Leveraging Synthetic Discourse Data via Multi-task Learning for Implicit Discourse Relation Recognition

Man Lan and **Yu Xu**
Department of Computer Science and Technology
East China Normal University
Shanghai, P.R.China
mlan@cs.ecnu.edu.cn
51101201049@ecnu.cn

Zheng-Yu Niu
Baidu Inc.
Beijing, P.R.China
niuzhengyu@baidu.com

Abstract

To overcome the shortage of labeled data for implicit discourse relation recognition, previous works attempted to automatically generate training data by removing explicit discourse connectives from sentences and then built models on these synthetic implicit examples. However, a previous study (Sporleder and Lascarides, 2008) showed that models trained on these synthetic data do not generalize very well to natural (i.e. *genuine*) implicit discourse data. In this work we revisit this issue and present a multi-task learning based system which can effectively use synthetic data for implicit discourse relation recognition. Results on PDTB data show that under the multi-task learning framework our models with the use of the prediction of explicit discourse connectives as auxiliary learning tasks, can achieve an averaged F_1 improvement of 5.86% over baseline models.

1 Introduction

The task of implicit discourse relation recognition is to identify the type of discourse relation (a.k.a. *rhetorical relation*) hold between two spans of text, where there is no discourse connective (a.k.a. *discourse marker*, e.g., *but*, *and*) in context to explicitly mark their discourse relation (e.g., *Contrast* or *Explanation*). It can be of great benefit to many downstream NLP applications, such as question answering (QA) (Verberne et al., 2007), information extraction (IE) (Cimiano et al., 2005), and machine translation (MT), etc. This task is quite challenging due to two reasons. First, without discourse connective in text, the task is quite difficult in itself. Second, implicit discourse relation is quite frequent in text. For example, almost half the sentences in the British National Corpus held implicit discourse relations (Sporleder and Lascarides, 2008). Therefore, the task of implicit discourse relation recognition is the key to improving end-to-end discourse parser performance.

To overcome the shortage of manually annotated training data, (Marcu and Echihabi, 2002) proposed a pattern-based approach to automatically generate training data from raw corpora. This line of research was followed by (Sporleder and Lascarides, 2008) and (Blair-Goldensohn, 2007). In these works, sentences containing certain words or phrases (e.g. *but*, *although*) were selected out from raw corpora using a pattern-based approach and then these words or phrases were removed from these sentences. Thus the resulting sentences were used as synthetic training examples for implicit discourse relation recognition. Since there is ambiguity of a word or phrase serving for discourse connective (i.e., the ambiguity between discourse and non-discourse usage or the ambiguity between two or more discourse relations if the word or phrase is used as a discourse connective), the synthetic implicit data would contain a lot of noises. Later, with the release of manually annotated corpus, such as Penn Discourse Treebank 2.0 (PDTB) (Prasad et al., 2008), recent studies performed implicit discourse relation recognition on natural (i.e., *genuine*) implicit discourse data (Pitler et al., 2009) (Lin et al., 2009) (Wang et al., 2010) with the use of linguistically informed features and machine learning algorithms.

(Sporleder and Lascarides, 2008) conducted a study of the pattern-based approach presented by (Marcu and Echihabi, 2002) and showed that the model built on synthetical implicit data has not generalize well on natural implicit data. They found some evidence that this behavior is largely independent of the classifiers used and seems to lie in the data itself (e.g., marked and unmarked examples may be too dissimilar linguistically and

Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, pages 476–485,
Sofia, Bulgaria, August 4-9 2013. ©2013 Association for Computational Linguistics

removing unambiguous markers in the automatic labelling process may lead to a meaning shift in the examples). We state that in some cases it is true while in other cases it may not always be so. A simple example is given here:

(**E1**) a. We can't win.
b. [***but***] We must keep trying.

We may find that in this example whether the insertion or the removal of connective *but* would not lead to a redundant or missing information between the above two sentences. That is, discourse connectives can be inserted between or removed from two sentences without changing the semantic relations between them in some cases. Another similar observation is in the annotation procedure of PDTB. To label implicit discourse relation, annotators inserted connective which can best express the relation between sentences without any redundancy[1]. We see that there should be some linguistical similarities between explicit and implicit discourse examples. Therefore, the first question arises: can we exploit this kind of linguistic similarity between explicit and implicit discourse examples to improve implicit discourse relation recognition?

