Connective Prediction using Machine Learning for Implicit Discourse Relation Classification

Yu Xu, Man Lan*, Yue Lu
East China Normal University
Email: 51101201049@ecnu.cn
mlan,ylu@cs.ecnu.cn

Zheng Yu Niu
Baidu.com Inc.,Beijing, PRC.
Email: niuzhengyu@baidu.com

Chew Lim Tan
National University of Singapore
Email: tancl@comp.nus.edu.sg

***Abstract*—Implicit discourse relation classification is a challenge task due to missing discourse connective. Some work directly adopted machine learning algorithms and linguistically informed features to address this task. However, one interesting solution is to automatically predict implicit discourse connective. In this paper, we present a novel two-step machine learning-based approach to implicit discourse relation classification. We first use machine learning method to automatically predict the discourse connective that can best express the implicit discourse relation. Then the predicted implicit discourse connective is used to classify the implicit discourse relation. Experiments on Penn Discourse Treebank 2.0 (PDTB) and Biomedical Discourse Relation Bank (BioDRB) show that our method performs better than the baseline system and previous work.**

I. Introduction

Discourse relations, such as *Contrast* and *Cause*, are to describe how two sentences or clauses are semantically connected with each other. It can be helpful in many natural language processing(NLP) applications, such as QA [1], text summarization [2] and text generation [3]. For example, in QA system, detecting *Cause* relation in text can help to answer *Why* question. Moreover, recognizing *Restatement* relation is useful for document summarization.

Discourse relation can sometimes be marked lexically by words and expressions (i.e., cue words) in the texts, such as *but* or *since*. For example, *because* indicates a *Contingency* relation in E1. In this case, the discourse relation and connectives are named as *explicit discourse relation* and *explicit discourse connectives* (known as *discourse markers*).

(**E1**) Longer maturities are thought to indicate declining interest rates **because** they permit portfolio managers to retain relatively higher rates for a longer period.

(Contingency-wsj_0004)

However, connectives are sometimes absent between sentences. In E2, *for instance* does not appear in real text, but it can be manually inserted by annotator into the text to express the *Expansion* relation between sentences without any redundancy[1]. In this case, the discourse relation (here *Expansion* relation) and the discourse connective (here *for instance*) are called as *implicit discourse relation* and *implicit discourse connective*, respectively.

(**E2**) Typically, money-fund yields beat comparable short-term investments because portfolio managers can vary maturities and go after the highest rates. [**for instance**] The top money funds are currently yielding well over 9%.

(Expansion-wsj_0004)

Previous work proved that using just the connectives in text, the accuracy of explicit discourse relation classification can reach 93% [5]. However, without the information provided by explicit connectives, the task of implicit discourse relation classification has become quite a challenge [6]. Furthermore, in real world texts, discourse connectives are often missing between sentences. Actually, almost half the sentences in the British National Corpus have no discourse connective entirely [7]. Among a total of 40,600 annotated relations in PDTB, 16,053 (40%) are annotated as implicit discourse relation [4]. Therefore, improving the performance of implicit relation classification is the key to the overall performance of discourse relation analysis. In this paper, we focus on the task of implicit discourse relation classification.

Similar to explicit discourse connective, implicit discourse connective is also a useful feature for implicit discourse relation classification. [8] proved that the F-score of implicit discourse relation classification achieves 91.8% on PDTB by simply mapping the hand-annotated implicit connectives to their most frequent sense. Furthermore, they used a language model built on unlabeled corpus to "insert" an implicit connective between two sentences, and the inserted connective is regraded as an important feature for implicit relation classification [8] [9]. It indicates that the prediction of implicit discourse connective can bring help to the classification of implicit discourse relation.

The basic idea of our new approach is to perform a two-step classification using machine learning for implicit discourse relation classification. Specifically, in the first step, classifiers are built to predict implicit discourse connectives between two text spans using linguistically informed features extracted from context sentences. In the second step, the connectives predicted in the first step are used as features to construct implicit discourse relation classifier. By doing so, the implicit

[1]According to the PDTB Annotation Manual [4], if the insertion of connective leads to "redundancy", the relation is annotated as Alternative lexicalizations(AltLex), not implicit.

connectives predicted in the first step act as a bridge to be used by the second step for implicit relation classification.

The organization of the paper is as follows. Section II summarizes related work and introduces the corpora we used in the work. Section III presents our new proposed approach. Section IV presents our experiments. Section V reports the results of our experiment. Section VI discusses the experimental results. Finally, Section VII concludes this work.

II. Related work

A. Previous work

Labeled and/or unlabeled data have been used to recognize implicit discourse relation in previous work. According to the data used in the method, existing work on implicit discourse classification falls into three categories. Note that although some studies have focused on both implicit and explicit discourse relations, we only focus on the part of their work relevant to implicit discourse relation classification.

