Detecting Speculations and their Scopes in Scientific Text

Arzucan Özgür
Department of EECS
University of Michigan
Ann Arbor, MI 48109, USA
ozgur@umich.edu

Dragomir R. Radev
Department of EECS and
School of Information
University of Michigan
Ann Arbor, MI 48109, USA
radev@umich.edu

Abstract

Distinguishing speculative statements from factual ones is important for most biomedical text mining applications. We introduce an approach which is based on solving two sub-problems to identify speculative sentence fragments. The first sub-problem is identifying the speculation keywords in the sentences and the second one is resolving their linguistic scopes. We formulate the first sub-problem as a supervised classification task, where we classify the potential keywords as real speculation keywords or not by using a diverse set of linguistic features that represent the contexts of the keywords. After detecting the actual speculation keywords, we use the syntactic structures of the sentences to determine their scopes.

1 Introduction

Speculation, also known as hedging, is a frequently used language phenomenon in scientific articles, especially in experimental studies, which are common in the biomedical domain. When researchers are not completely certain about the inferred conclusions, they use speculative language to convey this uncertainty. Consider the following example sentences from abstracts of articles in the biomedical domain. The abstracts are available at the U.S. National Library of Medicine PubMed web page[1]. The PubMed Identifier (PMID) of the corresponding article is given in parenthesis.

1. *We showed that the Roaz protein bound specifically to O/E-1 by using the yeast two-hybrid system. (PMID: 9151733)*

2. *These data suggest that p56lck is physically associated with Fc gamma RIIIA (CD16) and functions to mediate signaling events related to the control of NK cellular cytotoxicity. (PMID: 8405050)*

The first sentence is definite, whereas the second one contains speculative information, which is conveyed by the use of the word *"suggest"*. While speculative information might still be useful for biomedical scientists, it is important that it is distinguished from the factual information.

Recognizing speculations in scientific text has gained interest in the recent years. Previous studies focus on identifying speculative sentences (Light et al., 2004; Medlock and Briscoe, 2007; Szarvas, 2008; Kilicoglu and Bergler, 2008). However, in many cases, not the entire sentence, but fragments of a sentence are speculative. Consider the following example sentences.

1. *The mature mitochondrial forms of the erythroid and housekeeping ALAS isozymes are predicted to have molecular weights of 59.5 kd and 64.6 kd, respectively. (PMID: 2050125)*

2. *Like RAD9, RAD9B associates with HUS1, RAD1, and RAD17, suggesting that it is a RAD9 paralog that engages in similar biochemical reactions. (PMID: 14611806)*

Both sentences are speculative, since they contain speculative information, which is signaled by the use of the word *"predicted"* in the first sentence and the word *"suggesting"* in the second sentence. The scope of the speculation keyword *"predicted"* in the first sentence spans the entire sentence. Therefore, classifying the sentence as speculative does not cause information loss. However, the scope of the speculation keyword *"suggesting"* in the second sentence applies only to the second clause of the sentence. In other words, only the statement *"RAD9B is a RAD9 paralog that engages in similar biochemical reactions"* is speculative. The statement *"Like RAD9, RAD9B associates with HUS1, RAD1, and RAD17"* conveys factual information. Therefore, classifying

[1] http://www.ncbi.nlm.nih.gov/pubmed/

Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, pages 1398–1407,
Singapore, 6-7 August 2009. ©2009 ACL and AFNLP

the entire sentence as speculative will result in information loss.

In this paper, we aim to go beyond recognizing speculative sentences and tackle the problem of identifying speculative fragments of sentences. We propose an approach which is based on solving two sub-problems: (1) detecting the real speculation keywords, (2) resolving their linguistic scopes in the sentences. As the previous examples demonstrated speculations are signaled with speculation keywords (e.g. *might, suggest, likely, hypothesize, could, predict, and etc.*). However, these keywords are not always used in a speculative context. In other words, they are not always real speculation keywords. Unlike previous approaches which classify sentences as speculative or not, we formulate the problem as classifying the keywords as real speculation keywords or not. We extract a diverse set of features such as linguistic features that represent the context of the keyword and positional features of the sentence in which the keyword occurs. We use these features with Support Vector Machines (SVM) to learn models to classify whether the occurrence of a keyword is in a speculative context or not. After detecting the real speculation keywords, we use the syntactic structures of the sentences to identify their linguistic scopes.

2 Related Work

Although hedging in scientific articles has been studied from a linguistics perspective since the 1990s (e.g. (Hyland, 1998)), it has only gained interest from a natural language processing perspective in the recent years.