In this paper, we propose a multi-task learning based method to improve the performance of implicit discourse relation recognition (as main task) with the help of relevant auxiliary tasks. Specifically, the main task is to recognize the implicit discourse relations based on genuine implicit discourse data and the auxiliary task is to recognize the implicit discourse relations based on synthetic implicit discourse data. According to the principle of multi-task learning, the learning model can be optimized by the shared part of the main task and the auxiliary tasks without bring unnecessary noise. That means, the model can learn from synthetic implicit data while it would not bring unnecessary noise from synthetic implicit data.

Although (Sporleder and Lascarides, 2008) did not mention, we speculate that another possible reason for the reported worse performance may result from noises in synthetic implicit discourse data. These synthetic data can be generated from two sources: (1) raw corpora with the use of pattern-based approach in (Marcu and Echihabi, 2002) and (Sporleder and Lascarides, 2008), and (2) manually annotated explicit data with the removal of explicit discourse connectives. Obviously, the data generated from the second source is cleaner and more reliable than that from the first source. Therefore, the second question to address in this work is: whether synthetic implicit discourse data generated from explicit discourse data source (i.e., the second source) can lead to a better performance than that from raw corpora (i.e., the first source)? To answer this question, we will make a comparison of synthetic discourse data generated from two corpora, i.e., the BILLIP corpus and the explicit discourse data annotated in PDTB.

The rest of this paper is organized as follows. Section 2 reviews related work on implicit discourse relation classification and multi-task learning. Section 3 presents our proposed multi-task learning method for implicit discourse relation classification. Section 4 provides the implementation technique details of the proposed multi-task method. Section 5 presents experiments and discusses results. Section 6 concludes this work.

2 Related Work

2.1 Implicit discourse relation classification

2.1.1 Unsupervised approaches

Due to the lack of benchmark data for implicit discourse relation analysis, earlier work used unlabeled data to generate synthetic implicit discourse data. For example, (Marcu and Echihabi, 2002) proposed an unsupervised method to recognize four discourse relations, i.e., *Contrast*, *Explanation-evidence*, *Condition* and *Elaboration*. They first used unambiguous pattern to extract explicit discourse examples from raw corpus. Then they generated synthetic implicit discourse data by removing explicit discourse connectives from sentences extracted. In their work, they collected word pairs from synthetic data set as features and used machine learning method to classify implicit discourse relation. Based on this work, several researchers have extended the work to improve the performance of relation classification. For example, (Saito et al., 2006) showed that the use of phrasal patterns as additional features can help a word-pair based system for discourse relation prediction on a Japanese corpus. Furthermore, (Blair-Goldensohn, 2007) improved previous work with the use of parameter optimization,

[1] According to the PDTB Annotation Manual (PDTB-Group, 2008), if the insertion of connective leads to "redundancy", the relation is annotated as Alternative lexicalizations (AltLex), not implicit.

topic segmentation and syntactic parsing. However, (Sporleder and Lascarides, 2008) showed that the training model built on a synthetic data set, like the work of (Marcu and Echihabi, 2002), may not be a good strategy since the linguistic dissimilarity between explicit and implicit data may hurt the performance of a model on natural data when being trained on synthetic data.

2.1.2 Supervised approaches

This line of research work approaches this relation prediction problem by recasting it as a classification problem. (Soricut and Marcu, 2003) parsed the discourse structures of sentences on RST Bank data set (Carlson et al., 2001) which is annotated based on Rhetorical Structure Theory (Mann and Thompson, 1988). (Wellner et al., 2006) presented a study of discourse relation disambiguation on GraphBank (Wolf et al., 2005). Recently, (Pitler et al., 2009) (Lin et al., 2009) and (Wang et al., 2010) conducted discourse relation study on PDTB (Prasad et al., 2008) which has been widely used in this field.

2.1.3 Semi-supervised approaches

Research work in this category exploited both labeled and unlabeled data for discourse relation prediction. (Hernault et al., 2010) presented a semi-supervised method based on the analysis of co-occurring features in labeled and unlabeled data. Very recently, (Hernault et al., 2011) introduced a semi-supervised work using structure learning method for discourse relation classification, which is quite relevant to our work. However, they performed discourse relation classification on both explicit and implicit data. And their work is different from our work in many aspects, such as, feature sets, auxiliary task, auxiliary data, class labels, learning framework, and so on. Furthermore, there is no explicit conclusion or evidence in their work to address the two questions raised in Section 1.

Unlike their previous work, our previous work (Zhou et al., 2010) presented a method to predict the missing connective based on a language model trained on an unannotated corpus. The predicted connective was then used as a feature to classify the implicit relation.