The earlier work has performed discourse relation classification on unlabeled data. For example, [10] proposed an unsupervised method to recognize discourse relations, i.e., *Contrast*, *Explanation-evidence*, *Condition* and *Elaboration*, between two arbitrary segments of text. They used unambiguous patterns to generate synthetic implicit discourse relation data set from unlabeled data automatically. And they found that the word pairs from the two sentences can reflect their relation. Based on the work of [10], some studies attempted to extend the work to improve the performance of relation classification. For example, [11] proved that phrasal patterns are useful in relation classification and therefore they combined word pairs with phrasal pattern to predict the discourse relations on a Japanese corpus. In addition, [12] presented a refined work on training and classification process using parameter optimization, topic segmentation and syntactic parsing. However, recently [7] pointed out that training model built on a synthetic data set like what [10] did may not be a good strategy because the model learned from the synthetic data may not perform as well on natural data set.

The second research work is based on labeled data. For example, [13] parsed the discourse structures within the same sentence using RST Bank [14], which is annotated based on Rhetorical Structure Theory [15]. [16] presented relation disambiguation on GraphBank [17], which is annotated with discourse relation.

Recently, after the release of PDTB [18], this hand-annotated corpus has become a new benchmark data set and has been widely used for discourse relation classification in open-domain [6] [19] [20]. These studies have constructed relation classifiers using machine learning algorithms, e.g., SVM and Naïve Bayes, with different linguistically informed features, such as verb, context and modality, to automatically predict the implicit discourse relation between two spans of text.

The third research work performs semi-supervised framework on both labeled and unlabeled data. [21] proposed a semi-supervised method based on the analysis of co-occurring features in unlabeled data, which was used to extend feature vectors. In addition, they adopted structure learning method with the help of unlabeled data [22].

Unlike all above work which performed implicit discourse relation classification directly without considering implicit discourse connectives, [8] [9] proposed a method based on predicting implicit connectives, which are quite relevant to our work. They predicted discourse connectives with the use of language model trained on large amount of unlabeled data. Then they combined the predicted connectives with other linguistically informed features to classify discourse relation. In contrast to the language model based method in [8] [9], we present a two-step classification method using linguistically informed feature and machine learning. The major benefit of our approach is that it can take advantage of two strengths from these above research work. First, since discourse connectives are good indicators for discourse relation, the predicted implicit discourse connectives are supposed to help in implicit discourse relation classification. Second, the various linguistically informed features from context sentences are supposed to help in implicit discourse connective prediction rather than a language model built on unlabeled corpus.

B. Benchmark Corpora

We evaluated our method in both open domain and biomedical domain. To our knowledge, little published work on implicit discourse relation classification performed experiments in biomedical domain. Specifically, we used two benchmark corpora, i.e., PDTB and BioDRB, in our experiments.

1) Penn Discourse Treebank 2.0: PDTB [18] is an open-domain corpus of discourse relation, which contains 2312 Wall Street Journal articles. It is the largest corpora of discourse relation so far. The tag set of senses is organized hierarchically with three levels, i.e., *class*, *type* and *subtype*. The *class* level contains four major semantic classes: Comparison, Contingency, Expansion and Temporal. In this paper, we work on the *class* level.

For implicit discourse relation annotation, as shown in E3, an annotator first identified two text spans and labeled them as *Arg1* and *Arg2*. Then an implicit discourse connective (here *for instance*) that best expresses the relation (here *Expansion.Instantiation*) between the text spans was inserted manually by the annotator. At last, a hierarchical sense label was assigned as *Expansion* class and *Instantiation* type.

(**E3**) **Relation**: Implicit
Arg1:Typically, money-fund yields beat comparable short-term investments.
Arg2:The top money funds are currently yielding well over 9%.
Connective:for instance
Sense:Expansion.Instantiation
(Expansion-wsj_0004)

2) Biomedical discourse relation bank: BioDRB [23] is a biomedical domain corpus of discourse relation, which adapts the annotation framework of PDTB. It has 24 articles from GENIA corpus [24]. Different from PDTB, BioDRB only

has a two-level tagset of sense. However, we can follow the instruction offered by [23] to reconstruct the PDTB *class* level sense from the BioDRB sense types. In this way, we can compare the performance of our method in the same semantic level on the two benchmark corpora.

III. Proposed Method

A. Motivation

Discourse connectives are important indicators for both explicit and implicit discourse relation. As previous work shown, using just discourse connectives as feature, the accuracy of explicit relation classification can reach 93% [5] and the F-score of implicit discourse relation can reach 91.8% [8].

Generally, for explicit discourse relation classification task, it makes full use of discourse connectives in the following way. First is to identify the explicit discourse connectives and tell it from other words in text. Second is to use the connectives to classify the discourse relation. However, for implicit discourse relation classification, because of missing connectives in text, most previous work built implicit discourse relation classifier without the step of identifying implicit discourse connectives [25] [19] [6] [10].

