS tags Cue phrase words not replaced with POS tags	HedgeCue (CRF) to identify cue Phrases CRF to mark scope
		POS tags Cue phrase words not replaced with POS tags	BaselineCue (regular expression) to identify cue phrases CRF to mark scope
		POS tags Cue phrase words replaced with custom tag ‘CUE’	HedgeCue (CRF) to identify cue phrases CRF to mark scope
		POS tags Cue phrase words replaced with custom tag ‘CUE’	BaselineCue (regular expression) to identify cue phrases CRF to mark scope
BaselineScope	Scope of a hedge cue	Words	BaselineCue to identify cue phrases; scope marked till the first occurrence of a comma or period
		Words	BaselineCue to identify cue phrases; scope marked till the first occurrence of a period

Table 2

Performance of HedgeCue and BaselineCue systems at identifying hedge cue phrases in the BioScope testing set

	Clinical sentences		**Biomedical sentences**		**Both clinical and biomedical sentences**	
	HedgeCue	**BaselineCue**	**HedgeCue**	**BaselineCue**	**HedgeCue**	**BaselineCue**
Recall	88.69 ±0.05	95.5 ±0.02	82.23 ±0.02	94.69 ±0.01	87.22 ±0.01	96.79 ±0.01
Precision	98.79 ±0.01	95.24 ±0.02	94.83 ±0.01	68.83 ±0.02	94.39 ±0.01	71.5 ±0.02
F1-score	93.46 ±0.03	**95.37 ±0.02**	**88.08 ±0.01**	79.71 ±0.01	**90.66 ±0.01**	82.24 ±0.01
Accuracy	98.89 ±0.01	**99.17 ±0.01**	**99.41 ±5.0×10^{-4}**	98.73 ±7.0×10^{-4}	**99.47 ±4.0×10^{-4}**	98.76 ±9.0×10^{-4}

Table 3

Performance of HedgeScope and BaselineScope at predicting the scope of a hedge cue. The systems were trained and tested on sentences from both the biological sub-corpus and clinical sub-corpus of the BioScope corpus

	HedgeScope					**BaselineScope**	
Features Used	**Words**	**Part of speech**	**Part of speech**	**Part of speech**	**Part of speech**	**Words**	**Words**
Cue phrase identified using	-	HedgeCue	HedgeCue	Baseline Cue	Baseline Cue	Baseline Cue	Baseline Cue
Cue phrase replaced	-	No	Yes	No	Yes	-	-
Scope limited by	-	-	-	-	-	Comma and period	Period only
Recall	77.89 ±0.01	84.14 ±0.01	84.5 ±0.01	91.35 ±0.01	91.73 ±0.01	80.96 ±0.01	93.11 ±0.01
Precision	87.38 ±0.01	89.52 ±0.01	89.59 ±0.01	73.48 ±0.02	73.44 ±0.02	76.59 ±0.02	70.31 ±0.02
F1-score	82.36 ±0.01	86.75 ±0.01	**86.97 ±0.01**	81.45 ±0.01	81.57 ±0.01	78.71 ±0.02	80.12 ±0.01
Accuracy	90.2 ±0.01	92.45 ±0.01	**92.56 ±0.01**	87.78 ±0.01	87.83 ±0.01	87.13 ±0.01	86.43 ±0.01
PCS	78.68 ±2.26	81.18 ±1.71	**81.41 ±1.51**	71.6 ±2.20	71.45 ±2.11	66.16 ±2.69	69.64 ±2.34

Table 4

Performance of HedgeScope and BaselineScope at predicting the scope of a hedge cue. The systems were trained and tested on biological sentences from the BioScope corpus

	HedgeScope					**BaselineScope**	
Features Used	**Words**	**Part of speech**	**Part of speech**	**Part of speech**	**Part of speech**	**Words**	**Words**
Cue phrase identified using	-	HedgeCue	HedgeCue	Baseline Cue	Baseline Cue	Baseline Cue	Baseline Cue
Cue phrase replaced	-	No	Yes	No	Yes	-	-
Scope limited by	-	-	-	-	-	Comma and period	Period only
Recall	78.81 ±0.02	82.47 ±0.02	83.91 ±0.02	90.78 ±0.01	91.59 ±0.01	79.24 ±0.02	92.28 ±0.01
Precision	84.82 ±0.01	88.98 ±0.01	88.54 ±0.01	74.46 ±0.04	74.6 ±0.06	77.69 ±0.04	71.96 ±0.05
F1-score	81.7 ±0.02	85.6 ±0.01	**86.16 ±0.01**	81.81 ±0.02	82.23 ±0.03	78.46 ±0.02	80.87 ±0.03
Accuracy	88.92 ±0.01	91.29 ±0.01	**91.54 ±0.01**	87.34 ±0.01	87.58 ±0.02	86.33 ±0.01	86.28 ±0.02
PCS	76.79 ±3.32	**80.0** ±2.27	79.73 ±2.02	70.57 ±3.55	70.23 ±3.04	63.55 ±2.44	68.55 ±2.81

Table 5

Performance of HedgeScope and BaselineScope at predicting the scope of a hedge cue. The systems were trained and tested on sentences from the clinical sub-corpus of the BioScope corpus

	HedgeScope					**BaselineScope**	
Features Used	**Words**	**Part of speech**	**Part of speech**	**Part of speech**	**Part of speech**	**Words**	**Words**
Cue phrase identified using	-	HedgeCue	HedgeCue	Baseline Cue	Baseline Cue	Baseline Cue	Baseline Cue
Cue phrase replaced	-	No	Yes	No	Yes	-	-
Scope limited by	-	-	-	-	-	Comma and period	Period only
Recall	83.17 ±0.05	85.36 ±0.03	82.52 ±0.02	90.33 ±0.01	88.09 ±0.03	86.59 ±0.02	89.49 ±0.02
Precision	89.54 ±0.06	91.6 ±0.02	92.29 ±0.02	90.03 ±0.03	90.73 ±0.02	88.17 ±0.03	85.49 ±0.03
F1-score	86.24 ±0.02	88.37 ±0.02	87.13 ±0.02	90.18 ±0.02	89.39 ±0.02	87.38 ±0.02	87.44 ±0.02
Accuracy	90.93 ±0.02	92.33 ±0.02	91.68 ±0.02	**93.28 ±0.02**	92.86 ±0.02	91.46 ±0.02	91.22 ±0.02
PCS	81.75 ±3.52	83.74 ±3.93	81.53 ±4.87	**85.38 ±3.89**	83.05 ±4.74	80.35 ±5.72	80.47 ±5.34

Table 6

Performance of HedgeCue and BaselineCue at predicting the hedge status of sentences in the test set provided by Medlock and Briscoe

	Clinical sentences		Biomedical sentences		Both clinical and biomedical sentences	
	HedgeCue	**BaselineCue**	**HedgeCue**	**BaselineCue**	**HedgeCue**	**BaselineCue**
Recall	70.26	75.26	92.11	97.89	87.89	97.89
Precision	69.53	68.75	83.53	57.94	81.86	57.67
F1-score	69.90	71.86	**87.61**	72.80	84.77	72.59
Accuracy	85.04	85.43	**93.56**	81.91	92.19	81.72

J Biomed Inform. Author manuscript; available in PMC 2011 December 1.