The problem of identifying speculative sentences in biomedical articles has been introduced by Light *et al.* (2004). The authors discussed the possible application areas of recognizing speculative language and investigated whether the notion of speculative sentences can be characterized to enable manual annotation. The authors developed two automated systems to classify sentences as speculative or not. The first method is based on substring matching. A sentence is classified as speculative if it contains one of the 14 predefined strings (*suggest, potential, likely, may, at least, in part, possibl, further investigation, unlikely, putative, insights, point toward, promise, propose*). The second method is based on using SVM with bag-of-words features. The substring matching method performed slightly better than the SVM with bag-of-words features approach.

Medlock and Briscoe (2007) extended the work of Light *et al.* (2004) by refining their annotation guidelines and creating a publicly available data set (FlyBase data set) for speculative sentence classification. They proposed a weakly supervised machine learning approach to classify sentences as speculative or not with the aim of minimizing the need for manually labeled training data. Their approach achieved 76% precision/recall break-even point (BEP) performance on the FlyBase data set, compared to the BEP of 60% obtained by Light *et al.*'s (2004) substring matching approach on the same data set. Szarvas (2008) extended the weakly supervised machine learning methodology of Medlock and Briscoe (2007) by applying feature selection to reduce the number of candidate keywords, by using limited manual supervision to filter the features, and by extending the feature representation with bigrams and trigrams. In addition, by following the annotation guidelines of Medlock and Briscoe (2007), Szarvas (2008) made available the BMC Bioinformatics data set, by annotating four full text papers from the open access BMC Bioinformatics website. They achieved a BEP performance of 85.29% and an F-measure of 85.08% on the FlyBase data set. The F-measure performance achieved on the BMC Bioinformatics data set was 74.93% when the FlyBase data set was used for training. Kilicoglu and Bergler (2008) compiled a list of speculation keywords from the examples in (Hyland, 1998) and extended this list by using WordNet (Fellbaum, 1998) and UMLS SPECIALIST Lexicon (McCray et al., 1994). They used manually crafted syntactic patterns to identify speculative sentences and achieved a BEP and an F-measure of 85% on the FlyBase data set and a BEP and an F-measure of 82% on the BMC Bioinformatics data set.

Unlike pervious studies, which treat the problem of identifying speculative language as a sentence classification task, we tackle the more challenging problem of identifying the portions of sentences which are speculative. In other words, we allow a sentence to include both speculative and non-speculative parts. We introduce and evaluate a diverse set of features that represent the context of a keyword and use these features in a supervised machine learning setting to classify the keywords as real speculation keywords or not. Then, we develop a rule-based method to determine their linguistic scopes by considering the keyword-specific features and the syntactic structures of the sentences. To the best of our knowledge, the BioScope corpus (Vincze et al., 2008) is the only available data set that has been annotated for speculative sentence fragments and we report the first results on this corpus.

3 Corpus

The BioScope corpus[2] has been annotated at the token level for speculation keywords and at the sentence level for their linguistic scopes (Vincze et al., 2008). The corpus consists of three sub-corpora: medical free texts (radiology reports), biomedical article abstracts, and biomedical full text articles. In this paper we focus on identifying speculations in scientific text. Therefore, we use the biomedical article abstracts and the biomedical full text articles in our experiments. The statistics (number of documents, number of sentences, and number of occurrences of speculation keywords) for these two sub-corpora are given in Table 1. The scientific abstracts in the BioScope corpus were included from the Genia corpus (Collier et al., 1999). The full text papers consist of five articles from the FlyBase data set and four articles from the open access BMC Bioinformatics website. The sentences in the FlyBase and BMC Bioinformatics data sets were annotated as speculative or not and made available by Medlock and Briscoe (2007) and Szarvas (2008), respectively and have been used by previous studies in identifying speculative sentences (Medlock and Briscoe, 2007; Kilicoglu and Bergler, 2008; Szarvas, 2008). Vincze *et al.* (2008) annotated these full text papers and the Genia abstracts for speculation keywords and their scopes and included them to the BioScope corpus. The keywords were annotated with a minimalist strategy. In other words, the minimal unit that expresses speculation was annotated as a keyword. A keyword can be a single word (e.g. suggest, predict, might) or a phrase (complex keyword), if none of the words constituting the phrase expresses a speculation by itself. For example the phrase *"no evidence of‘"* in the sentence *"Direct sequencing of the viral genomes and reinfection kinetics showed no evidence of wild-type reversion even after prolonged infection with the Tat- virus."* is an example of a complex keyword, since the words forming the phrase can only express speculation together.

Data Set	Documents	Sentences	Hedge Keywords
Abstracts	1273	11871	2694
Full Papers	9	2670	682

Table 1: Summary of the biomedical scientific articles sub-corpora of the BioScope corpus

In contrast to the minimalist strategy followed when annotating the keywords, the annotation of scopes of the keywords was performed by assigning the scope to the largest syntactic unit possible by including all the elements between the keyword and the target word to the scope (in order to avoid scopes without a keyword) and by including the modifiers of the target word to the scope (Vincze et al., 2008). The reader can refer to (Vincze et al., 2008) for the details of the corpus and the annotation guidelines.

