# Implicit Discourse Relation Recognition for English and Chinese with Multiview Modeling and Effective Representation Learning

HAORAN LI and JIAJUN ZHANG, National Laboratory of Pattern Recognition, Institute of Automation, University of Chinese Academy of Sciences, Chinese Academy of Sciences
CHENGQING ZONG, National Laboratory of Pattern Recognition, Institute of Automation, University of Chinese Academy of Sciences, CAS Center for Excellence in Brain Science and Intelligence Technology, Chinese Academy of Sciences

Discourse relations between two text segments play an important role in many Natural Language Processing (NLP) tasks. The connectives strongly indicate the sense of discourse relations, while in fact, there are no connectives in a large proportion of discourse relations, that is, implicit discourse relations. Compared with explicit relations, implicit relations are much harder to detect and have drawn significant attention. Until now, there have been many studies focusing on English implicit discourse relations, and few studies address implicit relation recognition in Chinese even though the implicit discourse relations in Chinese are more common than those in English. In our work, both the English and Chinese languages are our focus. The key to implicit relation prediction is to properly model the semantics of the two discourse arguments, as well as the contextual interaction between them. To achieve this goal, we propose a neural network based framework that consists of two hierarchies. The first one is the model hierarchy, in which we propose a max-margin learning method to explore the implicit discourse relation from multiple views. The second one is the feature hierarchy, in which we learn multilevel distributed representations from words, arguments, and syntactic structures to sentences. We have conducted experiments on the standard benchmarks of English and Chinese, and the results show that compared with several methods our proposed method can achieve the best performance in most cases.



CCS Concepts: • **Computing methodologies → Discourse, dialogue and pragmatics;**

Additional Key Words and Phrases: Implicit discourse relation, neural network, multilevel features, max-margin learning

**ACM Reference Format:**
Haoran Li, Jiajun Zhang, and Chengqing Zong. 2017. Implicit discourse relation recognition for English and Chinese with multiview modeling and effective representation learning. ACM Trans. Asian Low-Resour. Lang. Inf. Process. 16, 3, Article 19 (March 2017), 21 pages.
DOI: http://dx.doi.org/10.1145/3028772

## 1. INTRODUCTION

Automatic discourse relation inference is a pivotal task for discourse analysis and is beneficial to many Natural Language Processing (NLP) applications [Zong 2013], such

The research work has been funded by the Natural Science Foundation of China under Grants No. 61333018 and No. 91520204 and also supported by the Strategic Priority Research Program of the CAS under Grant No. XDB02070007.
Authors' addresses: H. Li, J. Zhang, and C. Zong (Corresponding author), Intelligence Building, No. 95, Zhongguancun East Road, Haidian District, Beijing, 100190, China; emails: {haoran.li, jjzhang, cqzong}@nlpr.ia.ac.cn.
Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies show this notice on the first page or initial screen of a display along with the full citation. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, to republish, to post on servers, to redistribute to lists, or to use any component of this work in other works requires prior specific permission and/or a fee. Permissions may be requested from Publications Dept., ACM, Inc., 2 Penn Plaza, Suite 701, New York, NY 10121-0701 USA, fax +1 (212) 869-0481, or permissions@acm.org.
© 2017 ACM 2375-4699/2017/03-ART19 $15.00
DOI: http://dx.doi.org/10.1145/3028772

as information extraction [Cimiano et al. 2005], sentiment analysis [Somasundaran et al. 2009], machine translation [Tu et al. 2013], and question answering [Jansen et al. 2014]. For instance, contrast relation recognition can eliminate intrasentence polarity ambiguities, and contingency relation detection can improve question answering systems and event relation extraction.

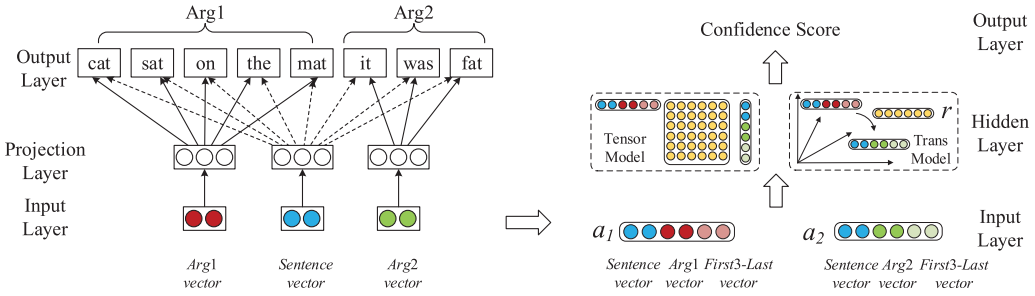
Depending on whether there are connectives, discourse relations can be categorized into explicit and implicit relations. Connectives play a crucial role in the inference of explicit discourse relations. For example, "but" is a strong indicator of a COMPARISON relation. However, implicit discourse relations are much more difficult to identify due to the absence of distinct relation cues, such as connectives. Therefore, this article focuses on implicit relation inference.

