Frontiers of Information Technology & Electronic Engineering
www.jzus.zju.edu.cn; engineering.cae.cn; www.springerlink.com
ISSN 2095-9184 (print); ISSN 2095-9230 (online)
E-mail: jzus@zju.edu.cn

Cross-lingual implicit discourse relation recognition with co-training*

Yao-jie LU[1,2,3], Mu XU[1], Chang-xing WU[1], De-yi XIONG[4], Hong-ji WANG[1], Jin-song SU[†‡1,2]

[1]*School of Software, Xiamen University, Xiamen 361005, China*

[2]*State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing 210023, China*

[3]*Institute of Software, Chinese Academy of Sciences, Beijing 100190, China*

[4]*Provincial Key Laboratory for Computer Information Processing Technology, Soochow University, Suzhou 215006, China*

[†]E-mail: jssu@xmu.edu.cn

Received Dec. 24, 2016; Revision accepted Apr. 10, 2017; Crosschecked May 8, 2018

Abstract: A lack of labeled corpora obstructs the research progress on implicit discourse relation recognition (DRR) for Chinese, while there are some available discourse corpora in other languages, such as English. In this paper, we propose a cross-lingual implicit DRR framework that exploits an available English corpus for the Chinese DRR task. We use machine translation to generate Chinese instances from a labeled English discourse corpus. In this way, each instance has two independent views: Chinese and English views. Then we train two classifiers in Chinese and English in a co-training way, which exploits unlabeled Chinese data to implement better implicit DRR for Chinese. Experimental results demonstrate the effectiveness of our method.

Key words: Cross-lingual; Implicit discourse relation recognition; Co-training
https://doi.org/10.1631/FITEE.1601865 **CLC number:** TP391.1

1 Introduction

As a crucial task for discourse analysis, discourse relation recognition (DRR) aims to identify automatically the internal structures and logical relations of coherent texts. According to whether an explicit connective exists between a pair of textual spans or not, discourse relations can be divided into explicit and implicit ones. The example illustrates a discourse relation instance (https://catalog.ldc.upenn.edu/LDC2008T05): "Three of its 17 Bay-area branches were closed yesterday; however, the company expects all branches to reopen today." With the presence of the discourse connective 'however', these two sentences display an explicit discourse relation comparison which can be inferred easily. Once this discourse connective is removed, the discourse relation becomes implicit.

DRR provides important clues to many other natural language processing (NLP) tasks, such as question answering (Verberne et al., 2007), information extraction (Cimiano et al., 2005), and machine translation (Guzmán et al., 2014). Despite the great progress in explicit DRR where discourse connectives (e.g., 'because' and 'but') are given explicitly in a text (Miltsakaki et al., 2005; Pitler and Nenkova, 2009), implicit DRR remains a great challenge. Previous works (Lin et al., 2009; Pitler et al., 2009; Louis et al., 2010; Wang et al., 2010; Biran and McKeown, 2013; Rutherford and Xue, 2014)

‡ Corresponding author

* Project supported by the National Natural Science Foundation of China (No. 61672440), the Natural Science Foundation of Fujian Province, China (No. 2016J05161), the Research Fund of the State Key Laboratory for Novel Software Technology in Nanjing University, China (No. KFKT2015B11), the Scientific Research Project of the National Language Committee of China (No. YB135-49), and the Fundamental Research Funds for the Central Universities, China (No. ZK1024)

ORCID: Yao-jie LU, http://orcid.org/0000-0002-5842-7715

studied most methods which usually rely heavily on labeled corpora. However, such resources in different languages are very unbalanced. For example, existing resources were developed mainly for English research while Chinese labeled discourse corpora are rare. Therefore, implicit DRR for Chinese is still a difficult task.

Inspired by the success of cross-lingual NLP tasks (Wan, 2009; Qian et al., 2014), we investigate the problem of cross-lingual implicit DRR in this study, using labeled English and unlabeled Chinese corpora for Chinese implicit DRR. Specifically, we propose an adapted co-training approach to improve the performance of implicit DRR for Chinese. Following Zhou et al. (2012) and Laali and Kosseim (2014), we assume that discourse relations are retained after translation and then we apply machine translation services to translate English training instances into Chinese and translate Chinese test instances and additional unlabeled instances into English. In this way, we can view each instance from two independent perspectives: a Chinese view with only Chinese features and an English view with only English features. Then we employ the proposed co-training approach to exploit the two redundant views of features. Compared with the previous works which exploited only monolingual views, our approach further uses training data in another language for model training. Hence, it has the potential to achieve better performance. We simulate a cross-lingual implicit DRR experiment environment in Section 4, where we train a classifier for Chinese implicit DRR using the English labeled data and Chinese unlabeled data. Experimental results show that our approach can outperform inductive and transductive classifiers.

2 Discourse relation definitions and corpora

2.1 English discourse relation recognition

Various labeled English corpora have been developed based on different discourse theories, such as the rhetorical structure theory discourse treebank (Carlson et al., 2001) and the Penn discourse treebank (PDTB) (Prasad et al., 2008). In this study, we focus on the latter, which is the largest corpus with manually labeled discourse relation labels so far. The PDTB contains discourse annotations over 2312 *Wall Street Journal* articles with 40 600 instances. For every instance, the discourse relation is labeled between two arguments and can be further classified as either an explicit or an implicit relation depending upon the presence or absence of connective words. With regard to DRR, discourse relation labels are organized hierarchically at three levels. Existing works focused mainly on the recognition of four top-level relations: comparison, contingency, expansion, and temporal.