The inter-annotator agreement rate was measured as the F-measure of the annotations of the first annotator by considering the annotations of the second one as the gold standard. The agreement rate for speculation keyword annotation is reported as 92.05% for the abstracts and 90.81% for the full text articles and the agreement rate for speculation scope resolution is reported as 94.04% for the abstracts and 89.67% for the full text articles (Vincze et al., 2008). These rates can be considered as the upper bounds for the automated methods proposed in this paper.

4 Identifying Speculation Keywords

Words and phrases such as *"might", "suggest", "likely", "no evidence of", and "remains to be elucidated"* that can render statements speculative are called speculation keywords. Speculation keywords are not always used in speculative context. For instance, consider the following sentences:

1. *Thus, it appears that the T-cell-specific activation of the proenkephalin promoter is mediated by NF-kappa B. (PMID: 91117203)*

2. *Differentiation assays using water soluble phorbol esters reveal that differentiation becomes irreversible soon after AP-1 appears. (PMID: 92088960)*

The keyword *"appears"* in the first sentence renders it speculative. However, in the second sentence, *"appears"* is not used in a speculative context.

[2] Available at: http://www.inf.u-szeged.hu/rgai/bioscope

The first sub-problem that we need to solve in order to identify speculative sentence fragments is identifying the real speculation keywords in a sentence (i.e. the keywords which convey speculative meaning in the sentence). We formulate the problem as a supervised classification task. We extract the list of keywords from the training data which has been labeled for speculation keywords. We match this list of keywords in the unlabeled (test data) and train a model to classify each occurrence of a keyword in the unlabeled test set as a real speculation keyword or not. The challenge of the task can be demonstrated by the following statistics from the Genia Abstracts of the BioScope corpus. There are 1273 abstracts in the corpus. There are 138 unique speculation keywords and the total number of their occurrence in the abstracts is 6125. In only 2694 (less than 50%) of their occurrences they are used in speculative context (i.e., are real speculation keywords).

In this study we focus on identifying the features that represent the context of a speculation keyword and use SVM with linear kernel (we used the SVM^{light} package (Joachims, 1999)) as our classification algorithm. The following subsection describes the set of features that we propose.

4.1 Feature Extraction

We introduce a set of diverse types of features including keyword specific features such as the stem and the part-of-speech (POS) of the keyword, and keyword context features such as the words surrounding the keyword, the dependency relation types originating at the keyword, the other keywords that occur in the same sentence as the keyword, and positional features such as the section of the paper in which the keyword occurs. While designing the features, we were inspired by studies on other natural language processing problems such as Word Sense Disambiguation (WSD) and summarization. For example, machine learning methods with features based on part-of-speech tags, word stems, surrounding and co-occurring words, and dependency relationships have been successfully used in WSD (Montoyo et al., 2005; Ng and Lee, 1996; Dligach and Palmer, 2008) and positional features such as the position of a sentence in the document have been used in text summarization (e.g. (Radev et al., 2004)).

4.1.1 Keyword Features

Statistics from the BioScope corpus suggest that different keywords have different likelihoods of being used in a speculative context (Vincze et al., 2008). For example, the keyword *"suggest"* has been used in a speculative context in all its occurrences in the abstracts and in the full papers. On the other hand, *"appear"* is a real speculation keyword in 86% of its occurrences in the abstracts and in 83% of its occurrences in the full papers, whereas *"can"* is a real speculation keyword in 12% of its occurrences in the abstracts and in 16% of its occurrences in the full papers. POS of a keyword might also play a role in determining whether it is a real speculation keyword or not. For example, consider the keyword *"can"*. It is more likely to have been used in a speculative context when it is a modal verb, than when it is a noun. Based on these observations, we hypothesize that features specific to a keyword such as the keyword itself, the stem of the keyword, and the POS of the keyword might be useful in discriminating the speculative versus non-speculative use of it. We use Porter's Stemming Algorithm (Porter, 1980) to obtain the stems of the keywords and Stanford Parser (de Marneffe et al., 2006) to get the POS of the keywords. If a keywords consists of multiple words, we use the concatenation of the POS of the words constituting the keyword as a feature. For example, the extracted POS feature for the keywords *"no evidence"* and *"no proof"* is *"DT.NN"*
.

4.1.2 Dependency Relation Features

Besides the occurrence of a speculation keyword, the syntactic structure of the sentence also plays an important role in characterizing speculations. Kilicoglu and Bergler (2008) showed that manually identified syntactic patterns are effective in classifying sentences as speculative or not. They identified that, while some keywords do not indicate hedging when used alone, they might act as good indicators of hedging when used with a clausal complement or with an infinitival clause. For example, the *"appears"* keyword in the example sentences, which are given in the beginning of Section 4, is not a real speculation keyword in the second example *"...soon after AP-1 appears."*
, whereas it is a real speculation keyword in the first example, where it is used with a *that* clausal complement *"...it appears that..."*. Similarly, *"appears"* is used in a speculative context in the fol-

lowing sentence, where it is used with an infinitival clause: *"Synergistic transactivation of the BMRF1 promoter by the Z/c-myb combination appears to involve direct binding by the Z protein."*.