On the other hand, although implicit discourse connectives do not appear in text, they can be figured out according to the context. For example, as we mentioned above, PDTB annotators "insert" implicit discourse connective manually and the implicit discourse connective can help them decide the implicit discourse relation. Following this idea, automatically predicting implicit discourse connective is an interesting and promising solution to solve the problem of implicit discourse relation classification. This idea was first presented by [8], which predicted connectives with the use of language model trained by unlabeled data.

As we all know, people follow the grammatical usage of connectives to express the semantic relation between arguments. For example, *neither* is paired with *nor* in a sentence rather than with *or*. It means that the linguistically informed features, such as lexical features and syntactic features, can act as good indicators for implicit discourse connective prediction. However, the connective prediction method proposed by [8] only adopted language model, which depends on the word sequence only and cannot incorporate with richer linguistically informed features. In consideration of the above analysis and previous work, one question consequently arises there, can we use linguistically informed feature to predict implicit discourse connectives?

The basic idea of our new approach is to address the implicit discourse relation classification by predicting implicit discourse connectives between sentences using machine learning methods worked with various linguistically informed features. To the best of our knowledge, this work is the first to predict implicit discourse connectives using machine learning. Unlike previous work in [6] [19] [20] that constructed automatic discourse relation classifiers using machine learning algorithms with different linguistically informed features, this work presents a two-step classification approach instead of performing one-step discourse relation classification. Moreover, unlike previous work in [8] [9] that predicted discourse connectives using a language model, this work makes use of linguistically informed features extracted from context sentences, which are supposed to provide more information than a language model built on unlabeled corpus. The difference between the previous method and our method is shown in Figure 1.

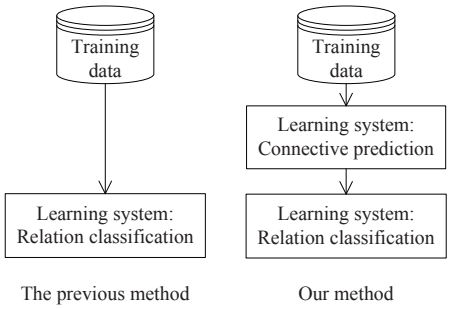

Fig. 1. The left figure shows the general procedure of previous studies on implicit discourse relation classification. The right figure shows the two-step procedure of our proposed approach.

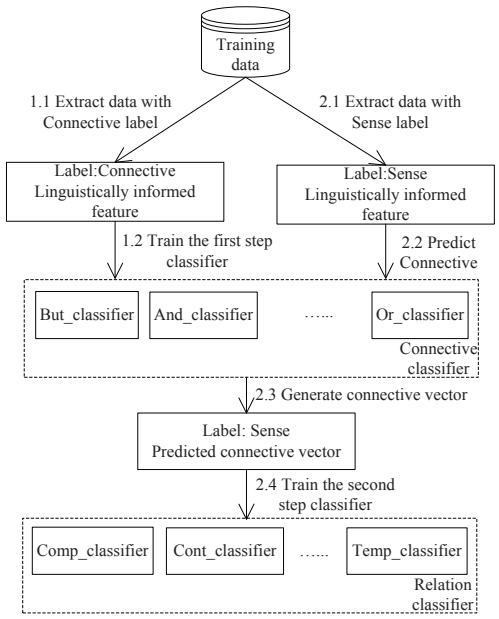

Fig. 2. The steps of training procedure

B. Our Proposed Method

To address the classification of implicit discourse relation, we propose a two-step machine learning method. The classifier in the first step, called *connective classifier*, is to predict implicit discourse connectives that can best express the relation between arguments. In the second step, the classifier, called *relation classifier*, is to classify implicit relation based on the connectives predicted by the first layer of classifier. In this

method, implicit discourse connective is added as a "bridge" to support implicit discourse relation classification.

1) Training procedure: Figure 2 depicts the general training procedure of our method. From the description in Section II-B, we know that each training example contains both a *connective* label and a *sense* (i.e., *relation*) label. Since the tasks of the two steps have different purposes, these two labels are used to build different classifiers in the two steps. That is, *connective* label is used to build connective classifier in the first step and *sense* label is used to construct relation classifier in the second step.

Step 1: build connective classifier The purpose of the first step is to predict implicit discourse connective. To do so, we construct connective classifier using machine learning method worked with linguistically informed features, such as syntactic feature and lexical feature (we will describe these features in the following subsection). The first step consists of two sub steps indicated as step 1.1 and step 1.2 in Figure 2.

Step 1.1: In this step, each training instance is generated according to its *connective* label rather than its *relation* (i.e., *sense*) label. That is, in the first step the connective attached to each training example is granted as its category label. After that, for each instance, we extract linguistically informed features from two spans of text. These extracted features and connective label are combined together as feature vectors for each training instance.

Step 1.2: In this step, we use multiple independent binary classifiers to perform implicit discourse prediction. That is, for each connective, we construct a binary classifier for it. In the test stage, each connective classifier is used to predict whether this connective can best express the implicit discourse relation in the test sample or not. For example, the *But_classifier* is used to predict whether *But* can best express the implicit discourse relation in two arguments or not.

