Learning to Detect Hedges and their Scope Using CRF

Qi Zhao, Chengjie Sun, Bingquan Liu, Yong Cheng
Harbin Institute of Technology, HIT
Harbin, PR China
{qzhao, cjsun, liubq, ycheng}@insun.hit.edu.cn

Abstract

Detecting speculative assertions is essential to distinguish the facts from uncertain information for biomedical text. This paper describes a system to detect hedge cues and their scope using CRF model. HCDic feature is presented to improve the system performance of detecting hedge cues on BioScope corpus. The feature can make use of cross-domain resources.

1 Introduction

George Lakoff (1972) first introduced linguistic hedges which indicate that speakers do not back up their opinions with facts. Later other linguists followed the social functions of hedges closely. Interestingly, Robin Lakoff (1975) introduces that hedges might be one of the "women's language features" as they have higher frequency in women's languages than in men's.

In the natural language processing domain, hedges are very important, too. Along with the rapid development of computational and biological technology, information extraction from huge amount of biomedical resource becomes more and more important. While the uncertain information can be a noisy factor sometimes, affecting the performance of information extraction. Biomedical articles are rich in speculative, while 17.70% of the sentences in the abstracts section of the BioScope corpus and 19.44% of the sentences in the full papers section contain hedge cues (Vincze et al., 2008). In order to distinguish facts from uncertain information, detecting speculative assertions is essential in biomedical text.

Hedge detection is paid attention to in the biomedical NLP field. Some researchers regard the problem as a text classification problem (a sentence is speculative or not) using simple machine learning techniques. Light et al. (2004) use substring matching to annotate speculation in biomedical text. Medlock and Briscoe (2007) create a hedging dataset and use an SVM classifier and get to a recall/precision Break-Even Point (BEP) of 0.76. They report that the POS feature performs badly, while lemma feature works well. Szarvas (2008) extends the work of Medlock and Briscoe with feature selection, and further improves the result to a BEP of 0.85 by using an external dictionary. Szarvas concludes that scientific articles contain multiword hedging cues more commonly, and the portability of hedge classifiers is limited. Halil Kilicoglu and Sabine Bergler (2008) propose an algorithm to weight hedge cues, which are used to evaluate the speculative strength of sentences. Roser Morante and Walter Daelemans (2009) introduce a metalearning approach to process the scope of negation, and they identify the hedge cues and their scope with a CRF classifier based on the original work. They extract a hedge cues dictionary as well, but do not combine it with the CRF model.

In the CoNLL-2010 shared task (Farkas et al., 2010), there are two subtasks for worldwide participants to choose:

- Task 1: learning to detect sentences contain-ing uncertainty.
- Task 2: learning to resolve the in-sentence scope of hedge cues.

This paper describes a system using CRF model for the task, which is partly based on Roser Morante and Walter Daelemans' work.

2 Hedges in the training dataset of BioScope and Wikipedia Corpus

Two training datasets, the BioScope and Wikipedia corpus are provided in the CoNLL-2010 shared task. BioScope consists of two parts, full articles and abstracts collected from biomedical papers. The latter is analyzed for having larger scale and more information of hedges.

In Table 1, the percentage of the speculative sentences in the abstracts section of BioScope corpus is the same as Vincze et al. (2008) reported. We can estimate 1.28 cue words per sentence, meaning that each sentence usually just has one hedge cue. The statistics in Table 1 also

Proceedings of the Fourteenth Conference on Computational Natural Language Learning: Shared Task, pages 100–105,
Uppsala, Sweden, 15-16 July 2010.

indicate that a hedge cue appears 26.7 times on average.

Dataset	ITEM	#
Abstracts of BioScope	Sentences	11871
	Certain sentences	9770
	Uncertain sentences	2101 (17.7%)
	Hedge cues	2694
	cues# per sentence	1.28
	Different hedge cues	143
	Max length of the cues	4
Wikipedia	Sentences	11111
	Certain sentences	8627
	Uncertain sentences	2484 (22.4%)
	weasel cues	3133
	Different weasel cues	1984
	Max length of the cues	13 words

Table 1: Statistics about the abstracts section of the BioScope corpus and Wikipedia corpus.