Another observation is that, some keywords act as real speculation keywords only when used with a negation. For example, words such as *"know", "evidence", and "proof"* express certainty when used alone, but express a speculation when used with a negation (e.g., *"not known", "no evidence", "no proof"*).

Auxiliaries in verbal elements might also give clues for the speculative meaning of the main verbs. Consider the example sentence: *"Our findings may indicate the presence of a reactivated virus hosted in these cells."*. The modal auxiliary *"may"* acts as a clue for the speculative context of the main verb *"indicate"*.

We defined boolean features to represent the syntactic structures of the contexts of the keywords. We used the Stanford Dependency Parser (de Marneffe et al., 2006) to parse the sentences that contain a candidate speculation keyword and extracted the following features from the dependency parse trees.

Clausal Complement: A Boolean feature which is set to 1, if the keyword has a child which is connected to it with a clausal complement or infinitival clause dependency type.

Negation: A Boolean feature which is set to 1, if the keyword (1) has a child which is connected to it with a negation dependency type (e.g. "not known": "not" is a child of "known", and the Stanford Dependency Type connecting them is "neg") or (2) the determiner "no" is a child of the keyword (e.g., "no evidence": "no" is a child of "evidence" and the Stanford Dependency Type connecting them is "det").

Auxiliary: A Boolean feature which is set to 1, if the keyword has a child which is connected to it with an auxiliary dependency type (e.g. "may indicate": "may" is a child of "indicate", and the Stanford Dependency Type connecting them is "aux").

If a keyword consists of multiple-words, we examine the children of the word which is the ancestor of the other words constituting the keyword. For example, "no evidence" is a multi-word keyword, where "evidence" is the parent of "no". Therefore, we extract the dependency parse tree features for the word "evidence".

4.1.3 Surrounding Words

Recent studies showed that using machine learning with variants of the "bag-of-words" feature representation is effective in classifying sentences as speculative vs. non-speculative (Light et al., 2004; Medlock and Briscoe, 2007; Szarvas, 2008). Therefore, we also decided to include bag-of-words features that represent the context of the speculation keyword. We extracted the words surrounding the keyword and performed experiments both with and without stemming, and with window sizes of one, two, and three. Consider the sentence: *"Our findings may indicate the presence of a reactivated virus hosted in these cells."*. The bag-of-words features for the keyword "indicate", when a window size of three and no stemming is used are: *"our", "findings", "may", "indicate", "the", "presence", "of"*. In other words, the feature set consists of the keyword, the three words to the left of the keyword, and the three words to the right of the keyword.

4.1.4 Positional Features

Different parts of a scientific article might have different characteristics in terms of the usage of speculative language. For example, Hyland (1998) analyzed a data set of molecular biology articles and reported that the distribution of speculations is similar between abstracts and full text articles, whereas the Results and Discussion sections tend to contain more speculative statements compared to the other sections (e.g. Materials and Methods or Introduction and Background sections). The analysis of Light *et al.* (2004) showed that the last sentence of an abstract is more likely to be speculative than non-speculative.

For the scientific abstracts data set, we defined the following boolean features to represent the position of the sentence the keyword occurs in. Our intuition is that titles and the first sentences in the abstract tend to be non-speculative, whereas the last sentence of the abstract tends to be speculative.

Title: A Boolean feature which is set to 1, if the keyword occurs in the title.

First Sentence: A Boolean feature which is set to 1, if the keyword occurs in the first sentence of the abstract.

Last Sentence: A Boolean feature which is set to 1, if the keyword occurs in the last sentence of the abstract.

For the scientific full text articles data set, we defined the following features that represent the position of the sentence in which the keyword occurs. Our assumption is that the "Results and Discussion" and the "Conclusion" sections tend to contain more speculative statements than the "Materials and Methods" and "Introduction and Background" sections. We also assume that figure and table legends are not likely to contain speculative statements.

Title: A Boolean feature which is set to 1, if the keyword occurs in the title of the article, or in the title of a section or sub-section.

First Sentence: A Boolean feature which is set to 1, if the keyword occurs in the first sentence of the abstract.

Last Sentence: A Boolean feature which is set to 1, if the keyword occurs in the last sentence of the abstract.

Background: A Boolean feature which is set to 1, if the keyword occurs in the Background or Introduction section.

Results: A Boolean feature which is set to 1, if the keyword occurs in the Results or in the Discussion section.

Methods: A Boolean feature which is set to 1, if the keyword occurs in the Materials and Methods section.

Conclusion: A Boolean feature which is set to 1, if the keyword occurs in the Conclusion section.