Step 2: build relation classifier The purpose of Step 2 is to perform implicit discourse relation classification based on the implicit connectives predicted from Step 1. The procedure of the second step consists of four steps as shown in Figure 2 (from step 2.1 to step 2.4).

Step 2.1: In this step, each training instance is generated according to its *sense* label from corpus. These extracted features and connective label are combined together as feature vectors for each training instance.

Step 2.2: We use the connective classifier to predict implicit discourse connectives. From the description in step 1.2, we know that connective classifier contains independent binary classifiers(suppose the number of binary classifiers is N). Therefore, we can get N binary outputs from these connective classifiers.

Step 2.3: We combine the N binary outputs obtained in step 2.2 into a vector, called *predicted connective vector*(PCV), which reflects the result of connective prediction.

Step 2.4: The relation classifier is trained to classify implicit discourse relation using the PCV. Similar to the step 1.2, we use independent binary classifiers to perform the multi-class classification problem. Every binary classifier is trained to distinguish one relation from the others. According to the Yes or No prediction, we can output the final relation for each test instance. For example, in order to classify *Comparison* relation, we construct a binary classifier, named *comp_classifier* to distinguish *Comparison* relation and other relations in text.

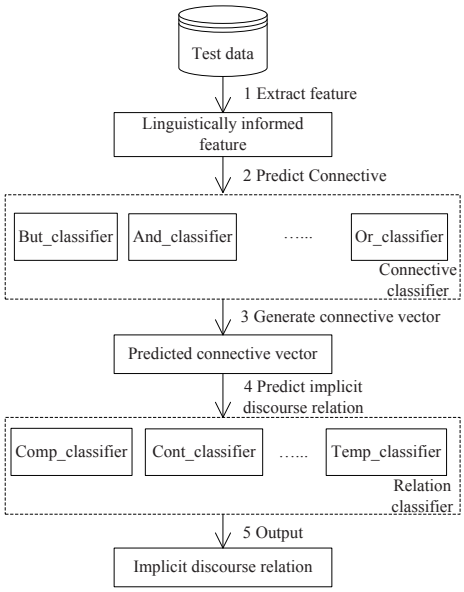

Fig. 3. The steps of test procedure

2) Test procedure: In test stage, we use our two-step method to classify the implicit discourse relation. First, we using the connective classifier to predict the implicit discourse connective. The outputs of connective classifier are combined to generate the predicted connective vector. Then, the predicted connective vector is used in the relation classifier to classify implicit discourse relation. The test procedure is described in Figure 3.

IV. Experiments

A. Dataset

PDTB and BioDRB are used in our experiment to evaluate the performance of our method in open-domain and biomedical domain.

From the description in section III, we find that two datesets are generated. In the first step, we extract data with *connective* label(connective prediction dataset) to train connective classifier. In the second step, data with *sense* label(relation classification dataset) is used to train relation classifier.

1) PDTB: Following the previous work [6] [8] [9], we used the sections 2-20 for training and sections 21-22 for testing. Sections 00-01 worked as the development set for parameter optimization.

Connective prediction dataset: In the training dataset, there were 93 hand-annotated implicit connectives. Therefore, we constructed 93 binary classifiers for connective prediction. We grouped the training data into 93 sub datasets according

to their *connective* labels. In every sub dataset, we added the equal number of negative samples randomly.

For example, Because(1477 each), While(516 each), Thus(140 each).

Relation classification dataset: We work on the *class* level in PDTB, which contains four most general semantic classes: *Comparison* (Comp.), *Contingency* (Cont.), *Expansion* (Exp.) and *Temporal* (Temp.). Therefore, four binary classifiers were constructed for them. The number of training and test instances is listed in Table I.

TABLE I
RELATION CLASSIFICATION DATASET IN PDTB(POSITIVE SAMPLE/NEGATIVE SAMPLE)

	Comp.	Cont.	Exp.	Temp.	NoRel.
train	1939/1939	3333/3333	6316/6316	736/736	-
test	152/907	279/780	567/492	83/976	13/1046

2) BioDRB: In BioDRB, we used the two articles (GENIA_1421503 and GENIA_1513057) for testing and one article (GENIA_111020) for parameter optimization. The rest of the labeled articles were used for training.

Connective prediction dataset: The BioDRB training dataset had 49 hand-annotated implicit connectives. Therefore, we grouped training data into 49 sub datasets. In every sub dataset, the numbers of positive sample and negative samples were equal.

Relation classification dataset: Following the instruction of [23], we group the training data into *class* level. The numbers of training and test instances are listed in Table II.