2.2 Chinese discourse relation recognition

Compared with English resources for DRR, Chinese labeled discourse corpora are rare. In this study, we use two well-known labeled corpora for Chinese implicit DRR. One is Harbin Institute of Technology Chinese Discourse Treebank (HIT-CDTB) (Zhang et al., 2014), containing 525 articles from four domains (broad news, magazine, newwire, and web). In this corpus, each instance is labeled with one of the six kinds of relations: temporal, causality, contingency, comparison, expansion, and coordination. The other is Soochow University Chinese Discourse Treebank (Soochow-CDTB) (Li Y et al., 2014), consisting of 500 news articles, where each instance is classified as one of the following four kinds of relations: causality, coordination, transition, and explanation.

Note that PDTB, HIT-CDTB, and Soochow-CDTB have different discourse relation definitions. Following Xue (2005), Huang and Chen (2011), and Zhou et al. (2012), we adopt directly the lexically ground approach of the PDTB while making systematic adaptations motivated by the characteristics of Chinese texts. Specifically, we primarily study how to recognize Chinese texts with PDTB-style relations including comparison (Comp), contingency (Cont), expansion (Expa), and temporal (Temp), leaving the recognition of other relations as our future work. As implemented in Pitler et al. (2009), we run four binary classification tasks to identify each of the main relations from the rest. To exploit the instances of Soochow-CDTB, we analyze the discourse relations defined in different corpora, and then map the discourse relations of Soochow-CDTB into PDTB-style ones. Table 1 shows the details of discourse relation mapping. In particular, for some instances that may correspond to

different PDTB-style relations (such as the continuing subtype relation), we manually classify them into different PDTB-style relations.

3 Our approach

In this study, we devote ourselves to exploiting the labeled English corpora and unlabeled Chinese corpora for Chinese implicit DRR in a semi-supervised framework.

Given labeled English instances and unlabeled Chinese instances, there are two direct approaches to perform cross-lingual implicit DRR. The first is to learn an English classifier on labeled English corpora and use the classifier to recognize the implicit discourse relations of the unlabeled Chinese instances' English translation. The second is to translate labeled English instances into Chinese ones and train a Chinese classifier, which can be used directly to classify unlabeled Chinese instances. However, the original and translated instances are different in underlying distributions, so we think that the above two methods may not perform well.

To settle the above problem, we propose a co-training approach for the task. The co-training algorithm (Blum and Mitchell, 1998) is a typical bootstrapping method, starting with a set of labeled data and training by increasing the amount of labeled data using some amounts of unlabeled data in an incremental way. In practical use, two views which are not identical are required for co-training to work. So far, the co-training algorithm has been applied successfully in many NLP tasks, such as parsing (Sarkar, 2001), conference resolution (Ng and Cardie, 2003), part-of-speech (POS) tagging (Clark et al., 2003), and cross-lingual sentiment analysis (Wan, 2009). Inspired by Wan (2009), we apply the conventional co-training algorithm to implement cross-lingual implicit DRR, which exploits the English and Chinese features in a unified framework.

Fig. 1 shows the framework of the proposed approach, which involves the following three steps:

1. Using machine translation services, we translate two arguments of labeled English instances into labeled Chinese ones and two arguments of unlabeled Chinese instances into unlabeled English ones. For this process, we use the public state-of-the-art commercial machine translation systems, such as Baidu Translate, for both English-to-Chinese translation and Chinese-to-English translation. After translation, we obtain an English and a Chinese version for each instance. In this way, the English and Chi-

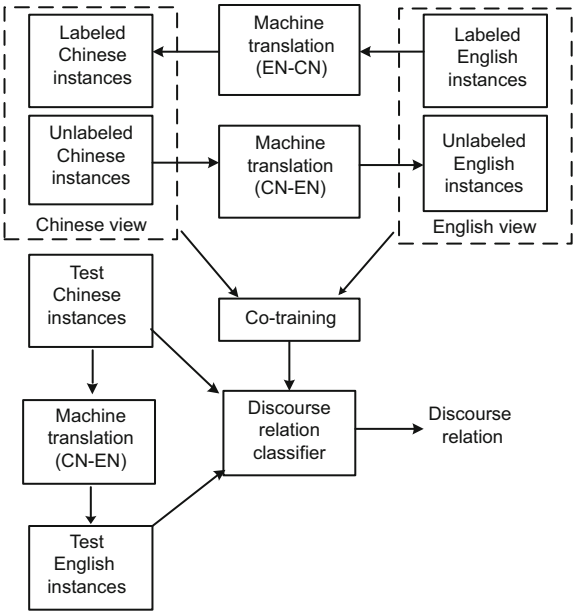

Fig. 1 Framework of co-training for cross-lingual implicit discourse relation recognition

Table 1 Discourse relation mapping details

Soochow-CDTB		PDTB-style relation
Top-level relation	Second-level relation	
Causality	Gause-result, inference, hypothesis	Contingency
	purpose, condition	
	Background	Comparison/contingency/temporal
Coordination	Coordination, progressive, selection	Expansion
	Continue	Temporal/expansion
Transition	Inverse	Comparison
	Transition, concessive	
Explanation	Explanation, summary-elaboration	Expansion
	Example, evaluation	

nese features for instances can be considered as two independent and redundant views for them.