Legend: A Boolean feature which is set to 1, if the keyword occurs in a table or figure legend.

4.1.5 Co-occurring Keywords

Speculation keywords usually co-occur in the sentences. Consider the sentence: *"We, therefore, wished to determine whether T3SO4 could mimic the action of thyroid hormone in vitro."*. Here, *"whether"* and *"could"* are speculation keywords and their co-occurence might be a clue for their speculative context. Therefore, we decided to include the co-occurring keywords to the feature set of a keyword.

5 Resolving the Scope of a Speculation

After identifying the real speculation keywords, the next step is determining their scopes in the sentences, so that the speculative sentence fragments can be detected. Manual analysis of sample sentences from the BioScope corpus and their parse trees suggests that the scope of a keyword can be characterized by its part-of-speech and the syntactic structure of the sentence in which it occurs. Consider the example sentence whose parse tree is shown in Figure 1. The sentence contains three speculation keywords, "or" and two occurrences of "might". The scope of the conjunction "or", extends to the "VP" whose children it coordinates. In other words, the scope of "or" is "[might be one of the earliest crucial steps in the lysis of normal and dex-resistant CEM cells, *or* might serve as a marker for the process]". Here, "or" conveys a speculative meaning, since we are not certain which of the two sub-clauses (sub-clause 1: [might be one of the earliest crucial steps in the lysis of normal and dex-resistant CEM cells] *or* sub-clause 2: [might serve as a marker for the process]) is correct. The scope of both occurrences of the modal verb "might" is the parent "VP". In other words, the scope of the first occurrence of "might" is "[*might* be one of the earliest crucial steps in the lysis of normal and dex-resistant CEM cells]" and the scope of the second occurrence of "might" is "[*might* serve as a marker for the process]". By examining the keywords, sample sentences and their syntactic parse trees we developed the following rule-based approach to resolve the scopes of speculation keywords. The examples given in this section are based on the syntactic structure of the Penn Tree Bank. But, the rules are generic (e.g. "the scope of a verb followed by an infinitival clause, extends to the whole sentence").

The scope of a conjunction or a determiner (e.g. or, and/or, vs) is the syntactic phrase to which it is attached. For example, the scope of "or" in Figure 1 is the "VP" immediately dominating the "CC".

The scope of a modal verb (e.g. may, might, could) is the "VP" to which it is attached. For example, the scope of "might" in Figure 1 is the "VP" immediately dominating the "MD".

The scope of an adjective or an adverb starts with the keyword and ends with the last token of the highest level "NP" which dominates the adjective or the adverb. Consider the sentence "The endocrine events that are rapidly expressed (seconds) are due to a [possible interaction with cellular membrane]." The scope of the speculation keyword "possible" is enclosed in rectangular brackets. The sub-tree that this scope maps to is: "(NP (NP (DT a) (JJ possible) (NN interaction)) (PP (IN with) (NP (JJ cellular) (NN membrane))))". If there does not exist a "NP" dominating the adverb or adjective keyword, the scope extends to the whole sentence. For example the scope of the speculation adverb "probably" in the sentence "[The remaining portion of the ZFB motif was probably lost in TPases of insect Transib transposons]" is the whole sentence.

The scope of a verb followed by an infinitival

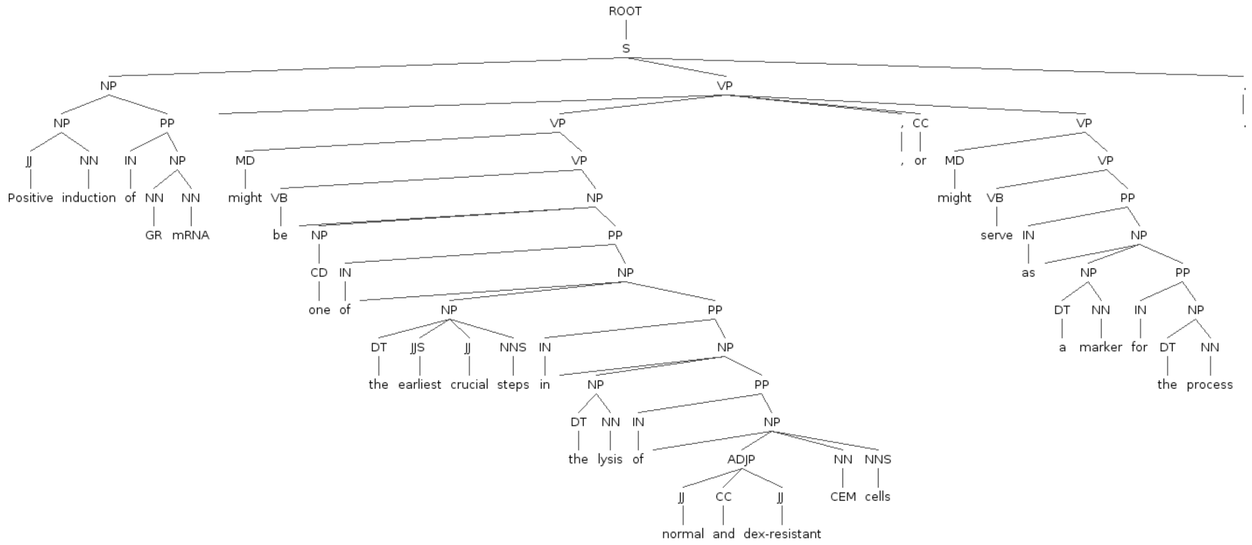
ROOT
S
NP VP .
NP PP VP , CC VP .
JJ NN IN NP MD VP , or MD VP
Positive induction of NN NN might VB NP might VB PP
GR mRNA be NP PP serve IN NP
CD IN NP as NP PP
one of NP PP DT NN IN NP
DT JJS JJ NNS IN NP a marker for DT NN
the earliest crucial steps in NP PP the process
DT NN IN NP
the lysis of ADJP NN NNS
JJ CC JJ CEM cells
normal and dex-resistant