TABLE II
RELATION CLASSIFICATION DATASET IN BIODRB(POSITIVE SAMPLE/NEGATIVE SAMPLE)

	Comp.	Cont.	Exp.	Temp.	NoRel.
train	134/134	90/90	2084/2176	320/320	-
test	7/242	5/244	204/45	30/219	3/246

B. Linguistically informed features

In this paper, we extract 5 features, including semantic features, lexical features and the syntactic features, due to their superior performance in previous work [6] [19] [8].

Verbs: We count the number of verb pairs in arguments which come from the same class in Levin verb class [26]. Moreover, the average length of verb phrases in each argument and the part of speech tags of the main verb are included.

Polarity: The feature includes the number of positive, negated positive, negative and neutral words in the arguments.

Modality: This feature indicates whether there is a modal verb in the arguments and what is the model verb.

Production Rule: Similar to the work of [19], we extract all the production rules in the arguments. For every production rule, three binary features are used to present whether this rule appears in Arg1, Arg2 and both of them.

Word-pair: Following the work of [10], we collect the word pairs from the two arguments and remove those which appear less than 10 times in the training data set.

Due to domain specificity in PDTB and BioDRB, we used different feature sets. According to their performance in development set, the feature set for PDTB contains *Polarity*, *Modality*, *Production Rule* and *Word-pair*. In BioDRB, *Verbs*, *Polarity* and *Modality* are included in feature set.

C. Performance Evaluation and Machine Learning Algorithms

We report *F-score* and *accuracy*(acc) to evaluate the performance of our method. *F-score* is the harmonic mean of recall and precision. *Accuracy* is the number of test samples classified correctly divided by the total number of test data.

Many machine learning algorithms have been adopted in natural language processing (NLP), such as SVM, *k*NN, Naïve Bayes, decision tree, etc. In this paper, we used Naïve Bayes and Maximum Entropy(MaxEnt) classifiers implemented in MALLET toolkit [27] for our experiment.

D. Baseline System

To make the performance comparison reasonable and reliable, we construct a baseline system that used machine learning algorithm and linguistically informed features to classify the implicit discourse relation without connective prediction.

E. Experiments of two-step Classifier

1) Connective Classifier: We first evaluate the performance of classifier for connective prediction, ie. connective classifier. Then, we chose the classifier which achieving the best performance as the *best connective classifier*.

2) Relation Classifier: For the experiment on relation classifier, we conduct two experiments.

In the first experiment (EXPERIMENT I), the relation classifier uses the connective classifier that was trained by the same machine learning algorithm, ie. the two step classifiers were trained by the same algorithm.

In the second experiment (EXPERIMENT II), we use the best connective classifier to predict implicit discourse connective.

For example, we use MaxEnt to construct the relation classifier. In EXPERIMENT I, connective classifier trained by MaxEnt is used to predict implicit discourse connective. However, the best connective classifier is used in EXPERIMENT II

TABLE III
F-SCORE(ACCURACY) OF CONNECTIVE CLASSIFIER USING DIFFERENT MACHINE LEARNING ALGORITHMS

	NaïveBayes	MaxEnt
PDTB	**53.22(57.62)**	50.00(53.77)
BioDRB	36.31(44.26)	**38.79(46.36)**

V. RESULTS

A. Results on PDTB

1) Results of Connective Prediction: Table III summarizes the results of connective classifier using Naïve Bayes and MaxEnt. From this table, we observe that the average F-score and accuracy of connective classifier (93 binary classifiers) trained by Naïve Bayes are 53.22% and 57.62%, respectively, which are better than the performance of the connective classifier trained by MaxEnt. Thus, the connective classifier trained by Naïve Bayes is regarded as the best connective classifier.

2) Results of relation classification: Figure 4 compares the F-score of Naïve Bayes and MaxEnt. It contains the results of baseline system and two experiments in relation classification, ie. EXPERIMENT I and EXPERIMENT II. Note that the best connective is trained by Naïve Bayes, so the result of EXPERIMENT II of it is not shown in Figure 4

From Figure 4, we can find that our method improved the F-score in *Comparison* and *Expansion* relation. For example, our method achieved 4.22% and 5.23% improvement in the experiment of MaxEnt. For *Contingency* relation, in the experiment of Naïve Bayes, we achieve comparable F-score to baseline system. But, in the experiment of MaxEnt, our method provides a 5.21% improvement. However, the result shows that our method can not perform well in *Temporal* relation. The possible reason is discussed in Section VI in detail. Further, using the best connective classifier,

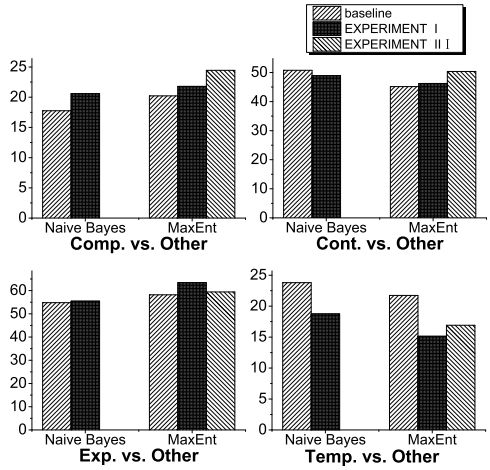

Fig. 4. The performance(F-socre) of different relations and different machine learning algorithms in PDTB

EXPERIMENT II performed better than EXPERIMENT I in *Comparison*, *Contingency* and *Temporal* relation. It means that the performance of implicit relation classification depend on the performance of connective prediction. Improving the performance of connective prediction is a promising solution to implicit discourse relation classification.