The goal of implicit relation inference is to recognize implicit discourse relations given two discourse arguments. Most of the research focuses on English implicit discourse relation recognition and regards this as a classification problem. Typically, many machine learning technologies such as Naïve Bayes, Maximum Entropy (ME) model, and Support Vector Machine (SVM) are used to perform this task. Recently, deep neural networks, such as the Recursive Neural Network (RNN) or Convolutional Neural Network (CNN), are employed to boost the recognition performance [Ji and Eisenstein 2015; Braud and Denis 2015; Zhang et al. 2015].

In addition to the model architecture, the features are also important. The prior work usually resorts to exploring diverse linguistically informed features, starting with lexical features introduced by Pitler et al. [2009], to syntactic features that have been proven to be more effective [Lin et al. 2009]. Subsequent work focuses on introducing more features [Louis et al. 2010; Biran and McKeown 2013] or selecting an optimal feature set [Park and Cardie 2012]. There are two limitations to these features. First, they strongly depend on the external linguistic resources, which are not generalizable to other languages. Second, these methods adopt discrete features, which lead to severe data sparsity, and cannot explore the similarities between discrete features. Recently, many studies demonstrate that the distributed feature representations can relieve the preceding problems and substantially improve implicit relation prediction [Ji and Eisenstein 2015; Braud and Denis 2015; Zhang et al. 2015]. These distributed representations usually focus on certain aspects, such as surface words or arguments, without modeling multilevel features, which should be more effective.

Compared with English, implicit discourse relations account for a much higher proportion in Chinese text. A total of 78% of the samples in the Chinese Discourse Treebank (CDTB) are annotated with implicit discourse relations [Kang et al. 2016], compared to 54% in the English Penn Discourse Treebank (PDTB). Therefore, inferring Chinese implicit discourse relations is more important. In our work, both the English and Chinese languages are our focus.

In this article, we tackle implicit discourse relation recognition for both English and Chinese, and propose a neural network based framework, which is shown in Figure 1. Our framework consists of two hierarchies. One is the model hierarchy, which is shown in Figure 1(b), and the other is the feature hierarchy, which is shown in Figure 1(a). The model is based on a max-margin neural network, which considers two views: the relation classification view and the relation transformation view. The feature hierarchy proposes to learn and apply distributed representations from different levels, namely, from words, arguments, and syntactic structures, to sentences. We also explore language-dependent features for English and Chinese (e.g., punctuations are used in Chinese implicit relation prediction). Extensive experiments show that our proposed method can significantly improve the quality of English and Chinese implicit relation recognition.




(a) Paragraph Vector Model (b) Discourse Relation Scoring Model

Fig. 1. The architecture of our model. To simplify the figure we neglect the word vectors of the input layer in the Paragraph Vector Model.

In summary, we make the following contributions in this article.

—We design a multiview based neural network model to recognize implicit discourse relations that takes into account the interactions between two discourse arguments and the relation transformation property.
—Instead of discrete features and distributed features of certain aspects, we propose learning and applying distributed feature representations in a multilevel manner, from words, arguments, and syntactic structures to sentences.
—The experiments show that we can obtain the best performance in most cases on the standard test sets in both English and Chinese.

## 2. DATASET

For English implicit discourse relation recognition task, we use PDTB 2.0 [Prasad et al. 2008], which is the largest available English discourse corpus, annotated with five main types of discourse relation labels: Explicit, Implicit, AltLex, EntRel, and NoRel. The types of discourse relations are organized into a hierarchical structure in which the first level contains four major classes: COMPARISON, CONTINGENCY, EXPANSION, and TEMPORAL. The next two levels consist of more fine-grained relation types. For example, "Comparison.Contrast.Juxtaposition" indicates that the second-level discourse relation is "Contrast" and the third-level is "Juxtaposition," Following Zhou et al. [2010], we do not regard the samples of EntRel relation as EXPANSION.

A discourse relation instance consists of two arguments: *Arg*1 and *Arg*2. For an implicit discourse relation argument pair, with implicit connectives manually inserted into the annotation as evidence for the sense of relation, *Arg*1 is to the left of the discourse connective and *Arg*2 is to the right. We give some examples of the implicit discourse relations in PDTB in Table I.