Figure 1: The syntactic parse tree of the sentence "*Positive induction of GR mRNA might be one of the earliest crucial steps in the lysis of normal and dex-resistant CEM cells, or might serve as a marker for the process.*"

clause extends to the whole sentence. For example, the scope of the verb "appears" followed by the "to" infinitival clause is the whole sentence in "[The block of pupariation appears to involve signaling through the adenosine receptor (AdoR)]".

The scope of a verb in passive voice extends to the whole sentence such as the scope of "suggested" in "[The existence of such an independent mechanism has also been suggested in mammals]".

If none of the above rules apply, the scope of a keyword starts with the keyword and ends at the end of the sentence (or clause). An example is the scope of "suggested" in "This [suggested that there is insufficient data currently available to determine a reliable ratio for human]".

6 Evaluation

We evaluated our approach on two different types of scientific text from the biomedical domain, namely the scientific abstracts sub-corpus and the full text articles sub-corpus of the BioScope corpus (see Section 3). We used stratified 10-fold cross-validation to evaluate the performance on the abstracts. In each fold, 90% of the abstracts are used for training and 10% are used to test. To facilitate comparison with future studies the PubMed Identifiers of the abstracts that we used as a test set in each fold are provided[3]. The full text papers sub-corpus consists of nine articles. We used leave-one-out cross-validation to evaluate the performance on the full text papers. In each iteration eight articles are used for training and one article is used to test. We report the average results over the runs for each data set.

6.1 Evaluation of Identifying Speculation Keywords

To classify whether the occurrence of a keyword is in speculative context or not, we built linear SVM models by using various combinations of the features introduced in Section 4.1. Tables 2 and 3 summarize the results obtained for the abstracts and the full text papers, respectively. *BOW N* is the bag-of-words features obtained from the words surrounding the keyword (see Section 4.1.3). *N* is the window size. We experimented both with the stemmed and non-stemmed versions of this feature type. The non-stemmed versions performed slightly better than the stemmed versions. The reason might be due to the different likelihoods of being used in a speculative context of different inflected forms of words. For example, consider the words "appears" and "appearance". They have the same stems, but "appearance" is less likely to be a real speculation keyword than "appears". Another observation is that, decreasing the window size led to improvement in performance. This suggests that the words right before and right after the candidate speculation keyword are more effective in distinguishing its speculative vs. non-speculative context compared to a wider local context. Wider local context might create sparse data and degrade

[3] http://belobog.si.umich.edu/clair/bioscope/

performance. Consider the example, “it appears that TP53 interacts with AR”. The keyword “appears”, and BOW1 (“it” and “that”) are more relevant for the speculative context of the keyword than “TP53”, “interacts”, and “with”. Therefore, for the rest of the experiments we used the *BOW 1* version, i.e., the non-stemmed surrounding bag-of-words with window size of 1. *KW* stands for the keyword specific features, i.e., the keyword, its stem, and its part-of-speech (discussed in Section 4.1.1). *DEP* stands for the dependency relation features (discussed in Section 4.1.2). *POS* stands for the positional features (discussed in Section 4.1.4) and *CO-KW* stands for the co-occurring keywords feature (discussed in Section 4.1.5).

Our results are not directly comparable with the prior studies about identifying speculative sentences (see Section 2), since we attempted to solve a different problem, which is identifying speculative parts of sentences. Only the substring matching approach that was introduced in (Light et al., 2004) could be adapted as a keyword classification task, since the substrings are keywords themselves and we used this approach as a baseline in the keyword classification sub-problem. We compare the performances of our models with two baseline methods, which are based on the substring matching approach. Light *et al.* (2004) have shown that the substring matching method with a predefined set of 14 strings performs slightly better than an SVM model with bag-of-words features in classifying sentences as speculative vs. non-speculative (see Section 2). In baseline 1, we use the 14 strings identified in (Light et al., 2004) and classify all the keywords in the test set that match any of them as real speculation keywords. Baseline 2 is similar to baseline 1, with the difference that rather than using the set of strings in (Light et al., 2004), we extract the set of keywords from the training set and classify all the words (or phrases) in the test set that match any of the keywords in the list as real speculation keywords.