Table IV reports the best result of our method and previous work [9]. The first line is the best result of using Naïve Bayes and the second line list the result of MaxEnt. The last line in Table IV shows the result of [9]. They introduced a language model trained on a large amount of unlabeled data to predict implicit discourse connective. From this table, we can find that our method performs better than the previous work.

B. Results on BioDRB

1) Results of Connective Prediction: Table III shows the results of connective classifiers in BioDRB.

In BioDRB, MaxEnt achieves the best result, the F-score and accuracy of MaxEnt classifier are 38.79% and 46.36%, respectively. Thus, the best connective classifier is the classifier trained by MaxEnt.

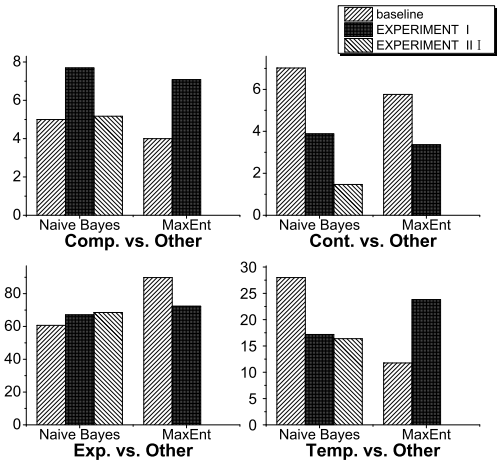

Fig. 5. The performance (F-score) of different relations and different machine learning algorithms in BioDRB

2) Results of relation classification: Figure 5 shows the f-score of Naïve Bayes and MaxEnt in BioDRB. Since the best connective classifier is trained by MaxEnt, the result of EXPERIMENT II using MaxEnt is not shown in Figure 5.

From Table II, we know that the data distribution in BioDRB is skewed. The size of training data in *Expansion* relation is much more than the other relations. Thus, the performance in Expansion relation is much better than the other relations.

Table V summarizes the best results of implicit relation classification. From this table, we find that the performance of MaxEnt is better than Naïve Bayes.

VI. ANALYSIS AND DISCUSSION

A. The size of training data set

In this paper, we perform an error analysis on our method. We find that the size of training data set has great impact on the performance of our method.

Connective classifier: PDTB has 102 implicit discourse connectives in total. Over 15% of them appear only once and 36% of them occur less than 10 times. For example, *what's more* appears only once and *yet* appears 6 times in PDTB as implicit discourse relation [4]. In BioDRB, 54.72% implicit discourse connectives appear less than 10 times.

TABLE IV
THE PERFORMANCE OF DIFFERENT RELATIONS AND DIFFERENT MACHINE LEARNING ALGORITHMS IN PDTB

Method	Comp. vs. Other F-score(acc)	Cont. vs. Other F-score(acc)	Exp. vs. Other F-score(acc)	Temp. vs. Other F-score(acc)	Average
Best result of Naïve Bayes	20.60(72.33)	48.91(60.15)	55.48(55.15)	18.75(38.62)	35.94(56.56)
Best result of MaxEnt	24.45(50.99)	50.37(62.42)	63.44(59.40)	16.91(35.98)	38.79(52.20)
Results of just using language model to predict implicit discourse relation in [9]	21.91(52.84)	39.53(50.85)	68.84(52.93)	11.91(6.33)	35.55(40.74)

TABLE V
THE PERFORMANCE OF DIFFERENT RELATIONS AND DIFFERENT MACHINE LEARNING ALGORITHMS IN BIODRB

Method	Comp. vs. Other F-score(acc)	Cont. vs. Other F-score(acc)	Exp. vs. Other F-score(acc)	Temp. vs. Other F-score(acc)	Average
Best result of Naïve Bayes	7.69(61.45)	3.88(60.24)	68.54(55.02)	17.19(57.43)	24.33(58.54)
Best result of MaxEnt	7.08(57.83)	3.36(53.82)	72.32(60.64)	23.81(61.45)	26.64(58.44)

Due to the lack of enough training data, the performance of the connective classifiers, which are used to predict the low-frequency connective is worse than the others. It decreases the overall performance of connective classifier. Thus, the general performance of connective classifier may be further improved by increasing the number of training data, especially the training data of low-frequency connectives.