We extract all the hedge cues from the abstracts section of BioScope corpus, getting 143 different hedge cues and 101 cues with ignoring morphological changes. The maximum length of the cues is 4, with 1.44 words per hedge cue. This suggests that most hedge cues happen to be a single word. We assume that hedge cues set is a limited one in BioScope corpus. Most hedge cues could be identified if the known dataset of hedge cues is large enough. The cue words collected from the BioScope corpus play an important role in the speculative sentences detection.

In contrast to the biomedical abstracts, the weasel cues on Wikipedia corpus make a little difference. Most weasel cues consist of more than one word, and usually appear once. This leads to different results in our test.

A hedge cue word may appear in the non-speculative sentences. Occurrences of the four typical words in speculative and non-speculative sentences are counted.

As shown in Table 2, the cue words can be divided into two classes generally. The hedge cue words “feel” and “suggesting”, which are grouped as one class, only act as hedge cues with never appearing in the non-speculative sentences. While “may” and “or” appear both in the speculative and non-speculative sentences, which are regard as the other one. Moreover, we treat the words “may” and “or” in the same class differently, while “may” is more likely to be a hedge cue than “or”. The treatment is also unequal between “feel” and “suggesting”. In the training datasets, the non-S#/S# ratio can give a weight to distinguish the words in each class. After all, we can divide the hedge cues into 4 groups.

word	S#	non-S#
feel	1	0
suggesting	150	0
may	516	1
or	118	6218

Table 2: Statistics of cue words. (S# short for the occurrence times in speculative sentences, non-S# for the count in non-speculative ones)

3 Methods

Conditional random fields (CRF) model was firstly introduced by Lafferty et al. (2001). CRF model can avoid the label bias problem of HMMs and other learning approaches. It was applied to solve sequence-labeling problems, and has shown good performance in NER task. We consider hedge cues detection as some kind of sequence-labeling problem, and the model will contribute to a good result.

We use CRF++ (version 0.51) to implement the CRF model. Cheng Yong, one of our team members has evaluated the several widespread used CRF tool kits, and he points out that CRF++ has better precision and recall but longer training time. Fortunately, the training time cost of BioScope corpus is acceptable. In our system, all the data training and testing processing step can be completed within 8 minutes (Intel Xeon 2.0GHz CPU, 6GB RAM). It is likely due to the small scale of the training dataset and the limited types of the annotation.

To identify sentences in the biomedical texts that contain unreliable or uncertain information (CoNLL-2010 shared task1), we start with hedge cues detection:

- If one or more than one hedge cues are detected in the sentence, then it will be annotated “uncertain”
- If not, the sentence will be tagged as “certain”.

3.1 Detecting hedge cues

The BioScope corpus annotation guidelines [1] show that most typical instances of keywords can be grouped into 4 types as Auxiliaries, Verbs of hedging or verbs with speculative content, Adjectives or adverbs, and Conjunctions. So the POS (part-of-speech) is thought to be the feature reasonably. Lemma feature of the word and chunk features are also considered to improve system performance. Chunk features may help to the recognition of biomedical entity boundaries. GENIA Tagger (Tsuruoka et al., 2005) is employed to obtain part-of-speech (POS) features, chunk features and lemma features. It works well for biomedical documents.

In the biomedical abstracts section of BioScope corpus, the hedge cues are collected into a dictionary (HCDic, short for the Hedge Cues Dictionary). As mentioned in section 2, one hedge cue appears 26.7 times on average, and we assume the set of hedge cues is limited. The HCDic consist of 143 different hedge cues extracted from the abstracts. The dictionary (HCDic) extracted from the corpus is very valuable for the system. We can focus on whether the word such as "or" listed in table 2 is a hedge cue or not. The cue words in HCDic are divided into 4 different levels with the non-S#/S# ratio.

The four types are described as "L", "H", "FL" and "FH". "L" shows low confidence of the cue word being a hedge cue, while "H" indicates high confidence about it. The prefix 'F' for "FL"/"FH" shows false negatives may happen to the cue word in HCDic. The threshold for the non-S#/S# ratio to distinguish "FL" type from "FH" is set 1.0. As the non-S#/S# ratio of "L" and "H" is always zero, we set the hedge cue whose S# is more than 5 as "H" type as shown in table 3. The four types are added into the HCDic along with the hedge cues,

In our experiment, HCDic types of word sequence are tagged as follows:

- If words are found in HCDic using maximum matching method, label them with their types in HCDic. For hedges of multi-word, label them with *BI* scheme which will be described later.
- If not, tag the words as 'O' type.