2. We adopt the co-training algorithm to our scenario and train two classifiers in English and Chinese, separately. The algorithm adopted is shown in Algorithm 1. The basic intuition behind our approach is that if one classifier can predict confidently the class of an example, it can provide one more training example for the classifier in the other language. L_0 denotes the English corpus and U_0 denotes the Chinese corpus. Note that both English and Chinese instances have two versions.

Algorithm 1 Cross-lingual implicit DRR based on co-training

Input:

Iteration number N_{iter};
Feature set in English view F_{en};
Feature set in Chinese view F_{cn};
Initial set of labeled data L_0;
Initial set of unlabeled data U_0;
Number of positive samples in each iteration N_{p};
Number of negative samples in each iteration N_{n};
The most confident instance set in English view E_{en};
The most confident instance set in Chinese view E_{cn};

Procedure:

$L = L_0, U = U_0$
for $i = [1, N_{\text{iter}}]$ **do**
  $E_{\text{en}} = \varnothing$, $E_{\text{cn}} = \varnothing$;
  Train an English classifier C_{en} with L based on F_{en};
  Use C_{en} to label instances from U based on F_{en};
  Add the most confidently predicted N_{p} positive instances and N_{n} negative instances from U to E_{en};
  Train a Chinese classifier C_{cn} with L based on F_{cn};
  Use C_{cn} to label instances from U based on F_{cn};
  Add the most confidently predicted N_{p} positive instances and N_{n} negative instances from U to E_{cn};
  Remove instances whose labels are conflicting with E_{cn} and E_{en};
  Update unlabeled instance set $U{=}U{-}(E_{\text{en}} \cup E_{\text{cn}})$;
  Update labeled instance set $L{=}L{+}(E_{\text{en}} \cup E_{\text{cn}})$;
end for

To begin our work, we use the labeled English instances and the Chinese translations to train two classifiers in two languages. Here, we take the widely used support vector machine (SVM) as our basic classifier because of its good performance in many NLP tasks. Using these two classifiers, we then classify the unlabeled Chinese instances and the corresponding English translations. Next, we choose the most confidently predicted instances, which had consistently predicted labels produced by the two classifiers. The selected Chinese instances and the English translations are used later to retrain the classifiers with primary training instances. The number of selected positive instances N_{p} and the number of selected negative instances N_{n} are optimized over the validation set. This procedure is repeated until the maximum number of iterations is reached.

3. We simultaneously apply the above mentioned two classifiers to implement implicit DRR for Chinese. For each Chinese instance, we translate it into an English instance, and then use the above two classifiers to predict labels for the Chinese and translate English instances, respectively. Then we obtain two prediction values for each instance, both of which can be normalized into $[-1, 1]$ by dividing the maximum absolute value. We then use the average of the normalized values as the final prediction value for the instance.

4 Experiments

We conducted experiments on a Chinese implicit DRR task to validate the effectiveness of our approach.

4.1 Experimental setup

In all experiments, we used SVM-light (http://svmlight.joachims.org) to train our SVM classifiers and Baidu Translate (http://fanyi.baidu.com) to obtain translations. We adopted the Stanford NLP toolkit (http://nlp.stanford.edu) to preprocess (i.e., word segmenting and parsing) English and Chinese instances, respectively.

4.1.1 Dataset

To investigate the cross-lingual transfer ability of the proposed approach, we used an English dataset as the training set, and a Chinese dataset as the val/test/unlabeled set. Considering the fact that the instances of different labels are distributed unevenly

in the training corpora, we trained classifiers using the same number of positive and negative instances, as implemented by Pitler et al. (2009). However, the relation distribution in PDTB is very imbalanced. The number of implicit instances with expansion relation in the training set is greater than those with all other relations. To build a balanced training set, we sampled negative instances from implicit and explicit instances with other relations. Table 2 shows the instance numbers of datasets.

1. English dataset

Following Lan et al. (2013) and Rutherford and Xue (2015), we used the implicit instances of sections 2–20 that contain 12 632 instances as English labeled training data in the PDTB 2.0 corpus. For those instances that had been labeled with multiple types, we used them as multiple training instances during training.

2. Chinese dataset

Chinese data come from HIT-CDTB 1.0 (Zhang et al., 2014) and Soochow-CDTB 1.0 (Li Y et al., 2014). To avoid domain difference, we chose only the implicit instances of HIT-CDTB belonging to broad news and newwire in our experiments. For Soochow-CDTB, we mapped the discourse relations of its implicit instances into PDTB-style ones, as mentioned in Section 2. In total, we obtained 19 074 implicit instances from the two corpora. Then we split these instances into three parts in the units of articles, unlabeled, validation, and test sets, in the proportion of 8:1:1.

4.1.2 Features

We used the following common features from both English and Chinese views: cross-argument word pairs, first last first3, polarity (Pitler et al., 2009), and production rules (Lin et al., 2009).

