Hedge Classification with Syntactic Dependency Features based on an Ensemble Classifier

Yi Zheng, Qifeng Dai, Qiming Luo, Enhong Chen
Department of Computer Science,
University of Science and Technology of China, Hefei, China.
{xiaoe, dqf2008}@mail.ustc.edu.cn
{luoq, cheneh}@ustc.edu.cn

Abstract

We present our CoNLL-2010 Shared Task system in the paper. The system operates in three steps: sequence labeling, syntactic dependency parsing, and classification. We have participated in the Shared Task 1. Our experimental results measured by the in-domain and cross-domain F-scores on the biological domain are 81.11% and 67.99%, and on the Wikipedia domain 55.48% and 55.41%.

1 Introduction

The goals of the Shared Task (Farkas et al., 2010) are: (1) learning to detect sentences containing uncertainty and (2) learning to resolve the in-sentence scope of hedge cues. We have participated in the in-domain and cross-domain challenges of Task 1. Specifically, the aim of Task 1 is to identify sentences in texts that contain unreliable or uncertain information, and it is formulated as a binary classification problem.

Similar to Morante et al. (2009), we use the BIO-cue labels for all tokens in a sentence to predict whether a token is the first one of a hedge cue (B-cue), inside a hedge cue (I-cue), or outside of a hedge cue (O-cue). Thus we formulate the problem at the token level, and our task is to label tokens in every sentence with BIO-cue. Finally, sentences that contain at least one B-cue or I-cue are considered as uncertain.

Our system operates in three steps: sequence labeling, syntactic dependency parsing, and classification. Sequence labeling is a preprocessing step for splitting sentence into tokens and obtaining features of tokens. Then a syntactic dependency parser is applied to obtain the dependency information of tokens. Finally, we employ an ensemble classifier based on combining CRF (conditional random field) and MaxEnt (maximum entropy) classifiers to label each token with the BIO-cue.

Our experiments are conducted on two training data sets: one is the abstracts and full articles from BioScope (biomedical domain) corpus (Vincze et al., 2008)[1], the other one is paragraphs from Wikipedia possibly containing weasel information. Both training data sets have been annotated manually for hedge/weasel cues. The annotation of weasel/hedge cues is carried out at the phrase level. Sentences containing at least one hedge/weasel cue are considered as uncertain, while sentences with no hedge/weasel cues are considered as factual. The results show that employing the ensemble classifier outperforms the single classifier system on the Wikipedia data set, and using the syntactic dependency information in the feature set outperform the system without syntactic dependency information on the biological data set (in-domain).

In related work, Szarvas (2008) extended the methodology of Medlock and Briscoe (2007), and presented a hedge detection method in biomedical texts with a weakly supervised selection of keywords. Ganter and Strube (2009) proposed an approach for automatic detection of sentences containing linguistic hedges using Wikipedia weasel tags and syntactic patterns.

The remainder of this paper is organized as follows. Section 2 presents the technical details of our system. Section 3 presents experimental results and performance analysis. Section 4 presents our discussion of the experiments. Section 5 concludes the paper and proposes future work.

2 System Description

This section describes the implementation of our system.

2.1 Information Flow of Our System

Common classification systems consist of two steps: feature set construction and classification. The feature set construction process of our sys-

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

Proceedings of the Fourteenth Conference on Computational Natural Language Learning: Shared Task, pages 151–156,
Uppsala, Sweden, 15-16 July 2010.

tem consists of sequence labeling and syntactic dependency parsing. Figure 1 shows the main information flow of our system.

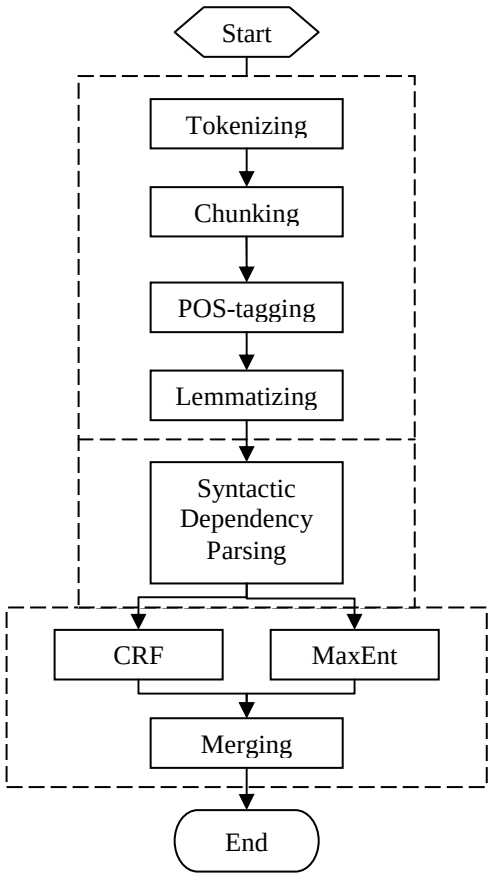

Figure 1: The main information flow of our system

2.2 Sequence labeling

The sequence labeling step consists of the following consecutive stages: (1) tokenizing, (2) chunking, (3) POS-tagging, (4) lemmatizing. Firstly, the PTBTokenizer[2] is employed to split sentence into tokens. Then, tokens are labeled with BIO-tags by the OpenNLP[3] chunker. Finally, Stanford Parser[4] is used to obtain the POS and lemma of tokens.