[1] http://www.inf.u-szeged.hu/rgai/bioscope

The processing assigns each token of a sentence with an HCDic type. The *BIO* types for each token are involved as features for the CRF.

The HCDic can be expanded to a larger scale. Hedge cues extracted from different corpora can be added into HCDic, and regular expression of hedge cues can be used, too. This will be helpful to the usage of cross-domain resources.

word	**S#**	**non-S#**	**type**
feel	1	0	L
suggesting	150	0	H
may	516	1	FH
or	118	6218	FL

Table 3: Types of the HCDic words. (S# and non-S# have the same meaning as in Table 2)

The features F (F stands for all the Features) including unigram, bigram, and trigram types is used for CRF as follows:

F(n)(n=-2,-1,0,+1,+2)
F(n-1)F(n)(n=-1,0,+1,+2)
F(n-2)F(n-1)F(n) (n=0,+1,+2)
Where F(0) is the current feature, F(-1) is the previous one, F(1) is the following one, etc.

We regard each word in a sentence as a token and each token is tagged with a cue-label. The *BIO* scheme is used for tagging multiword hedge cues, such as "whether or not" in our HCDic. where B-cue (tag for "whether") represents that the token is the start of a hedge cue, I-cue (tag for "or", "not") stands for the inside of a hedge cue, and O (tag for the other words in the sentence) indicates that the token does not belong to any hedge cue.

We also have the method tested on Wikipedia corpus with a preprocessing of the HCDic.

Section 2 reports that most weasel cues in Wikipedia corpus are multiword, and usually appear once. Different from our assumption in BioScope corpus, the set of weasel cues seems numerous. The HCDic of Wikipedia would be not so valuable if it tags few tokens for a new given text. To prevent these from happening, a preprocessing of the HCDic is taken.

Most of the hedge cues in Wikipedia corpus accord with the structure of "adjective + noun" e.g. "many persons". Although most cue words appear just once, the adjective usually happens to be the same, and we call them core words. Therefore, the hedge cue dictionary (HCDic) can be simplified with the core words. It helps to

reduce the scale of the hedges cues from 1984 cues down to 170. Then, we process the Wikipedia text the same way as the BioScope corpus.

3.2 Detecting scope of hedge cues

This phase (for CoNLL-2010 shared task 2) is based on Roser Morante and Walter Daelemans' scope detection system.

CRF model is applied in this part, too. The word, POS, lemma, chunk and HCDic tags are also applied to be the features as in the step of hedge cues detection. In section 3.1, we can obtain the hedge cues in a sentence. The scope relies on its cue vary much. We make the *BIO* schema of detected hedge cues to be the important features of this part. Besides, the sentences tagged as "certain" type are neglected in this step.

Here is an example of golden standard of scope label.

<sentence id="S5.149"> We <xcope id="X5.149.3"><cue ref="X5.149.3" type= "specula-tion"> propose </cue> that IL-10-producing Th1 cells <xcope id="X5.149.2"> <cue ref="X5.149.2" type= "speculation" >may</cue> be the essential regulators of acute infection-induced inflammation </xcope> and that such "self-regulating" Th1 cells <xcope id= "X5.149.1"> <cue ref= "X5.149.1" type= "speculation" >may</cue> be essential for the infection to be cleared without inducing immune-mediated pathology </xcope> </xcope>.

As shown, each scope is a block with a beginning and an end, and we refer to the beginning of scope as scope head (*<xcope…>*), and the end of the scope as scope tail (*</xcope>*).

The types of the scope are labeled as:

1. Label the token next to scope head as "xcope-H" (e.g. *propose, may*)
2. Tag the token before scope tail as "xcope-T"(e.g. *pathology* for both scopes)
3. The other words tag 'O' , including the words inside the scope and out of it. This is very different from the *BIO* scheme.

The template for each feature is the same as in section 3.1.