1. Cross-argument word pairs

We grouped all words from the first and second arguments (Arg1 and Arg2) into two sets W_1 and W_2, respectively, and then extracted any possible word pair $(w_i; w_j)$ $(w_i \in W_1; w_2 \in W_2)$ as features.

2. First last first3

We took the first and last words of each argument, the pair of the first words, the pair of the last words, and the first three words of each argument as features.

3. Polarity

We collected the count of positive, negated positive, negative, and neutral words in Arg1 and Arg2 according to the MPQA corpus (http://mpqa.cs.pitt.edu) (English) and HowNet database (http://www.keenage.com) (Chinese). Their cross products were used as features.

4. Production rules

We extracted all production rules from syntactic trees of the arguments; each rule was represented by three binary features to check whether this rule appeared in Arg1, Arg2, or both arguments.

Particularly, we used a cutoff frequency which is tuned on the validation set to remove infrequent features.

4.1.3 Baselines

We compared our approach against the following methods:

SVM (CN): We used the inductive SVM with only Chinese features for implicit DRR in the Chinese view. Only English-to-Chinese translations were required. Unlabeled data were not used in this case.

SVM (EN): The inductive SVM with only English features was used for implicit DRR in the English view. Only Chinese-to-English translations were required. We did not use unlabeled data either.

SVM (ENCN1): We employed the inductive SVM with both English and Chinese features for implicit DRR in two views. Both English-to-Chinese and Chinese-to-English translations were used. Unlabeled data were not used.

SVM (ENCN2): We combined the results of SVM (EN) and SVM (CN) by averaging prediction values in the same way as was done with the co-training approach.

TSVM (CN): We used the transudative SVM with only Chinese features for implicit DRR in the Chinese view. Only English-to-Chinese translations were needed and unlabeled data were used.

TSVM (EN): We applied the transudative SVM with only English features for implicit DRR in the English view. Only Chinese-to-English translations were needed and unlabeled data were used.

TSVM (ENCN1): We applied the transudative SVM with both English and Chinese features for implicit DRR in two views. Both English-to-Chinese and Chinese-to-English translations were required and unlabeled data were used.

TSVM (ENCN2): We combined the results of TSVM (EN) and TSVM (CN) by averaging their prediction values.

Self-training SVM (EN): This was a basic self-training method with only English features for implicit DRR in the English view. Data used in this method were the same as those in TSVM (EN).

Self-training SVM (CN): This was a basic self-training method with only English features for implicit DRR in the Chinese view. Data used in this method were the same as those in TSVM (CN).

The data used in baseline systems and the co-training method were described in Table 3.

4.1.4 Measurements

Because the validation and test sets are imbalanced, we chose the F_1 score and accuracy as our major evaluation metrics. Formally, we calculated the accuracy and F_1 score as follows:

$$\mathrm{Acc} = \frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN}}, \tag{1}$$

$$P = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}}, \tag{2}$$

$$R = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}, \tag{3}$$

$$F_1 = \frac{2P \cdot R}{P+R}, \tag{4}$$

where TP (FP) denotes the number of instances that are identified correctly (incorrectly) as positive ones, and TN (FN) is the number of instances which are classified correctly (incorrectly) as negative ones.

4.2 Hyperparameter optimization

Following Ji and Eisenstein (2015) and Liu et al. (2016), we trained classifiers on the training set and tuned parameters (such as the cut-off frequency and iteration number) on the validation set, and evaluated the algorithm's performance on the test set. We enumerated all combinations of N_p and N_n and tuned the optional hyperparameters according to the F_1 score on the validation set. Specifically, we set the maximum iteration number as 200, drew N_p or N_n from 1 to 20 with step 2, and set the other to 1, such as (1, 3) and (3, 1). The tuned hyperparameters are shown in Table 4. We can observe that the optimal N_p and N_n of co-training acting on expansion instances are different from those for the instances with other labels. The reason is that the number of positive instances in expansion is larger than that of negative ones, which differs from other relations.

Table 2 Instance numbers of datasets in our experiments

Discourse relation	Number of labeled data			Unlabeled data (CN)
	Train positive/negative (EN)	Val positive/negative (CN)	Test positive/negative (CN)	
Comp	1942/1942	15/944	24/1172	10 691
Cont	3342/3342	146/813	175/1021	
Expa	7004/7004	772/187	772/187	
Temp	760/760	26/933	41/1155	

CN: Chinese; EN: English

Table 3 Data used in baselines and the co-training method

Method	Training set		Val/Test set		Unlabeled set	
	English	E-to-C	Chinese	C-to-E	Chinese	C-to-E
SVM (EN)	√			√		
SVM (CN)		√	√			
SVM (ENCN1)	√	√	√	√		
SVM (ENCN2)	√	√	√	√		
TSVM (EN)	√			√		√
TSVM (CN)		√	√		√	
TSVM (ENCN1)	√	√	√	√	√	√
TSVM (ENCN2)	√	√	√	√	√	√
Self-training SVM (EN)	√			√		√
Self-training SVM (CN)		√	√		√	
Co-training	√	√	√	√	√	√

E-to-C: Chinese translations of English instances; C-to-E: English translations of Chinese instances

Similarly, we tuned the optimal parameters N_{p}, N_{n}, and N_{iter} of the self-training algorithm and num_{+} (which means the fraction of unlabeled examples to be classified into the positive class) in TSVM on the validation set.