Relation classifier: The size of training data for relation classifier is shown in Table I and Table II. We find that the data distributions of the four relations are significantly skewed. In PDTB, *Expansion* relation has the largest training data set, while *Temporal* makes up about only 5% of instances. In BioDRB, most instances are labeled as *Expansion* relation and the other three relations only count for about 20%.

From Table IV and Table V, we find that the performance of each relation closely depends on the size of training data.

B. Ambiguity

In this section, we analyze some ambiguity existing in implicit discourse relation which makes implicit relation classification a challenging task.

1) Ambiguity of implicit discourse relation: In PDTB, if the annotators have identified more than one simultaneous interpretation of implicit discourse relation, more than one sense tags are provided, called *multiple sense tag*. For example,

(**E4**) **Relation**: Implicit
Arg1: The gatherings in East Berlin and elsewhere were viewed as part of a government effort to stop activists from staging protests to press their demands.
Arg2: Dissidents in Czechoslovakia said the nation's pro-democracy movement was growing despite the government's move to crush a protest Saturday in Prague's Wenceslas Square.
Connective:Meanwhile
Sense:Expansion.Conjunction,
Temporal.Synchrony
(Expansion,Temporal-wsj_0587)

In E4, the annotators identify *Expansion* and *Temporal* relation in one instance.

In PDTB, 4.05% of Comparison relation have multiple sense tag. 2.73% Contingency relation and 3.36% Expansion relation have multiple sense tags. For Temporal, the number is over 21% [4]. This may be a possible reason for the weak performance in Temporal.

2) The ambiguity of implicit discourse connective: In some cases, an implicit discourse connective can be used to express more than one relation. For example, *while* appears 597 times in implicit discourse relation. Among them, it expresses Comparison relation for 179 times, Expansion relation for 343 times and Temporal relation for 7 times. For the remaining 68 times it is assigned multi sense tags [4].

For implicit discourse connectives with high degree of ambiguity, we cannot classify implicit discourse relation correctly by just using them.

C. Comparison between language model method and the proposed method

Using a language model to predict the connective [8] [9] has some shortcomings. First, the performance of this method is highly dependent to some parameters. For example, they selected the best N connectives that received the highest score in the language model as the predicted implicit connective. Obviously, the output of the model was highly relevant to the parameter N. Second, the result of the language model only depends on the word sequence of the sentence. Hence, the performance of the language model based method largely depends on the training data. In [8] and [9], the language model trained by different data performed differently. Third, it has nothing to do which the linguistically informed features, such as syntactic structure of the sentences, which are proved to be good indicators for implicit discourse relation [6] [19] and [10].

On the other hand, our method depends on the machine learning approach and linguistically informed features in use. Thus, it can overcome the above shortcomings. To begin with, our method uses machine learning to predict the connective and classify the relation. There is no parameter to choose in our method. Hence our model is more robust. Moreover, our method contains rich linguistically informed features, such as,

the syntactic structure of the span, verbs and context words, etc. Experiment has shown that our method achieved better performance than the language model method.

VII. Conclusion

Based on the importance of implicit discourse connective, we propose a two-step machine learning based classification approach. In the first step, connective classifiers are built to predict connectives using linguistically informed features. In the second step, a relation classifier is used to classify the discourse relation with the predicted connectives. The implicit connectives predicted in the first step act as a bridge to be used by the second step for implicit relation classification.

We evaluate the performance of our method in both open-domain (PDTB) and biomedical-domain (BioDRB). In PDTB, the experiment shows that our method can significantly improve the performance of implicit discourse relation classification and it achieves a 3.24% improvement over the previous work [9]. In BioDRB, we find that the performance is closely related to the size of training data. It can be improved if more training data are provided.

Acknowledgements

The authors would like to thank the reviewers for their helpful suggestions and comments, which improved the final version of this paper. This research is supported by grants from National Natural Science Foundation of China (No.60903093) and Doctoral Fund of Ministry of Education of China (No.20090076120029).

References

[1] S. Verberne, L. Boves, N. Oostdijk, and P. Coppen, "Evaluating discourse-based answer extraction for why-question answering." ACM, 2007, pp. 735–736, proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval.

[2] D. Marcu, "Improving summarization through rhetorical parsing tuning," 1998, pp. 206–215, the 6th Workshop on Very Large Corpora.

[3] E. Hovy, "Automated discourse generation using discourse structure relations," *Artificial intelligence*, vol. 63, no. 1-2, pp. 341–385, 1993.

[4] PDTB-Group, "The penn discourse treebank 2.0 annotation manual," Institute for Research in Cognitive Science, University of Pennsylvania, Tech. Rep., 2008.

[5] E. Pitler and A. Nenkova, "Using syntax to disambiguate explicit discourse connectives in text." Association for Computational Linguistics, 2009, pp. 13–16, proceedings of the ACL-IJCNLP 2009 Conference Short Papers.

[6] E. Pitler, A. Louis, and A. Nenkova, "Automatic sense prediction for implicit discourse relations in text." Association for Computational Linguistics, 2009, pp. 683–691, proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2-Volume 2.