For Chinese implicit discourse relation recognition task, we use CDTB 0.5 [Zhou and Xue 2015], which consists of 98 files taken from the Chinese Treebank. Unlike the hierarchical structure of PDTB, there is only one level of relation in CDTB, including eight classes: Causation, Conditional, Conjunction, Contrast, Expansion, Progression, Purpose, and Temporal. In this work, we do not consider EntRel and NoRel. We give some examples of the implicit discourse relation in CDTB in Table II.

The distribution of implicit discourse relation instances in PDTB and CDTB will be introduced in Section 6.

Table I. Implicit Discourse Relation Examples in PDTB

| Discourse relation | Arg1 | Arg2 |
|---|---|---|
| Comparison.Contrast | A figure above 50 indicates the economy is likely to expand. | One below 50 indicates a contraction may be ahead. |
| Contingency.Cause | It's going to be a tough league. | There will be a lot of malice. |
| Expansion.List | The average 6-month bill was sold with a yield of 8.04%, up from 7.90%. | The average 3-month issue rose to 8.05% from 7.77%. |
| Temporal.Asynchronous | After practicing law locally, he was elected to his first 10-year term as judge in 1971. | In 1981, he was effectively re-elected |

*Note*: For each implicit discourse relation example, we show only the first-level and the second-level discourse relations because we only conducted experiments on these two levels. For example, "Comparison.Contrast" denotes that the first-level discourse relation is "Comparison" and the second-level is "Contrast."

Table II. Implicit Discourse Relation Examples in CDTB

| Discourse relation | Arg1 | Arg2 |
|---|---|---|
| Causation | 路 通 财 就 通 | 宁波 的 腾飞 之 日 为时不远 |
| Conditional | 按 原 产地 原则 统计 | 产品 的 出口国 就 从 这些 国家 和 地区 转移 到 了 中国 |
| Conjunction | 今年 七月 中旬 以来 ， 淮河 中 、 下游 又 发生 大 面积 水污染 事故 ， 使 一些 地区 的 群众 饮水 发生 困难 | 直接 危及 了 下游 人民 的 生活 和 生产 |
| Contrast | 六十五 岁 的 香港 居民 卢光辉 五 年 前 回 广东 探亲 时 ， 从 深圳 坐 汽车 到 广州 花费 了 四 个 小时 | 今年 ， 他 从 深圳 回到 广州 仅 用 了 九十 分钟 |

## 3. MODEL

We propose a max-margin based neural network model for implicit discourse relation recognition. The discourse relation scoring model will be presented in Section 3.1, followed by the details of the max-margin learning framework in Section 3.2.

### 3.1. Discourse Relation Scoring Model

As shown in Figure 1(b), each discourse relation argument pair is represented as two dense embeddings $a_1$ and $a_2 \in \mathbb{R}^{H_1}$ where $H_1$ is the size of the embeddings. The representation learning of $a_1$ and $a_2$ will be introduced in Section 4. Then, $a_1$ and $a_2$ serve as the input of discourse relation scoring model in which we design multiple kinds of hidden layers. Above the hidden layer, the model outputs a confidence score for specific relations using a linear transformation of the following function:

$$f(a_1, a_2) = W^T h,$$

where $h \in \mathbb{R}^{H_2}$ denotes hidden layer representation and $W \in \mathbb{R}^{H_2}$ denotes the linear transformation vector. $H_2$ is the size of the hidden layer.

To better investigate the representation of the hidden layer $h$, we apply multiple types of networks: Single-layer Neural Network, Tensor Neural Network, and Transformation (Trans) Neural Network.

*3.1.1. Single-Layer (SL) Model.* This model is the simplest form of neural network containing only one hidden layer. It is defined as follows:

$$h = tanh(W_s[a_1; a_2] + b_s),$$

where $W_s \in \mathbb{R}^{H_2 \times 2H_1}$ and $b_s \in \mathbb{R}^{H_2}$. $[a_1; a_2] \in \mathbb{R}^{2H_1}$ denotes concatenation of $a_1$ and $a_2$.

This model concatenates $a_1$ with $a_2$ as input to a nonlinear hidden layer, providing weak interaction between the two arguments.

*3.1.2. Tensor Model.* A tensor is a multidimensional array that can connect two input vectors of every dimension. The tensor model has been widely used in many NLP tasks [Socher et al. 2013; Pei et al. 2014]. It can be defined as follows:

$$h = tanh\big(a_1^T W_t^{[1:H_2]} a_2 + W_s[a_1; a_2] + b_t\big),$$

where $W_t^{[1:H_2]} \in \mathbb{R}^{H_1 \times H_1 \times H_2}$ is a $H_2$-way tensor and $b_t \in \mathbb{R}^{H_2}$.