Baseline 1 achieves high precision, but low recall. Whereas, baseline 2 achieves high recall in the expense of low precision. All the SVM models in Tables 2 and 3 achieve more balanced precision and recall values, with F-measure values significantly higher than the baseline methods. We start with a model that uses only the keyword-specific features (KW). This type of feature alone achieved a significantly better performance than the baseline methods (90.61% F-measure for the abstracts and 80.57% F-measure for the full text papers), suggesting that the keyword-specific features are important in determining its speculative context. We extended the feature set by including the dependency relation (DEP), surrounding words (BOW 1), positional (POS), and co-occurring keywords (CO-KW) features. Each new type of included feature improved the performance of the model for the abstracts. The best F-measure (91.69%) is achieved by using all the proposed types of features. This performance is close to the upper bound, which is the human inter-annotator agreement F-measure of 92.05%.

Including the co-occurring keywords to the feature set for full text articles slightly improved precision, but deceased recall, which led to lower F-measure. The best F-measure (82.82%) for the full text articles is achieved by using all the feature types except the co-occurring keywords. The achieved performance is significantly higher than the baseline methods, but lower than the human inter-annotator agreement F-measure of 90.81%. The lower performance for the full text papers might be due to the small size of the data set (9 full text papers compared to 1273 abstracts).

Method	Recall	Precision	F-Measure
Baseline 1	52.84	92.71	67.25
Baseline 2	97.54	43.66	60.30
BOW 3 - stemmed	81.47	92.36	86.51
BOW 2 - stemmed	81.56	93.29	86.97
BOW 1 - stemmed	83.08	93.83	88.05
BOW 3	82.58	92.04	86.98
BOW 2	82.77	92.74	87.41
BOW 1	83.27	93.67	88.10
KW: kw, kw-stem, kw-pos	88.62	92.77	90.61
KW, DEP	88.77	92.67	90.64
KW, DEP, BOW 1	88.46	94.71	91.43
KW, DEP, BOW 1, POS	88.16	95.21	91.50
KW, DEP, BOW 1, POS, CO-KW	88.22	95.56	91.69

Table 2: Results for the Scientific Abstracts

Method	Recall	Precision	F-Measure
Baseline 1	33.77	86.75	47.13
Baseline 2	88.22	52.57	64.70
BOW 3 - stemmed	70.79	83.88	76.58
BOW 2 - stemmed	72.31	85.49	78.11
BOW 1 - stemmed	73.49	84.35	78.41
BOW 3	70.54	82.56	75.88
BOW 2	71.52	85.93	77.94
BOW 1	73.72	86.27	79.43
KW: kw, kw-stem, kw-pos	75.21	87.08	80.57
KW, DEP	75.02	89.49	81.53
KW, DEP, BOW 1	76.15	89.54	82.27
KW, DEP, BOW 1, POS	76.17	90.81	82.82
KW, DEP, BOW 1, POS, CO-KW	75.76	90.82	82.58

Table 3: Results for the Scientific Full Text Papers

6.2 Evaluation of Resolving the Scope of a Speculation

We compared the proposed rule-based approach for scope resolution with two baseline methods. Previous studies classify sentences as speculative or not, therefore implicitly assigning the scope of a speculation to the whole sentence (Light et al., 2004; Medlock and Briscoe, 2007; Szarvas, 2008; Kilicoglu and Bergler, 2008). Baseline 1 follows this approach and assigns the scope of a speculation keyword to the whole sentence. Szarvas (2008) suggest assigning the scope of a keyword from its occurrence to the end of the sentence. They state that this approach works accurately for clinical free texts, but no any results are reported (Szarvas, 2008). Baseline 2 follows the approach proposed in (Szarvas, 2008) and assigns the scope of a keyword to the fragment of the sentence that starts with the keyword and ends at the end of the sentence. Table 4 summarizes the accuracy results obtained for the abstracts and the full text papers.

The poor performance of baseline 1, emphasizes the importance of detecting the portions of sentences that are speculative, since less than 5% of the sentences that contain speculation keywords are entirely speculative. Classifying the entire sentences as speculative or not leads to loss in information for more than 95% of the sentences. The rule-based method significantly outperformed the two baseline methods, indicating that the part-of-speech of the keywords and the syntactic parses of the sentences are effective in characterizing the speculation scopes.