[7] C. Sporleder and A. Lascarides, "Using automatically labelled examples to classify rhetorical relations: An assessment," *Natural Language Engineering*, vol. 14, no. 03, pp. 369–416, 2008.

[8] Z. Zhou, Y. Xu, Z. Niu, M. Lan, J. Su, and C. Tan, "Predicting discourse connectives for implicit discourse relation recognition." Association for Computational Linguistics, 2010, pp. 1507–1514, proceedings of the 23rd International Conference on Computational Linguistics: Posters.

[9] Z. Zhou, M. Lan, Z. Niu, Y. Xu, and J. Su, "The effects of discourse connectives prediction on implicit discourse relation recognition." Association for Computational Linguistics, 2010, pp. 139–146, proceedings of the 11th Annual Meeting of the Special Interest Group on Discourse and Dialogue.

[10] D. Marcu and A. Echihabi, "An unsupervised approach to recognizing discourse relations." Association for Computational Linguistics, 2002, pp. 368–375, proceedings of the 40th Annual Meeting on Association for Computational Linguistics.

[11] M. Saito, K. Yamamoto, and S. Sekine, "Using phrasal patterns to identify discourse relations." Association for Computational Linguistics, 2006, pp. 133–136, proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers on XX.

[12] S. Blair-Goldensohn, "Long-answer question answering and rhetorical-semantic relations," Ph.D. dissertation, 2007.

[13] R. Soricut and D. Marcu, "Sentence level discourse parsing using syntactic and lexical information." Association for Computational Linguistics, 2003, pp. 149–156, proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1.

[14] L. Carlson, D. Marcu, and M. Okurowski, "Building a discourse-tagged corpus in the framework of rhetorical structure theory." Association for Computational Linguistics, 2001, pp. 1–10, proceedings of the Second SIGdial Workshop on Discourse and Dialogue-Volume 16.

[15] W. Mann and S. Thompson, "Rhetorical structure theory: Toward a functional theory of text organization," *Text-Interdisciplinary Journal for the Study of Discourse*, vol. 8, no. 3, pp. 243–281, 1988.

[16] B. Wellner, J. Pustejovsky, C. Havasi, A. Rumshisky, and R. Sauri, "Classification of discourse coherence relations: An exploratory study using multiple knowledge sources." Association for Computational Linguistics, 2006, pp. 117–125, proceedings of the 7th SIGdial Workshop on Discourse and Dialogue.

[17] F. Wolf, E. Gibson, A. Fisher, and M. Knight, "The discourse graphbank: A database of texts annotated with coherence relations," *Linguistic Data Consortium*, 2005.

[18] R. Prasad, N. Dinesh, A. Lee, E. Miltsakaki, L. Robaldo, A. Joshi, and B. Webber, "The penn discourse treebank 2.0," in *In Proceedings of LREC*, 2008.

[19] Z. Lin, M. Kan, and H. Ng, "Recognizing implicit discourse relations in the penn discourse treebank." Association for Computational Linguistics, 2009, pp. 343–351, proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1.

[20] B. Wang, X. Wang, C. Sun, B. Liu, and L. Sun, "Modeling semantic relevance for question-answer pairs in web social communities." Association for Computational Linguistics, 2010, pp. 1230–1238, proceedings of the 48th Annual Meeting of the Association for Computational Linguistics.

[21] H. Hernault, D. Bollegala, and M. Ishizuka, "A semi-supervised approach to improve classification of infrequent discourse relations using feature vector extension." Association for Computational Linguistics, 2010, pp. 399–409, proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing.

[22] H. Hernault, D. Bollegala, and M. Ishizuka, "Semi-supervised discourse relation classification with structural learning," in *Proceedings of the 12th international conference on Computational linguistics and intelligent text processing - Volume Part I*, ser. CICLing'11. Berlin, Heidelberg: Springer-Verlag, 2011, pp. 340–352.

[23] R. Prasad, S. McRoy, N. Frid, A. Joshi, and H. Yu, "The biomedical discourse relation bank," *BMC bioinformatics*, vol. 12, no. 1, p. 188, 2011.

[24] J. Kim, T. Ohta, Y. Tateisi, and J. Tsujii, "Genia corpus - semantically annotated corpus for bio-textmining," *Bioinformatics*, vol. 19, no. suppl 1, pp. i180–182, 2003.

[25] W. Wang, J. Su, and C. Tan, "Kernel based discourse relation recognition with temporal ordering information." Association for Computational Linguistics, 2010, pp. 710–719, proceedings of the 48th Annual Meeting of the Association for Computational Linguistics.

[26] B. Levin, *English verb classes and alternations: A preliminary investigation*. University of Chicago press Chicago, IL:, 1993, vol. 348.

[27] A. K. McCallum, "Mallet: A machine learning for language toolkit," 2002, http://mallet.cs.umass.edu.