2.2 Multi-task learning

Multi-task learning is a kind of machine learning method, which learns a main task together with other related auxiliary tasks at the same time, using a shared representation. This often leads to a better model for the main task, because it allows the learner to use the commonality among the tasks. Many multi-task learning methods have been proposed in recent years, (Ando and Zhang, 2005a), (Argyriou et al., 2008), (Jebara, 2004), (Bonilla et al., 2008), (Evgeniou and Pontil, 2004), (Baxter, 2000), (Caruana, 1997), (Thrun, 1996). One group uses task relations as regularization terms in the objective function to be optimized. For example, in (Evgeniou and Pontil, 2004) the regularization terms make the parameters of models closer for similar tasks. Another group is proposed to find the common structure from data and then utilize the learned structure for multi-task learning (Argyriou et al., 2008) (Ando and Zhang, 2005b).

3 Multi-task Learning for Discourse Relation Prediction

3.1 Motivation

The idea of using multi-task learning for implicit discourse relation classification is motivated by the observations that we have made on implicit discourse relation.

On one hand, since building a hand-annotated implicit discourse relation corpus is costly and time consuming, most previous work attempted to use synthetic implicit discourse examples as training data. However, (Sporleder and Lascarides, 2008) found that the model trained on synthetic implicit data has not performed as well as expected in natural implicit data. They stated that the reason is linguistic dissimilarity between explicit and implicit discourse data. This indicates that straightly using synthetic implicit data as training data may not be helpful.

On the other hand, as shown in Section 1, we observe that in some cases explicit discourse relation and implicit discourse relation can express the same meaning with or without a discourse connective. This indicates that in certain degree they must be similar to each other. If it is true, the synthetic implicit relations are expected to be helpful for implicit discourse relation classification. Therefore, what we have to do is to find a way to train a model which has the capabilities to learn from their similarity and to ignore their dissimilarity as well.

To solve it, we propose a multi-task learning method for implicit discourse relation classification, where the classification model seeks the shared part through jointly learning main task and multiple auxiliary tasks. As a result, the model can be optimized by the similar shared part without bringing noise in the dissimilar part. Specifically, in this work, we use alternating structure optimization (ASO) (Ando and Zhang, 2005a) to construct the multi-task learning framework. ASO has been shown to be useful in a *semi-supervised learning* configuration for several NLP applications, such as, text chunking (Ando and Zhang, 2005b) and text classification (Ando and Zhang, 2005a).

3.2 Multi-task learning and ASO

Generally, multi-task learning(MTL) considers m prediction problems indexed by $\ell \in \{1, ..., m\}$, each with n_ℓ samples (X_i^ℓ, Y_i^ℓ) for $i \in \{1, ...n_\ell\}$ (X_i are input *feature vectors* and Y_i are corresponding *classification labels*) and assumes that there exists a common predictive structure shared by these m problems. Generally, the joint linear model for MTL is to predict problem ℓ in the following form:

$$f_\ell(\Theta, X) = w_\ell^T X + v_\ell^T \Theta X, \Theta\Theta^T = I, \quad (1)$$

where I is the identity matrix, w_ℓ and v_ℓ are weight vectors specific to each problem ℓ, and Θ is the structure matrix shared by all the m predictors. The main goal of MTL is to learn a common good feature map ΘX for all the m problems. Several MTL methods have been presented to learn ΘX for all the m problems. In this work, we adopt the ASO method.

Specifically, the ASO method adopted *singular value decomposition* (SVD) to obtain Θ and m predictors that minimize the empirical risk summed over all the m problems. Thus, the problem of optimization becomes the minimization of the joint empirical risk written as:

$$\sum_{\ell=1}^{m} \Big(\sum_{i=1}^{n_\ell} \frac{L(f_\ell(\Theta, X_i^\ell), Y_i)}{n_\ell} + \lambda ||W_\ell||^2 \Big) \quad (2)$$

where loss function $L(.)$ quantifies the difference between the prediction $f(X_i)$ and the true output Y_i for each predictor, and λ is a regularization parameter for square regularization to control the model complexity. To minimize the empirical risk, ASO repeats the following alternating optimization procedure until a convergence criterion is met:

1) Fix (Θ, V_ℓ), and find m predictors f_ℓ that minimize the above joint empirical risk.

2) Fix m predictors f_ℓ, and find (Θ, V_ℓ) that minimizes the above joint empirical risk.

3.3 Auxiliary tasks

There are two main principles to create auxiliary tasks. First, the auxiliary tasks should be automatically labeled in order to reduce the cost of manual labeling. Second, since the MTL model learns from the shared part of main task and auxiliary tasks, the auxiliary tasks should be quite relevant/similar to the main task. It is generally believed that the more the auxiliary tasks are relevant to the main task, the more the main task can benefit from the auxiliary tasks. Following these two principles, we create the auxiliary tasks by generating automatically labeled data as follows.