Method	Accuracy-Abstracts	Accuracy-Full text
Baseline 1	4.82	4.29
Baseline 2	67.60	42.82
Rule-based method	79.89	61.13

Table 4: Scope resolution results

7 Conclusion

We presented an approach to identify speculative sentence fragments in scientific articles. Our approach is based on solving two sub-problems. The first one is identifying the keywords which are used in speculative context and the second one is determining the scopes of these keywords in the sentences. We evaluated our approach for two types of scientific texts, namely abstracts and full text papers from the BioScope corpus.

We formulated the first sub-problem as a supervised classification task, where the aim is to learn models to classify the candidate speculation keywords as real speculation keywords or not. We focused on identifying different types of linguistic features that capture the contexts of the keywords. We achieved a performance which is significantly better than the baseline methods and comparable to the upper-bound, which is the human inter-annotator agreement F-measure.

We hypothesized that the scope of a speculation keyword can be characterized by its part-of-speech and the syntactic structure of the sentence and developed rules to map the scope of a keyword to the nodes in the syntactic parse tree. We achieved a significantly better performance compared to the baseline methods. The considerably lower performance of the baseline of assigning the scope of a speculation keyword to the whole sentence indicates the importance of detecting speculative sentence portions rather than classifying the entire sentences as speculative or not.

Acknowledgements

This work was supported in part by the NIH Grant U54 DA021519 to the National Center for Integrative Biomedical Informatics.

References

Nigel Collier, Hyun S. Park, Norihiro Ogata, Yuka Tateishi, Chikashi Nobata, Tomoko Ohta, Tateshi Sekimizu, Hisao Imai, Katsutoshi Ibushi, and Jun I. Tsujii. 1999. The GENIA project: corpus-based knowledge acquisition and information extraction from genome research papers. In *Proceedings of the ninth conference on European chapter of the Association for Computational Linguistics*, pages 271–272. Association for Computational Linguistics.

Marie-Catherine de Marneffe, Bill MacCartney, and Christopher D. Manning. 2006. Generating Typed Dependency Parses from Phrase Structure Parses. In *Proceedings of LREC-06.*

Dmitriy Dligach and Martha Palmer. 2008. Novel Semantic Features for Verb Sense Disambiguation. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies.*

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database (Language, Speech, and Communication).* The MIT Press.

Ken Hyland. 1998. *Hedging in Scientific Research Articles.* John Benjamins Publishing Co.

T. Joachims, 1999. *Advances in Kernel Methods-Support Vector Learning*, chapter Making Large-Scale SVM Learning Practical. MIT-Press.

Halil Kilicoglu and Sabine Bergler. 2008. Recognizing speculative language in biomedical research articles: a linguistically motivated perspective. *BMC Bioinformatics*, 9(Suppl 11).

Marc Light, Xin Ying Qiu, and Padmini Srinivasan. 2004. The language of bioscience: Facts, speculations, and statements in between. In Lynette Hirschman and James Pustejovsky, editors, *HLT-NAACL 2004 Workshop: BioLINK 2004, Linking Biological Literature, Ontologies and Databases*, pages 17–24, Boston, Massachusetts, USA, May 6. Association for Computational Linguistics.

A. T. McCray, S. Srinivasan, and A. C. Browne. 1994. Lexical methods for managing variation in biomedical terminologies. *Proc Annu Symp Comput Appl Med Care*, pages 235–239.

Ben Medlock and Ted Briscoe. 2007. Weakly supervised learning for hedge classification in scientific literature. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 992–999, Prague, Czech Republic, June. Association for Computational Linguistics.

Andres Montoyo, Armando Suarez, German Rigau, and Manuel Palomar. 2005. Combining knowledge- and corpus-based word-sense-disambiguation methods. *Journal of Artificial Intelligence Research*, 23:299–330.

H. T. Ng and H. B Lee. 1996. Integrating multiple knowledge sources to disambiguate word senses: An exemplar-based approach. In *Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics*.

M. F. Porter. 1980. An algorithm for suffix stripping. *Program*, 3(14):130–137.

Dragomir R. Radev, Timothy Allison, Sasha Blair-Goldensohn, John Blitzer, Arda Celebi, Stanko Dimitrov, Elliott Drabek, Ali Hakim, Wai Lam, Danyu Liu, Jahna Otterbacher, Hong Qi, Horacio Saggion, Simone Teufel, Adam Winkel, and Zhang Zhu. 2004. Mead - a platform for multidocument multilingual text summarization. In *Proceedings of LREC 2004*.

Gyorgy Szarvas. 2008. Hedge classification in biomedical texts with a weakly supervised selection of keywords. In *ACL 2008*.

Veronika Vincze, Gyorgy Szarvas, Richard Farkas, Gyorgy Mora, and Janos Csirik. 2008. The BioScope corpus: biomedical texts annotated for uncertainty, negation and their scopes. *BMC Bioinformatics*, 9(Suppl 11).