2.3 Syntactic Dependency Parsing

In the syntactic dependency parsing stage, we use the Stanford Parser again to obtain dependency information of tokens. Based on the Stanford typed dependencies manual (Marneffe and Manning 2008), we have decided to obtain the tree dependencies structure. During the process of parsing, we found that the parser may fail due to either empty sentences or very long sentences. To deal with very long sentences, we decided to allocate more memory. To deal with empty sentences, we decided to simply label them as certain ones because there are only a few empty sentences in the training and test data sets and we could ignore their influence.

2.4 Features

After sequence labeling and syntactic dependency parsing, we obtain candidate features. In our system, all the features belong to the following five categories: (1) token features, (2) dependency features, (3) neighbor features, (4) data features, (5) bigram and trigram features.

Token features of the current token are listed below:

- token: the current token.
- index: index of the current token in the sentence
- pos: POS of the current token.
- lemma: lemma of the current token.
- chunk: BIO-chunk tags of the current token.

Dependency features of the current token are listed below:

- parent_index: the index of the parent token of the current token.
- parent_token: the parent token of the current token.
- parent_lemma: the lemma of the parent token of the current token.
- parent_pos: the POS of the parent token of the current token.
- parent_relation: the dependency relation of the current token and its parent token.

Neighbor features of the current token include token, lemma, pos, chunk tag of three tokens to the right and three to the left.

Data features of current token are listed below:

- type: indicating documentPart[5] type of the sentence which contains the current token, such as Text, SectionTitle and so on.
- domain: distinguishing the Wikipedia and biological domain.
- abstract_article: indicating document type of the sentence which contains the current token, abstract or article.

[2] a tokenizer from Stanford Parser.
[3] http://www.opennlp.org/
[4] http://nlp.stanford.edu/software/lex-parser.shtml

[5] documentPart, SectionTitle, Text and so on are tags in the training and test data sets.

We empirically selected some bigram features and trigram features as listed below:

- left_token_2+left_token_1
- left_token_1+token
- token+right_token_1
- right_token_1+right_token_2
- left_token_2+left_token_1+token
- left_token_1+token+right_token_1
- token+right_token_1+right_token_2

These are the complete set of features for our system. If the value of a feature is empty, we set it to a default value. In the ensemble classifier, we have selected different features for each individual classifier. Details of this are described in the next subsection.

2.5 Classification

In our system, we have combined CRF++[6] and OpenNLP MaxEnt[7] classifiers into an ensemble classifier. The set of features for each classifier are shown in the column named "system" of Table 6. And the two classifiers are used in training and prediction separately, based on their individual set of features. Then we merge the results in this way: for each token, if the two predictions for it are both O-cue, then we label the token with an O-cue; otherwise, we label the token with a B-cue (one of the predictions is B-cue) or an I-cue (no B-cue in the predictions). The motivation of the ensemble classifier approach is based on the observation of our internal experiments using 10-fold cross validation, which we describe in Section 3. In addition, the parameters of OpenNLP MaxEnt classifier are all set to default values (number of iterations is 100, cutoff is 0 and without smoothing). For CRF++, we only set the option "-f" as 3 and the option "-c" as 1.5, and the others are set to default values.

3 Experimental Results

We have participated in four subtasks, biological in-domain challenge (Bio-in-domain), biological cross-domain challenge (Bio-cross-domain), Wikipedia in-domain challenge (Wiki-in-domain) and Wikipedia cross-domain challenge (Wiki-cross-domain). In all the experiments, TP, FP, FN and F-Score for the uncertainty class are used as the performance measures. We have tested our system with the test data set and obtained official results as shown in Table 1. In addition, we have performed several internal experiments on the training data set and several experiments on the test data set, which we describe in the next two subsections. The feature sets used for each subtask in our system are shown in Table 6, where each column denotes a feature set named after the title of the column ("System", "dep", …). Actually, for different subtasks, we make use of the same feature set named "system".