Previous work (Marcu and Echihabi, 2002) and (Sporleder and Lascarides, 2008) adopted predefined pattern-based approach to generate synthetic labeled data, where each predefined pattern has one discourse relation label. In contrast, we adopt an automatic approach to generate synthetic labeled data, where each discourse connective between two texts serves as their relation label. The reason lies in the very strong connection between discourse connectives and discourse relations. For example, the connective *but* always indicates a *contrast* relation between two texts. And (Pitler et al., 2008) proved that using only connective itself, the accuracy of explicit discourse relation classification is over 93%.

To build the mapping between discourse connective and discourse relation, for each connective, we count the times it appears in each relation and regard the relation in which it appears most frequently as its most relevant relation. Based on this mapping between connective and relation, we extract the synthetic labeled data containing the connective as training data for auxiliary tasks.

For example, *and* appears $3,000$ times in PDTB as a discourse connective. Among them, it is manually annotated as an *Expansion* relation for $2,938$ times. So we regard the *Expansion* relation as its most relevant relation and generate a mapping pattern like: "*and* → *Expansion*". Then we extract all sentences which contain discourse "*and*" and remove this connective "*and*" from sentences to generate synthetic implicit data. The resulting sentences are used in auxiliary task and automatically marked as *Expansion* relation.

4 Implementation Details of Multi-task Learning Method

4.1 Data sets for *main* and *auxiliary* tasks

To examine whether there is a difference in synthetic implicit data generated from unannotated and annotated corpus, we use two corpora. One is a hand-annotated explicit discourse corpus, i.e., the explicit discourse relations in PDTB, denoted as *exp*. Another is an unannotated corpus, i.e., BLLIP (David McClosky and Johnson., 2008).

4.1.1 Penn Discourse Treebank

PDTB (Prasad et al., 2008) is the largest hand-annotated corpus of discourse relation so far. It contains $2,312$ Wall Street Journal (WSJ) articles. The sense label of discourse relations is hierarchically with three levels, i.e., *class*, *type* and *subtype*. The top level contains four major semantic *classes*: *Comparison* (denoted as *Comp.*), *Contingency* (*Cont.*), *Expansion* (*Exp.*) and *Temporal* (*Temp.*). For each class, a set of *types* is used to refine relation sense. The set of *subtypes* is to further specify the semantic contribution of each argument. In this paper, we focus on the top level (*class*) and the second level (*type*) relations because the *subtype* relations are too fine-grained and only appear in some relations.

Both explicit and implicit discourse relations are labeled in PDTB. In our experiment, the implicit discourse relations are used in the main task and for evaluation. While the explicit discourse relations are used in the auxiliary task. A detailed description of the data sources for different tasks is given below.

Data set for main task Following previous work in (Pitler et al., 2009) and (Zhou et al., 2010), the implicit relations in sections 2-20 are used as training data for the main task (denoted as *imp*) and the implicit relations in sections 21-22 are for evaluation. Table 1 shows the distribution of implicit relations. There are too few training instances for six second level relations (indicated by * in Table 1), so we removed these six relations in our experiments.

Data set for auxiliary task All explicit instances in sections 00-24 in PDTB, i.e., $18,459$ instances, are used for auxiliary task (denoted as *exp*). Following the method described in Section 3.3, we build the mapping patterns between connectives and relations in PDTB and generate synthetic labeled data by removing the connectives. According to the most relevant relation sense of connective removed, the resulting instances are grouped into different data sets.

Top level	Second level	train	test
Temp		736	83
	Synchrony	203	28
	Asynchronous	532	55
Cont		3333	279
	Cause	3270	272
	*Pragmatic Cause**	64	7
	*Condition**	1	0
	*Pragmatic condition**	1	0
Comp		1939	152
	Contrast	1607	134
	*Pragmatic contrast**	4	0
	Concession	183	17
	*Pragmatic concession**	1	0
Exp		6316	567
	Conjunction	2872	208
	Instantiation	1063	119
	Restatement	2405	213
	Alternative	147	9
	*Exception**	0	0
	List	338	12

Table 1: Distribution of implicit discourse relations in the top and second level of PDTB

4.1.2 BLLIP

BLLIP North American News Text (Complete) is used as unlabeled data source to generate synthetic labeled data. In comparison with the synthetic labeled data generated from the explicit relations in PDTB, the synthetic labeled data from BLLIP contains more noise. This is because the former data is manually annotated whether a word serves as discourse connective or not, while the latter does not manually disambiguate two types of ambiguity, i.e., whether a word serves as discourse connective or not, and the type of discourse relation if it is a discourse connective. Finally, we extract $26,412$ instances from BLLIP (denoted as *BLLIP*) and use them for auxiliary task.