Table 4 Optimal hyperparameters

Hyperparameter	Temp	Cont	Comp	Expa
N_{p}^{*}	1	1	1	9
N_{n}^{*}	9	7	9	1
N_{iter}^{*}	48	44	14	97

4.3 Overall performance

Table 5 shows the experimental results. The proposed method outperforms all the comparative systems on temporal, comparison, and expansion, and is comparable with the best comparative system on contingency.

Among the baselines, TSVMs do not always improve performance using the unlabeled data compared to the inductive SVMs. We conjecture the main reason for this result is that transductive classifiers aim at optimizing the distribution of all the unlabeled data in each iteration. However, the extreme imbalanced distributions of different discourse relations result in difficulty in training transductive classifiers. In contrast, the proposed approach and self-training based methods augment iteratively the training data with a small quantity of the most confidently predicted instances each time. Classifier performance is improved steadily after several iterations.

In this study, self-training based methods simply add pseudo labeled instances based on monolingual features. In contrast, the instances selected by the proposed approach are highly confident with the same labels from two views. The co-training method takes into account the English and Chinese bilingual feature sets simultaneously, which enables them to promote and restrict each other. Therefore, compared with self-training based methods, our approach simultaneously exploits different language views. This is the main reason why the co-training approach is superior to self-training based methods.

4.4 Effect of the iteration number

In this group of experiments, we investigated the effect of the iteration number on our approach. We tried different iteration numbers from 1 to 200 with the optimal N_{p} and N_{n} to train different models.

From Fig. 2, we can see that the co-training approach outperforms self-training based methods for the first several iterations and the best performance of the co-training approach is better than that of self-training based methods. The reason is that co-training exploits only the instances with the same predicted labels in two languages; thus, the noisy instances even with high confidence will be removed. In other words, the co-training method is able to identify more correct instances effectively.

However, the performance of co-training on four relations is not improved all the time, although the best performance is achieved after several iterations. This is mainly because the instances with temporal, contingency, and comparison labels are rare in the Chinese corpus. As for expansion recognition, we can observe that the proposed approach achieves the best performance after a large number of iterations. We speculate that what underlies is that unlabeled data are dominated by instances with expansion labels.

5 Related work

Recently, implicit DRR has become a hot research spot in NLP. Most studies were limited to English implicit DRR. Research emphasis of earlier work was on how to collect training data using pattern-based approaches (Marcu and Echihabi, 2002), which may be problematic for real implicit relations (Sporleder and Lascarides, 2008). Then the release of PDTB provided a large-scale high-quality resource for DRR in English, leading to a number of studies implementing implicit DRR via supervised classifiers. For example, Pitler et al. (2009) exploited several linguistical informed features. Along this line, more powerful features have been exploited: contexts, word pairs, discourse parse information (Lin et al., 2009; Wang et al., 2010), entities (Louis et al., 2010), event pairs (Chiarcos, 2012), Brown cluster pairs, co-reference patterns (Rutherford and Xue, 2014), ect. In different ways, Zhou et al. (2010) used a language model to generate automatically implicit connectives which are used for the recognition of implicit relations, while Park and Cardie (2012) performed feature set optimization for better feature combination. All above approaches were based on labeled data, and therefore

Table 5 Performance (*F*-score/accuracy) comparison on the test set

Method	F_1 (Acc)			
	Temp	Cont	Comp	Expa
SVM (EN)	7.46 (78.11)	19.51 (79.61)	10.45 (89.41)	68.26 (55.34)
SVM (CN)	7.12 (67.78)	19.38 (79.44)	6.14 (75.73)	66.84 (54.28)
SVM (ENCN1)	7.69 (70.34)	21.79 (82.26)	10.81 (88.35)	69.67 (57.28)
SVM (ENCN2)	9.18 (75.55)	20.48 (79.44)	10.53 (86.50)	63.41 (51.10)
TSVM (EN)	5.77 (91.35)	**23.17 (42.63)**	10.13 (93.73)	81.20 (69.55)
TSVM (CN)	5.77 (91.35)	20.99 (29.57)	4.17 (83.76)	82.40 (70.96)
TSVM (ENCN1)	8.47 (90.47)	21.29 (30.19)	9.64 (93.38)	82.37 (70.79)
TSVM (ENCN2)	8.47 (90.47)	21.29 (30.19)	9.64 (93.38)	82.37 (70.79)
Self-training SVM (EN)	8.42 (84.64)	21.79 (60.72)	10.45 (89.41)	83.87 (72.90)
Self-training SVM (CN)	8.21 (84.20)	22.00 (52.43)	9.62 (91.70)	87.24 (78.02)
Co-training	**12.50 (93.82)**	22.86 (83.32)	**14.00 (92.41)**	**88.61 (79.96)**