SubTask	TP	FP	FN	F-Score
Bio-in-domain	717	261	73	81.11
Bio-cross-domain	566	309	224	67.99
Wiki-in-domain	974	303	1260	55.48
Wiki-cross-domain	991	352	1243	55.41

Table 1: Official results of our system.

3.1 Internal Experiments

Initially we only used a single classifier instead of an ensemble classifier. We performed 10-fold cross validation experiments on the training data set at the sentence level with different feature sets. The results of these experiments are shown in Table 2 and Table 3.

In internal experiments, we mainly focus on the results of different models and different feature sets. In Table 2 and Table 3, CRF and ME (MaxEnt) indicate the two classifiers; ENSMB stands for the ensemble classifier obtained by combining CRF and MaxEnt classifiers; the three words "dep", "neighbor" and "together" indicate the feature sets for different experiments shown in Table 6, and "together" is the union set of "dep" and "neighbor".

The results of ME and CRF experiments (third column of Table 2 and Table 3) show that the individual classifier wrongly predicts many uncertain sentences as certain ones. The number of such errors is much greater than the number of errors of predicting certain ones as uncertain. In other words, FN is greater than FP in our experiments and the recall ratio is very low, especially for the Wikipedia data set.

[6] http://crfpp.sourceforge.net/
[7] http://maxent.sourceforge.net/

Experiment	Biological in-domain				Biological cross-domain			
	TP	FP	FN	F-Score	TP	FP	FN	F-Score
ME-dep	244	28	34	88.73	220	24	58	84.29
CRF-dep	244	20	34	90.04	230	19	48	87.29
ENSMB-dep	248	32	30	88.89	235	28	43	86.88
ME-neighbor	229	14	49	87.91	211	12	67	84.23
CRF-neighbor	244	16	34	90.71	228	21	50	86.53
ENSMB-neighbor	247	22	31	90.31	241	26	37	88.44
ME-together	234	11	44	89.48	205	12	73	82.83
CRF-together	247	13	31	91.82	234	21	44	87.80
ENSMB-together	253	17	25	**92.36**	242	26	36	**88.64**

Table 2: Results of internal experiments on the biological training data set.

Experiment	Wikipedia in-domain				Wikipedia cross-domain			
	TP	FP	FN	F-Score	TP	FP	FN	F-Score
ME-dep	131	91	117	55.74	145	108	103	57.88
CRF-dep	108	51	140	53.07	115	60	133	54.37
ENSMB-dep	148	103	100	59.32	153	119	95	58.85
ME-neighbor	106	52	142	52.22	130	77	118	57.14
CRF-neighbor	123	44	125	59.28	123	72	125	55.53
ENSMB-neighbor	145	71	103	**62.50**	154	116	94	**59.46**
ME-together	100	57	148	49.38	117	69	131	53.92
CRF-together	125	54	123	58.55	127	67	121	57.47
ENSMB-together	141	83	107	59.75	146	104	102	58.63

Table 3: Results of internal experiments on the Wikipedia training data set.

Experiment	Biological in-domain				Biological cross-domain			
	TP	FP	FN	F-Score	TP	FP	FN	F-Score
System-ME	650	159	140	81.30	518	265	272	65.86
System-CRF	700	197	90	**82.99**	464	97	326	**68.69**
System-ENSMB	717	261	73	81.11	566	309	224	67.99

Table 4: Results of additional experiment of biological test data set.

Experiment	Wikipedia in-domain				Wikipedia cross-domain			
	TP	FP	FN	F-Score	TP	FP	FN	F-Score
System-ME	794	235	1440	48.67	798	284	1436	48.13
System-CRF	721	112	1513	47.02	747	153	1487	47.67
System-ENSMB	974	303	1260	**55.48**	991	352	1243	**55.41**

Table 5: Results of additional experiment of Wikipedia test data set.

Based on this analysis, we propose an ensemble classifier approach to decrease FN in order to improve the recall ratio. The results of the ensemble classifier show that: along with the decreasing of FN, FP and TP are both increasing. Although the recall ratio increases, the precision ratio decreases at the same time. Therefore, the ensemble classifier approach is a trade-off between precision and recall. For data sets with low recall ratio, such as Wikipedia, the ensemble classifier outperforms each single classifier in terms of F-score, just as the ME, CRF and ENSMB experiments show in Table 2 and Table 3.

In addition, we have performed simple feature selection in the internal experiments. The comparison of "dep", "neighbor" and "together" experiments shown in Table 2 demonstrates that the dependency and neighbor features are both beneficial only for the biological in-domain experiment. This may be because that sentences of the biological data are more regular than those of the Wikipedia data.