Following are our rules to form the scope of a hedge:

1. Most hedge cues have only one scope tag, meaning there is one-to-one relationship between hedge cue and its scope.
2. The scope labels may be nested.
3. The scope head of the cue words appears nearest before hedge cue.
4. The scope tail appears far from the cue word.
5. The most frequent head/tail positions of the scope are shown in Table 4.
   a) The scope head usually is just before the cue words.
   b) The scope tail appears in the end of the sentence frequently.

Scopes of hedge cues in BioScope corpus should be found for the shared task. The training dataset of abstract part is analyzed for its larger scale

item	Following strings with high frequency	%
1 scope head	*<cue...>*(cue words)	0.861
2 scope tail	'.'(sentence end)	0.695
	</xcope> (another scope tail)	0.144
	',' ';' ':'	0.078

Table 4: Statistics of the strings nearby the scope head and tail. Item 1 shows the word follow scope head, and item 2 shows the frequent words next to the scope tail.

We analyze the words around the scope head and the scope tail. The item 1 in Table 4 shows that 86.1% of the following words of the scope head are hedge cues. Other following words not listed are less than 1%, according to our statistics. The item 2 lists the strings with high frequency next to the scope tail as well. The first 2 words in item 2 can be combined sometimes, so the percentage of scope tail at the end of the sentence can be more than 80%. The strings ahead of scope head and tail not listed are also counted, but they do not give such valuable information as the two items listed in Table 4.

Therefore, when the CRF model gives low confidence, we just set the most probable positions of scope head and tail.

For the one-to-one relationship between hedge cues and their scopes, we make rules to insure each cue has only one scope, including the scope head and scope tail.

Rule 1: if more than one scope heads or tails are predicted, we get rid of the farther head or nearer tail.

Rule 2: if none of scope head or tail is predicted, the head is set to the word just before the cue words; the tail is set at the end of the sentence.

Rule 3: if one scope head and one tail are predicted, we consider them the result of scope detection.

4 Results

Our experiments are based on the CoNLL-2010 shared task's datasets, including BioScope and Wikipedia corpus. All the experiments for BioScope use abstracts and full papers for training data and the provided evaluation for testing.

We employ CRF model to detect the hedge cues in the BioScope. The experiments are carried out on different feature sets: words sequence with the chunk feature only, lemma feature only and POS feature only. The effect of the HCDic feature is also evaluated.

Features	**prec.**	**recall**	**F-score**
Chunk only	0.7236	0.6275	0.6721
Lemma only	0.7278	0.6103	0.6639
POS only	0.7320	0.6208	0.6718
Without HCDic	0.7150	0.6447	0.6781
ALL	0.7671	0.7393	0.7529

Table 5: Results at hedge cue-level

As described in section 1 of this paper, the feature of POS may be not so significant as the lemma, but we do not agree with this point of view for given POS feature's better performance in F-score (in Table 5). The interesting cue-level result does not go into for time limitations. The F-score of the three features, chunk, lemma and POS are approximately equal. When all of the three features are used for CRF model, the performance is not improved so significantly. The recall rate is a bit low in the experiment without HCDic features. As shown in Table 5, the feature of HCDic is effective to get a better score both in precision rate and in recall rate. As our assumption, hedges in the evaluation dataset are limited, too. Most of them along with some non-hedges can be tagged with HCDic. Then the tag could contribute to a good recall. It also helps the classifier to focus on whether the words with "L", "FL", and "FH" are hedge cues or not, which will be good for a better precision.

With detected hedge cues, we can get sentences containing uncertainty for the shared task 1. A sentence is tagged as "uncertain" type if any hedge cue is found in it.

	precision	**recall**	**F-score**
Without HCDic	0.8965	0.7898	0.8398
ALL	0.8344	0.8481	0.8412

Table 6: Evaluation result of task 1

Statistics in Table 6 show that even poor performance in cue-level test can get a satisfactory F-score of speculative sentences detection as well. It seems that hedges detection at cue-level is not proportionate to the sentence-level. Think about instance of more than one cues in a sentence such as the example of golden standard in section 3.2, the sentence will be tagged even if only one hedge cue has been identified (lower recall at cue-level). Moreover, in the speculative sentence with one hedge cue, false positives (lower precision at cue-level) can also lead to the correct result at sentence-level.