The bold numbers are the best results in each relation

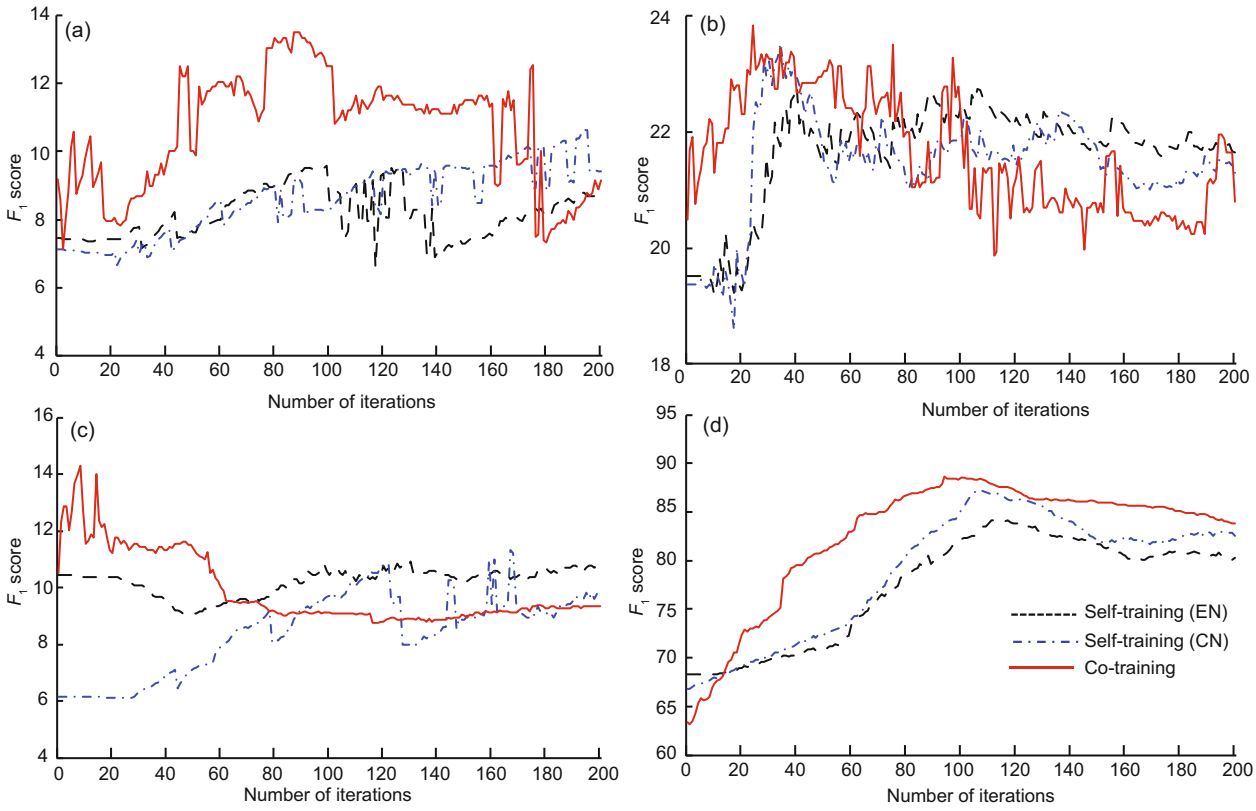

Fig. 2 Illustration of co-training and self-training on the test set with different iteration numbers: (a) temporal; (b) contingency; (c) comparison; (d) expansion

they were likely to suffer from data sparsity. To deal with this problem, some researchers adopted semi-supervised methods for DRR (Hernault et al., 2010), focused on the exploitation of synthetic data or explicit instances (Wang et al., 2012; Lan et al., 2013). Recently, neural networks have been used for implicit DRR due to their capabilities of representation learning and exploiting unlabeled data (Braud and Denis, 2015; Ji and Eisenstein, 2015; Zhang et al., 2015; Chen et al., 2016; Ji et al., 2016; Liu et al., 2016; Qin et al., 2016; Rutherford et al., 2016; Zhang et al., 2016).

Compared with the studies on English implicit DRR, previous studies on Chinese implicit DRR were quite limited. As for resource development, Chen (2006) and Ming (2008) used rhetorical struc-

ture theory (RST) to annotate Chinese discourse. Instead, Zhou et al. (2012) and Zhou and Xue (2015) used PDTB annotation guidelines to annotate Chinese discourse. Lately, HIT-CDTB (Zhang et al., 2014) and Soochow-CDTB (Li Y et al., 2014) emerged, both of which were developed according to the characteristics of Chinese. However, these corpora are small in size, and therefore the conventional discriminative classifiers did not work well (Zhang et al., 2013). To overcome this problem, Li J et al. (2014) projected English annotations available to Chinese via parallel corpora. This work was somewhat similar to Zhou et al. (2012), which preformed cross-lingual identification of ambiguous discourse connectives for Chinese. Besides, Rutherford et al. (2016) probed systematically the effectiveness of various neural networks for Chinese implicit DRR.

Compared with the above-mentioned studies, our work is significantly different because we focus mainly on leveraging English labeled data and Chinese unlabeled data via machine translation and co-training for cross-lingual implicit DRR. To the best of our knowledge, this has not been investigated before in this task.

6 Conclusions

In this paper, we have proposed a co-training approach to improve the accuracy of Chinese implicit DRR by exploiting English labeled data and Chinese unlabeled data. In our approach, each instance could be considered by two views of different languages, and therefore various syntactic features in different languages could be used for Chinese implicit DRR. The experimental results demonstrated the effectiveness of our approach.

In the future, we plan to investigate the effectiveness of our method for other languages and fine-grained subtype relations in PDTB. Additionally, inspired by cross-lingual NLP based on deep learning (Jain and Batra, 2015), we will attempt to use neural network architecture to perform cross-lingual implicit DRR.