4.2 Feature representation

For both main task and auxiliary tasks, we adopt the following three feature types. These features are chosen due to their superior performance in previous work (Pitler et al., 2009) and our previous work (Zhou et al., 2010).

Verbs: Following (Pitler et al., 2009), we extract the pairs of verbs from both text spans. The number of verb pairs which have the same highest

Levin verb class levels (Levin, 1993) is counted as a feature. Besides, the average length of verb phrases in each argument is included as a feature. In addition, the part of speech tags of the main verbs (e.g., base form, past tense, 3rd person singular present, etc.) in each argument, i.e., MD, VB, VBD, VBG, VBN, VBP, VBZ, are recorded as features, where we simply use the first verb in each argument as the main verb.

Polarity: This feature records the number of positive, negated positive, negative and neutral words in both arguments and their cross product as well. For negated positives, we first locate the negated words in text span and then define the closely behind positive word as negated positive. The polarity of each word in arguments is derived from Multi-perspective Question Answering Opinion Corpus (MPQA) (Wilson et al., 2009).

Modality: We examine six modal words (i.e., *can*, *may*, *must*, *need*, *shall*, *will*) including their various tenses or abbreviation forms in both arguments. This feature records the presence or absence of modal words in both arguments and their cross product.

4.3 Classifiers used multi-task learning

We extract the above linguistically informed features from two synthetic implicit data sets (i.e., *BLLIP* and *exp*) to learn the auxiliary classifier and from the natural implicit data set (i.e., *imp*) to learn the main classifier. Under the ASO-based multi-task learning framework, the model of main task learns from the shared part of main task and auxiliary tasks. Specifically, we adopt multiple binary classification to build model for main task. That is, for each discourse relation, we build a binary classifier.

5 Experiments and Results

5.1 Experiments

Although previous work has been done on PDTB (Pitler et al., 2009) and (Lin et al., 2009), we cannot make a direct comparison with them because various experimental conditions, such as, different classification strategies (multi-class classification, multiple binary classification), different data preparation (feature extraction and selection), different benchmark data collections (different sections for training and test, different levels of discourse relations), different classifiers with various parameters (MaxEnt, Naïve Bayes, SVM, etc) and even different evaluation methods (F_1, accuracy) have been adopted by different researchers.

Therefore, to address the two questions raised in Section 1 and to make the comparison reliable and reasonable, we performed experiments on the top and second level of PDTB using single task learning and multi-task learning, respectively. The systems using single task learning serve as baseline systems. Under the single task learning, various combinations of *exp* and *BLLIP* data are incorporated with *imp* data for the implicit discourse relation classification task.

We hypothesize that synthetical implicit data would contribute to the main task, i.e., the implicit discourse relation classification. Specifically, the natural implicit data (i.e., *imp*) are used to create main task and the synthetical implicit data (*exp* or *BLLIP*) are used to create auxiliary tasks for the purpose of optimizing the objective functions of main task. If the hypothesis is correct, the performance of main task would be improved by auxiliary tasks created from synthetical implicit data. Thus in the experiments of multi-task learning, only natural implicit examples (i.e., *imp*) data are used for main task training while different combinations of synthetical implicit examples (*exp* and *BLLIP*) are used for auxiliary task training.

We adopt precision, recall and their combination F_1 for performance evaluation. We also perform one-tailed t-test to validate if there is significant difference between two methods in terms of F_1 performance analysis.

5.2 Results

Table 2 summarizes the experimental results under single and multi-task learning on the top level of four PDTB relations with respect to different combinations of synthetic implicit data. For each relation, the first three rows indicate the results of using different single training data under single task learning and the last three rows indicate the results using different combinations of training data under single task and multi-task learning. The best F_1 for every relation is shown in bold font. From this table, we can find that on four relations, our multi-task learning systems achieved the best performance using the combination of *exp* and *BLLIP* synthetic data.