The method is also tested on Wikipedia corpus, using provided training dataset and evaluation data. The method has a bad performance in our close test. The results are listed in Table 7.

As talked in section 2, hedges in Wikipedia corpus are very different from in BioScope corpus. Besides, the string matching method for simplified HCDic is not so effective. The usefulness of HCDic is not so significant for a good recall in Wikipedia corpus.

dataset	**precision**	**recall**	**F-score**
Wikipedia	0.7075	0.2001	0.3120
BioScope	0.7671	0.7393	0.7529

Table 7: Results of weasel/hedge detection in Wikipedia and BioScope corpus.

In CoNLL-2010 shared task 2, the evaluation result shows our precision, recall and F-score are 34.8%, 41% and 37.6%. The performance of identifying the scope relies on the cue-level detection. Therefore, the false positive and false negatives of hedge cues can lead to recognition errors. The result shows that our lexical-level method for the semantic problem is limited. For the time constraints, we do not probe deeply.

5 Conclusions

This paper presents an approach for extracting the hedge cues and their scopes in BioScope corpus using two CRF models for CoNLL-2010 shared task. In the first task, the HCDic feature is proposed to improve the system performances, getting better performance (84.1% in F-score) than the baseline. The HCDic feature is also helpful to make use of cross-domain resources. The comparison of our methods based on between BioScope and Wikipedia corpus is given, which shows that ours are good at hedge cues detection in BioScope corpus but short at the in Wikipedia corpus. To detect the scope of hedge cues, we make rules to post process the text. For future work, we will look forward to constructing regulations for the HCDic to improve our system.

References

Richárd Farkas, Veronika Vincze, György Móra, János Csirik, and György Szarvas. 2010. The CoNLL-2010 Shared Task: Learning to Detect Hedges and their Scope in Natural Language Text. In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning (CoNLL-2010): Shared Task*, pages 1–12, Uppsala, Sweden, July. Association for Computational Linguistics.

Halil Kilicoglu, and Sabine Bergler. 2008. Recognizing speculative language in biomedical research articles: a linguistically motivated perspective. *BMC Bioinformatics*, 9(Suppl 11):S10.

John Lafferty, Andrew K. McCallum, and Fernando Pereira. 2001. Conditional random fields: probabilistic models for segmenting and labeling sequence data. In *ICML*, pages 282–289.

George Lakoff. 1972. Hedges: a study in meaning criteria and the logic of fuzzy concepts. *Chicago Linguistics Society Papers*, 8:183–228.

Marc Light, Xin Y. Qiu, and Padmini Srinivasan. 2004 The language of bioscience: facts,speculations, and statements in between. In *BioLINK 2004: Linking Biological Literature, Ontologies and Databases*, pages 17–24.

Ben Medlock, and Ted Briscoe. 2007. Weakly supervised learning for hedge classification in scientific literature. In *Proceedings of ACL 2007*, pages 992–999.

Roser Morante, and Walter Daelemans. 2009. Learning the scope of hedge cues in biomedical texts. In *Proceedings of the BioNLP 2009 Workshop*, pages 28-36, Boulder, Colorado, June 2009. Association for Computational Linguistics.

Roser Morante, and Walter Daelemans. 2009. A metalearning approach to processing the scope of negation. In *Proceedings of CoNLL-2009*. Boulder, Colorado.

György Szarvas. 2008. Hedge classification in biomedical texts with a weakly supervised selection of keywords. In *Proceedings of ACL 2008*, pages 281–289, Columbus, Ohio, USA. ACL.

Yoshimasa Tsuruoka, Yuka Tateishi, Jin-Dong Kim, Tomoko Ohta, John McNaught, Sophia Ananiadou, and Jun'ichi Tsujii. 2005. Developing a robust part-of-speech tagger for biomedical text. In: *Advances in Informatics, PCI 2005*, pages 382–392.

Veronika Vincze, György Szarvas, Richárd Farkas, György Móra, and János Csirik. 2008. The BioScope corpus: biomedical texts annotated for uncertainty, negation and their scopes. *BMC Bioinformatics*, 9(Suppl 11):S9.