References

Biran O, McKeown K, 2013. Aggregated word pair features for implicit discourse relation disambiguation. Proc 51st Annual Meeting of the Association for Computational Linguistics, p.69-73. https://doi.org/10.7916/D8PN9FZ4

Blum A, Mitchell T, 1998. Combining labeled and unlabeled data with cotraining. Proc 11th Annual Conf on Computational Learning Theory, p.92-100. https://doi.org/10.1145/279943.279962

Braud C, Denis P, 2015. Comparing word representations for implicit discourse relation classification. Proc Conf on Empirical Methods in Natural Language Processing, p.2201-2211. https://doi.org/10.18653/v1/d15-1262

Carlson L, Marcu D, Okurowski M, 2001. Building a discourse-tagged corpus in the framework of rhetorical structure theory. Proc 2nd SIGDIAL Workshop on Discourse and Dialogue, p.1-10. https://doi.org/10.3115/1118078.1118083

Chen J, Zhang Q, Liu P, et al., 2016. Implicit discourse relation detection via a deep architecture with gated relevance network. Proc 54th Annual Meeting of the Association for Computational Linguistics, p.1726-1735. https://doi.org/10.18653/v1/p16-1163

Chen L, 2006. English and Chinese Discourse Structure Dimension Theory and Practice. PhD Thesis, Shanghai International Studies University, China.

Chiarcos C, 2012. Towards the unsupervised acquisition of discourse relations. Proc 50th Annual Meeting of the Association for Computational Linguistics, p.213-217.

Cimiano P, Reyle U, Šarić J, 2005. Ontology-driven discourse analysis for information extraction. *Data Knowl Eng*, 55(1):59-83. https://doi.org/10.1016/j.datak.2004.11.009

Clark S, Curran J, Osborne M, 2003. Bootstrapping POS-taggers using unlabelled data. Proc 7th Conf on Natural Language Learning, p.49-55. https://doi.org/10.3115/1119176.1119183

Guzmán F, Joty S, Màrquez L, et al., 2014. Using discourse structure improves machine translation evaluation. Proc 52nd Annual Meeting of the Association for Computational Linguistics, p.687-698. https://doi.org/10.3115/v1/p14-1065

Hernault H, Bollegala D, Ishizuka M, 2010. Towards semi-supervised classification of discourse relations using feature correlations. Proc SIGDIAL Conf and the 11th Annual Meeting of the Special Interest Group on Discourse and Dialogue, p.55-58.

Huang H, Chen H, 2011. Chinese discourse relation recognition. Proc 5th Int Joint Conf on Natural Language Processing, p.1442-1446.

Jain S, Batra S, 2015. Cross lingual sentiment analysis using modified BRAE. Proc Conf on Empirical Methods in Natural Language Processing, p.159-168. https://doi.org/10.18653/v1/d15-1016

Ji Y, Eisenstein J, 2015. One vector is not enough: entity-augmented distributed semantics for discourse relations. *Trans Assoc Comput Ling*, 3:329-344.

Ji Y, Haffari G, Eisenstein J, 2016. A latent variable recurrent neural network for discourse relation language models. Proc Conf North American Chapter of the Association for Computational Linguistics on Human Language Technologies, p.332-342. https://doi.org/10.18653/v1/n16-1037

Laali M, Kosseim L, 2014. Inducing discourse connectives from parallel texts. Proc 25th Int Conf on Computational Linguistics, p.610-619.

Lan M, Xu Y, Niu Z, 2013. Leveraging synthetic discourse data via multi-task learning for implicit discourse relation recognition. Proc 51st Annual Meeting of the Association for Computational Linguistics, p.476-485.

Li J, Carpuat M, Nenkova A, 2014. Cross-lingual discourse relation analysis: a corpus study and a semi-supervised classification system. Proc 25th Int Conf on Computational Linguistics, p.577-587.

Li Y, Feng W, Sun J, et al., 2014. Building Chinese discourse corpus with connective-driven dependency tree structure. Proc Conf on Empirical Methods in Natural Language Processing, p.2105-2114. https://doi.org/10.3115/v1/d14-1224

Lin Z, Kan M, Ng H, 2009. Recognizing implicit discourse relations in the Penn discourse treebank. Proc Conf on Empirical Methods in Natural Language Processing, p.343-351. https://doi.org/10.3115/1699510.1699555

Liu Y, Li S, Zhang X, et al., 2016. Implicit discourse relation classification via multi-task neural networks. Proc 30th Conf on Artificial Intelligence, p.2750-2756.

Louis A, Joshi A, Prasad R, et al., 2010. Using entity features to classify implicit discourse relations. Proc 11th Annual Meeting of the Special Interest Group on Discourse and Dialogue, p.59-62.

Marcu D, Echihabi A, 2002. An unsupervised approach to recognizing discourse relations. Proc 40th Annual Meeting of the Association for Computational Linguistics, p.368-375. https://doi.org/10.3115/1073083.1073145

Miltsakaki E, Dinesh N, Prasad R, et al., 2005. Experiments on sense annotations and sense disambiguation of discourse connectives. Proc 4th Workshop on Treebanks and Linguistic Theories, p.1-13.