Table 3 summarizes the best single task and the best multi-task learning results on the second level of PDTB. For four relations, i.e., Synchrony, Con-

Level 1 class	Single-task				Multi-task				
	Data	P	R	F_1	Data (*main*)	Data (*aux*)	P	R	F_1
Comp.	*imp*	21.43	37.50	27.27	-	-	-	-	-
	BLLIP	12.68	53.29	20.48	-	-	-	-	-
	exp	15.25	50.66	23.44	-	-	-	-	-
	imp + exp	16.94	40.13	23.83	*imp*	*exp*	22.94	49.34	30.90
	imp + BLLIP	13.56	44.08	20.74	*imp*	*BLLIP*	20.47	63.16	30.92
	imp + exp + BLLIP	14.54	38.16	21.05	*imp*	*exp + BLLIP*	23.47	48.03	**31.53**
Cont.	*imp*	37.65	43.73	40.46	-	-	-	-	-
	BLLIP	33.72	31.18	32.40	-	-	-	-	-
	exp	35.24	26.52	30.27	-	-	-	-	-
	imp + exp	39.00	13.98	20.58	*imp*	*exp*	39.94	45.52	42.55
	imp + BLLIP	37.30	24.73	29.74	*imp*	*BLLIP*	37.80	63.80	47.47
	imp + exp + BLLIP	39.37	31.18	34.80	*imp*	*exp + BLLIP*	35.90	70.25	**47.52**
Exp.	*imp*	56.59	66.67	61.21	-	-	-	-	-
	BLLIP	53.29	40.04	45.72	-	-	-	-	-
	exp	57.97	58.38	58.17	-	-	-	-	-
	imp + exp	57.32	65.61	61.18	*imp*	*exp*	59.14	67.90	63.22
	imp + BLLIP	56.28	65.61	60.59	*imp*	*BLLIP*	53.80	99.82	69.92
	imp + exp + BLLIP	55.81	65.26	60.16	*imp*	*exp + BLLIP*	53.90	99.82	**70.01**
Temp.	*imp*	16.46	63.86	26.17	-	-	-	-	-
	BLLIP	17.31	43.37	24.74	-	-	-	-	-
	exp	15.46	36.14	21.66	-	-	-	-	-
	imp + exp	15.35	39.76	22.15	*imp*	*exp*	18.60	63.86	28.80
	imp + BLLIP	14.74	33.73	20.51	*imp*	*BLLIP*	18.12	67.47	28.57
	imp + exp + BLLIP	15.94	39.76	22.76	*imp*	*exp + BLLIP*	19.08	65.06	**29.51**

Table 2: Performance of precision, recall and F_1 for 4 Level 1 relation classes. “-” indicates N.A.

Level 2 type	Single-task				Multi-task				
	Data	P	R	F_1	Data (*main*)	Data (*aux*)	P	R	F_1
Asynchronous	*imp*	11.36	74.55	19.71	*imp*	*exp + BLLIP*	23.08	21.82	**22.43**
Synchrony	*imp*	-	-	-	*imp*	*exp + BLLIP*	-	-	-
Cause	*imp*	36.38	64.34	46.48	*imp*	*exp + BLLIP*	36.01	67.65	**47.00**
Contrast	*imp*	20.07	42.54	27.27	*imp*	*exp + BLLIP*	20.70	52.99	**29.77**
Concession	*imp*	-	-	-	*imp*	*exp + BLLIP*	-	-	-
Conjunction	*imp*	26.35	63.46	37.24	*imp*	*exp + BLLIP*	26.29	73.56	**38.73**
Instantiation	*imp*	22.78	53.78	32.00	*imp*	*exp + BLLIP*	22.55	57.98	**32.47**
Restatement	*imp*	23.11	67.61	34.45	*imp*	*exp + BLLIP*	26.93	53.99	**35.94**
Alternative	*imp*	-	-	-	*imp*	*exp + BLLIP*	-	-	-
List	*imp*	-	-	-	*imp*	*exp + BLLIP*	-	-	-

Table 3: Performance of precision, recall and F_1 for 10 Level 2 relation types. “-” indicates 0.00.

cession, Alternative and List, the classifier labels no instances due to the small percentages for these four types.

Table 4 summarizes the one-tailed t-test results on the top level of PDTB between the best single task learning system (i.e., *imp*) and three multi-task learning systems (*imp:exp+BLLIP* indicates that *imp* is used for main task and the combination of *exp* and *BLLIP* are for auxiliary task). The systems with insignificant performance differences are grouped into one set and ”>” and ”>>” denote better than at significance level 0.01 and 0.001 respectively.

5.3 Discussion

From Table 2 to Table 4, several findings can be found as follows.