3.2 Additional experiments on test data set

We have also performed experiments on the test data set, and the results are shown in Table 4 and Table 5. With the same set of features of our sys-

tem as shown in Table 6, we have performed three experiments: System-ME (ME denotes MaxEnt classifier), System-CRF (CRF denotes CRF classifier) and System-ENSMB (ENSMB denotes ensemble classifier), where "System" denotes the feature set in Table 6. The meanings of these words are similar to internal experiments.

As Table 4 and Table 5 show, for the Wikipedia test data set, the ensemble classifier outperforms each single classifier in terms of F-score by improving the recall ratio with a larger extent than the extent of the decreasing of the precision ratio. For the biological test data set, the ensemble classifier outperforms System-ME but underperforms System-CRF. This may be due to the relatively high values of the precision and recall ratios already obtained by each single classifier.

4 Discussion

The features in our experiments are selected empirically, and the performance of our system could be improved with more elaborate feature selection. From the experimental results, we observe that there are still many uncertain sentences predicted as certain ones. This indicates that the ability of learning uncertain information with the current classifiers and feature sets needs to be improved. We had the plan of exploring the ensemble classifier by combining CRF, MaxEnt and SVM (Support Vector Machine), but it was given up due to limited time. In addition, we were not able to complete experiments with MaxEnt classifier based on bigram and trigram features due to limited time. Actually only two labels I and O are needed for Task 1. We have not done the experiments with only I and O labels, and we plan to do it in the future.

According to our observation, the low F-score on the Wikipedia data set is due to many uncertain phrases. By contrast, for the biological data set, the uncertain information consists of mostly single words rather than phrases. It is difficult for a classifier to learn uncertain information consisting of 3 words or more. As we have observed, these uncertain phrases follow several patterns. A hybrid approach based on rule-based and statistical approaches to recognize them seems to be a promising.

5 Conclusion and Future Work

Our CoNLL-2010 Shared Task system operates in three steps: sequence labeling, syntactic dependency parsing, and classification. The results show that employing the ensemble classifier outperforms each single classifier for the Wikipedia data set, and using the syntactic dependency information in the feature set outperform the system without syntactic dependency information for the biological data set (in-domain). Our final system achieves promising results. Due to limited time, we have only performed simple feature selection empirically. In the future, we plan to explore more elaborate feature selection and explore ensemble classifier by combining more classifiers.

Acknowledgments

The work was supported by the National Natural Science Foundation of China (No.60775037), the National High Technology Research and Development Program of China (863 Program, No.2009 AA01Z123), and Specialized Research Fund for the Doctoral Program of Higher Education (No.20093402110017).

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 CoNLL-2010: Shared Task*, pages 1–12.

Viola Ganter, and Michael Strube. 2009. Finding hedges by chasing weasels: Hedge detection using Wikipedia tags and shallow linguistic features. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, pages 173–176.

Marie-Catherine de Marneffe, and Christopher D. Manning. 2008. *Stanford typed dependencies manual*, September 2008.

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

Roser Morante, and Walter Daelemans. 2009. Learning the scope of hedge cues in biomedical texts. In *Proc. of the BioNLP 2009 Workshop*, pages 28–36.

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

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.

Feature	System	Dep	Neighbor	Together
token	mc	mc	mc	mc
index	m	m	m	m
pos	mc	mc	mc	mc
lemma	mc	mc	mc	mc
chunk		mc	mc	mc
parent_index	mc	mc		mc
parent_token		mc		mc
parent_lemma	mc	mc		mc
parent_relation	mc	mc		mc
parent_pos	mc	mc		mc
left_token_1	c		c	c
left_lemma_1	mc		mc	mc
left_pos_1	mc		mc	mc
left_chunk_1			mc	mc
left_token_2	c		c	c
left_lemma_2	c		mc	mc
left_pos_2	mc		mc	mc
left_chunk_2			mc	mc
left_token_3				
left_lemma_3	mc		m	m
left_pos_3	mc		m	m
left_chunk_3			m	m
right_token_1	c		c	c
right_lemma_1	mc		mc	mc
right _pos_1	mc		mc	mc
right _chunk_1			mc	mc
right_token_2	c		c	c
right _lemma_2	mc		mc	mc
right _pos_2	c		mc	mc
right _chunk_2			mc	mc
right_token_3				
right _lemma_3	c		m	m
right _pos_3	mc		m	m
right _chunk_3			m	m
type	m	mc	mc	mc
domain	m	mc	mc	mc
abstract_article	m	mc	mc	mc
left_token_2+left_token_1	c		c	c
left_token_1+token	c		c	c
token+right_token_1	c		c	c
right_token_1+right_token_2	c		c	c
left_token_2+left_token_1+token	c		c	c
left_token_1+token+right_token_1	c		c	c
token+right_token_1+right_token_2	c		c	c

Table 6: Features selected for different experiments. The symbol m indicates MaxEnt classifier and c indicates CRF classifier.