Ming Y, 2008. Rhetorical structure annotation of Chinese news commentaries. *J Chin Inform Proc*, 22(4):19-23.

Ng V, Cardie C, 2003. Weakly supervised natural language learning without redundant views. Proc Conf North American Chapter of the Association for Computational Linguistics on Human Language Technology, p.94-101. https://doi.org/10.3115/1073445.1073468

Park J, Cardie C, 2012. Improving implicit discourse relation recognition through feature set optimization. Proc 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue, p.108-112.

Pitler E, Nenkova A, 2009. Using syntax to disambiguate explicit discourse connectives in text. Proc ACL-IJCNLP Conf, p.13-16. https://doi.org/10.3115/1667583.1667589

Pitler E, Louis A, Nenkova A, 2009. Automatic sense prediction for implicit discourse relations in text. Proc of the Joint Conf 47th Annual Meeting of the ACL and the 4th Int Joint Conf on Natural Language Processing of the AFNLP, p.683-691. https://doi.org/10.3115/1690219.1690241

Prasad R, Dinesh N, Lee A, et al., 2008. The Penn discourse treebank 2.0. Proc Int Conf on Language Resources and Evaluation, p.2961-2968.

Qian L, Hui H, Hu Y, et al., 2014. Bilingual active learning for relation classification via pseudo parallel corpora. Proc 52nd Annual Meeting of the Association for Computational Linguistics, p.582-592. https://doi.org/10.3115/v1/p14-1055

Qin L, Zhang Z, Zhao H, 2016. A stacking gated neural architecture for implicit discourse relation classification. Proc Conf on Empirical Methods in Natural Language Processing, p.2263-2270. https://doi.org/10.18653/v1/d16-1246

Rutherford A, Xue N, 2014. Discovering implicit discourse relations through brown cluster pair representation and coreference patterns. Proc 14th Conf European Chapter of the Association for Computational Linguistics, p.645-654. https://doi.org/10.3115/v1/e14-1068

Rutherford A, Xue N, 2015. Improving the inference of implicit discourse relations via classifying explicit discourse connectives. Proc Conf of the North American Chapter of the Association for Computational Linguistics on Human Language Technologies, p.799-808. https://doi.org/10.3115/v1/n15-1081

Rutherford A, Demberg V, Xue N, 2016. Neural network models for implicit discourse relation classification in English and Chinese without surface features. http://arxiv.org/abs/1606.01990

Sarkar A, 2001. Applying co-training methods to statistical parsing. Proc 2nd Meeting of the North American Chapter of the Association for Computational Linguistics on Language Technologies, p.1-8. https://doi.org/10.3115/1073336.1073359

Sporleder C, Lascarides A, 2008. Using automatically labelled examples to classify rhetorical relations: an assessment. *Nat Lang Eng*, 14:369-416. https://doi.org/10.1017/S1351324906004451

Verberne S, Boves L, Oostdijk N, et al., 2007. Evaluating discourse-based answer extraction for why-question answering. Proc 30th Annual Int ACM SIGIR Conf on Research and Development in Information Retrieval, p.735-736. https://doi.org/10.1145/1277741.1277883

Wan X, 2009. Co-training for cross-lingual sentiment classification. Proc of the Joint Conf 47th Annual Meeting of the ACL and the 4th Int Joint Conf on Natural Language Processing of the AFNLP, p.235-243.

Wang W, Su J, Tan C, 2010. Kernel based discourse relation recognition with temporal ordering information. Proc 48th Annual Meeting of the Association for Computational Linguistics, p.710-719.

Wang X, Li S, Li J, et al., 2012. Implicit discourse relation recognition by selecting typical training examples. Proc of COLING, p.2757-2772.

Xue N, 2005. Annotating discourse connectives in the Chinese treebank. Proc Workshop on Frontiers in Corpus Annotations II: Pie in the Sky, p.84-91. https://doi.org/10.3115/1608829.1608841

Zhang B, Su J, Xiong D, et al., 2015. Shallow convolutional neural network for implicit discourse relation recognition. Proc Conf on Empirical Methods in Natural Language Processing, p.2230-2235. https://doi.org/10.18653/v1/d15-1266

Zhang B, Xiong D, Su J, et al., 2016. Variational neural discourse relation recognizer. Proc Conf on Empirical Methods in Natural Language Processing, p.382-391. https://doi.org/10.18653/v1/d16-1037

Zhang M, Song Y, Qin B, et al., 2013. Chinese discourse relation recognition. *J Chin Inform Proc*, 27(6):51-57.

Zhang M, Qin B, Liu T, 2014. Chinese discourse relation hierarchy and annotation. *J Chin Inform Proc*, 28(2):28-36.

Zhou L, Gao W, Li B, et al., 2012. Cross-lingual identification of ambiguous discourse connectives for resource poor language. Proc COLING, p.1409-1418.

Zhou Y, Xue N, 2015. The Chinese discourse treebank: a Chinese corpus annotated with discourse relations. *Lang Res Eval*, 49(2):397-431. https://doi.org/10.1007/s10579-014-9290-3

Zhou Z, Xu Y, Niu Z, et al., 2010. Predicting discourse connectives for implicit discourse relation recognition. 23rd Int Conf on Computational Linguistics, p.1507-1514.