We can see that the multi-task learning systems perform consistently better than the single task learning systems for the prediction of implicit discourse relations. Our best multi-task learning system achieves an averaged F_1 improvement of 5.86% over the best single task learning system on the top level of PDTB relations. Specifically, for

Class	One-tailed t-test results
Comp.	(*imp:exp+BLLIP*, *imp:exp*, *imp:BLLIP*) >> (*imp*)
Cont.	(*imp:exp+BLLIP*, *imp:BLLIP*) >> (*imp:exp*) > (*imp*)
Exp.	(*imp:exp+BLLIP*, *imp:BLLIP*) >> (*imp:exp*) > (*imp*)
Temp.	(*imp:exp+BLLIP*, *imp:exp*, *imp:BLLIP*) >> (*imp*)

Table 4: Statistical significance tests results.

the relations *Comp.*, *Cont.*, *Exp.*, *Temp.*, our best multi-task learning system achieve 4.26%, 7.06%, 8.8% and 3.34% F_1 improvements over the best single task learning system. It indicates that using synthetic implicit data as auxiliary task greatly improves the performance of the main task. This is confirmed by the following t-tests in Table 4.

In contrast to the performance of multi-task learning, the performance of the best single task learning system has been achieved on natural implicit discourse data alone. This finding is consistent with (Sporleder and Lascarides, 2008). It indicates that under single task learning, directly adding synthetic implicit data to increase the number of training data cannot be helpful to implicit discourse relation classification. The possible reasons result from (1) the different nature of implicit and explicit discourse data in linguistics and (2) the noise brought from synthetic implicit data.

Based on the above analysis, we state that it is the way of utilizing synthetic implicit data that is important for implicit discourse relation classification.

Although all three multi-task learning systems outperformed single task learning systems, we find that the two synthetic implicit data sets have not been shown a universally consistent performance on four top level PDTB relations. On one hand, for the relations *Comp.* and *Temp.*, the performance of the two synthetic implicit data sets alone and their combination are comparable to each other and there is no significant difference between them. On the other hand, for the relations *Cont.* and *Exp.*, the performance of *exp* data is inferior to that of *BLLIP* and their combination. This is contrary to our original expectation that *exp* data which has been manually annotated for discourse connective disambiguation should outperform *BLLIP* which contains a lot of noise. This finding indicates that under the multi-task learning, it may not be worthy of using manually annotated corpus to generate auxiliary data. It is quite promising since it can provide benefits to reducing the cost of human efforts on corpus annotation.

5.4 Ambiguity Analysis

Although our experiments show that synthetic implicit data can help implicit discourse relation classification under multi-task learning framework, the overall performance is still quite low (44.64% in F_1). Therefore, we analyze the types of ambiguity in relations and connectives in order to motivate possible future work.

5.4.1 Ambiguity of implicit relation

Without explicit discourse connective, the implicit discourse relation instance can be understood in two or more different ways. Given the example E2 in PDTB, the PDTB annotators explain it as *Contingency* or *Expansion* relation and manually insert corresponding implicit connective *for one thing* or *because* to express its relation.

(**E2**) **Arg1**:Now the stage is set for the battle to play out
Arg2:The anti-programmers are getting some helpful thunder from Congress
Connective1:because
Sense1:Contingency.Cause.Reason
Connective2:for one thing
Sense2:Expansion.Instantiation
(wsj_0118)

Thus the ambiguity of implicit discourse relations makes this task difficult in itself.

5.4.2 Ambiguity of discourse connectives

As we mentioned before, even given an explicit discourse connective in text, its discourse relation still can be explained in two or more different ways. And for different connectives, the ambiguity of relation senses is quite different. That is, the most frequent sense is not always the only sense that a connective expresses. In example E3, "*since*" is explained by annotators to express *Temporal* or *Contingency* relation.

(**E3**) **Arg1**:MiniScribe has been on the rocks
Arg2:**since** it disclosed early this year that its earnings reports for 1988 weren't accurate.

Sense1:Temporal.Asynchronous.Succession
Sense2:Contingency.Cause.Reason
(wsj_0003)

In PDTB, "*since*" appears 184 times in explicit discourse relations. It expresses *Temporal* relation for 80 times, *Contingency* relation for 94 times and both *Temporal* and *Contingency* for 10 time (like example E3). Therefore, although we use its most frequent sense, i.e., *Contingency*, to automatically extract sentences and label them, almost less than half of them actually express *Temporal* relation. Thus the ambiguity of discourse connectives is another source which has brought noise to data when we generate synthetical implicit discourse relation.

6 Conclusions

In this paper, we present a multi-task learning method to improve implicit discourse relation classification by leveraging synthetic implicit discourse data. Results on PDTB show that under the framework of multi-task learning, using synthetic discourse data as auxiliary task significantly improves the performance of main task. Our best multi-task learning system achieves an averaged F_1 improvement of 5.86% over the best single task learning system on the top level of PDTB relations. Specifically, for the relations *Comp.*, *Cont.*, *Exp.*, *Temp.*, our best multi-task learning system achieves 4.26%, 7.06%, 8.8%, and 3.34% F_1 improvements over a state of the art baseline system. This indicates that it is the way of utilizing synthetic discourse examples that is important for implicit discourse relation classification.

Acknowledgements

This research is supported by grants from National Natural Science Foundation of China (No.60903093), Shanghai Pujiang Talent Program (No.09PJ1404500), Doctoral Fund of Ministry of Education of China (No. 20090076120029) and Shanghai Knowledge Service Platform Project (No. ZF1213).

References

R.K. Ando and T. Zhang. 2005a. A framework for learning predictive structures from multiple tasks and unlabeled data. *The Journal of Machine Learning Research*, 6:1817–1853.

R.K. Ando and T. Zhang. 2005b. A high-performance semi-supervised learning method for text chunking. pages 1–9. Association for Computational Linguistics. Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics.

A. Argyriou, C.A. Micchelli, M. Pontil, and Y. Ying. 2008. A spectral regularization framework for multi-task structure learning. *Advances in Neural Information Processing Systems*, 20:2532.

J. Baxter. 2000. A model of inductive bias learning. *J. Artif. Intell. Res. (JAIR)*, 12:149–198.

S.J. Blair-Goldensohn. 2007. *Long-answer question answering and rhetorical-semantic relations*. Ph.D. thesis.

E. Bonilla, K.M. Chai, and C. Williams. 2008. Multi-task gaussian process prediction. *Advances in Neural Information Processing Systems*, 20(October).

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

R. Caruana. 1997. Multitask learning. *Machine Learning*, 28(1):41–75.

P. Cimiano, U. Reyle, and J. Saric. 2005. Ontology-driven discourse analysis for information extraction. *Data and Knowledge Engineering*, 55(1):59–83.

Eugene Charniak David McClosky and Mark Johnson. 2008. Bllip north american news text, complete.

T. Evgeniou and M. Pontil. 2004. Regularized multi–task learning. pages 109–117. ACM. Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining.

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

H. Hernault, D. Bollegala, and M. Ishizuka. 2011. 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*, CICLing'11, pages 340–352, Berlin, Heidelberg. Springer-Verlag.

T. Jebara. 2004. Multi-task feature and kernel selection for svms. page 55. ACM. Proceedings of the twenty-first international conference on Machine learning.

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

Z. Lin, M.Y. Kan, and H.T. Ng. 2009. Recognizing implicit discourse relations in the penn discourse treebank. pages 343–351. Association for Computational Linguistics. Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1.

W.C. Mann and S.A. Thompson. 1988. Rhetorical structure theory: Toward a functional theory of text organization. *Text-Interdisciplinary Journal for the Study of Discourse*, 8(3):243–281.

D. Marcu and A. Echihabi. 2002. An unsupervised approach to recognizing discourse relations. pages 368–375. Association for Computational Linguistics. Proceedings of the 40th Annual Meeting on Association for Computational Linguistics.

PDTB-Group. 2008. The penn discourse treebank 2.0 annotation manual. Technical report, Institute for Research in Cognitive Science, University of Pennsylvania.

E. Pitler, M. Raghupathy, H. Mehta, A. Nenkova, A. Lee, and A. Joshi. 2008. Easily identifiable discourse relations. Citeseer. Proceedings of the 22nd International Conference on Computational Linguistics (COLING 2008), Manchester, UK, August.

E. Pitler, A. Louis, and A. Nenkova. 2009. Automatic sense prediction for implicit discourse relations in text. pages 683–691. Association for Computational Linguistics. 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.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. 2008. The penn discourse treebank 2.0. In *In Proceedings of LREC*.

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

R. Soricut and D. Marcu. 2003. Sentence level discourse parsing using syntactic and lexical information. pages 149–156. Association for Computational Linguistics. Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1.

C. Sporleder and A. Lascarides. 2008. Using automatically labelled examples to classify rhetorical relations: An assessment. *Natural Language Engineering*, 14(03):369–416.

S. Thrun. 1996. Is learning the n-th thing any easier than learning the first? *Advances in Neural Information Processing Systems*, pages 640–646.

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

W.T. Wang, J. Su, and C.L. Tan. 2010. Kernel based discourse relation recognition with temporal ordering information. pages 710–719. Association for Computational Linguistics. Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics.

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

T. Wilson, J. Wiebe, and P. Hoffmann. 2009. Recognizing contextual polarity: An exploration of features for phrase-level sentiment analysis. *Computational Linguistics*, 35(3):399–433.

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

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