Since the work of Pitler et al. [2009], pairwise features, such as word pairs, have been regarded as indispensable parts of lexical features for implicit discourse relation recognition. Pairwise features can appropriately reveal the interaction between two discourse arguments. In the SL, the input embedding pair is simply concentrated, which is hard to model the deep interactions. In contrast, the tensor model can be regarded as an effective tool to mine interactions between different features. Intuitively, different explicit interactions among argument pairs at the level of sentences, arguments, and words can be modeled by each slice of the tensor independently.

*3.1.3. Trans Model.* This model intends to explicitly explore relations between argument pairs by modeling the relative position information of two arguments in embedding space, which can be defined as follows:

$$h = tanh(W_e(a_1 + r - a_2) + W_s[a_1; a_2] + b_t),$$

where $W_e \in \mathbb{R}^{H_2 \times H_1}$ $b_e \in \mathbb{R}^{H_2}$ and $r \in \mathbb{R}^{H_1}$.

The transformation operation can be explained as follows: if $Arg1$ and $Arg2$ hold a relation $rel$, there should be a specific spatial relationship measure that captures the relation between these two arguments. To be straightforward, we expect a transformation embedding $r$ representing relation $rel$ so that $a_1$ can correlate to $a_2$ after adding $r$. The motivation comes from the work of Mikolov et al. [2013], in which the authors state that semantic relations existing between word pairs could be found in an embedding space such as $Paris$ - $France$ = $Rome$ - $Italy$. The work most related to our Trans model is the study of Bordes et al. [2013], in which the authors propose a transformation based model (TransE) to learn the relations between entities. Their score function is defined as follows:

$$||h + r - t||_{L_1/L_2},$$

where $h$, $r$, and $t$ denote head entity, relation, and tail entity, respectively.

The major difference is that the TransE model measures the entities relation transformation using dissimilarity measures, such as L1 or L2-norm, while in our Trans model, we apply transformation in the form of a vector to construct the hidden layer. In other words, the objective of the TransE model is to make $a_1$ as close to $a_2$ as possible after adding $r$, while our model intends to associate $a_1$ and $a_2$ with $r$, which can retain more relation transformation information for the subsequent training.

*3.1.4. TTNN Model.* We integrate Tensor and Trans Neural Network to create a new model called TTNN, which is defined as follows:

$$h = tanh\big(a_1{}^T W_t^{[1:H_2]} a_2 + W_e(a_1 + r - a_2) + W_s[a_1; a_2] + b_t\big).$$

This model can verify discourse relations from multiple perspectives. The tensor model focuses on modeling the discourse relations from the point of view of interaction between arguments, which can be regarded as an extreme form of feature combination. The Trans model intends to explore discourse relations between arguments from the point of view of relative position information of two arguments in embedding space.

Tensor and Trans models can learn discourse relations from different perspectives, therefore our combined multiview model should have much more expressive power than a single model.

### 3.2. Max-Margin Learning

After we obtain the relation score of the discourse argument pair, we apply the max-margin learning framework to optimize the neural network. We define two objective functions for different implicit discourse relation recognition tasks, that is, binary classification for English first-level discourse relations and multiclass classification for English second-level discourse relations and Chinese discourse relations.

For binary classification, given a training set $R$ of all the $(a_1, a_2)$ pairs with the specific discourse relations, we minimize an objective function defined as follows:

$$L_1(\theta) = \sum_{(a_1,a_2)\in R} \sum_{(a_1',a_2')\notin R} max\{0, 1 - f(a_1, a_2) + f(a_1', a_2')\} + \lambda \|\theta\|_2^2. \quad (1)$$

For each positive discourse argument pair $(a_1, a_2)$, we randomly sample a certain number of negative pairs $(a_1', a_2')$ that do not hold the same discourse relation as $(a_1, a_2)$. $L_2$ regularization is used to penalize the size of all the parameters to prevent overfitting, weighted by $\lambda$. The objective function $L_1$ favors a higher score for positive training pairs than for negative pairs.

In the testing phase, for each one of the four binary classification tasks, we first use the development set to obtain a threshold $T_{rel}$ for relation $rel$ so that for each argument pair in the testing set if $f(a_1, a_2) \geqslant T_{rel}$, then $(a_1, a_2)$ holds the relation $rel$.

For multiclass classification, we minimize an objective function defined as follows:

$$L_2(\theta) = \sum_{(a_1,a_2)\in R} \sum_{f':f'\neq f} max\{0, 1 - f^+(a_1, a_2) + f^-(a_1, a_2)\} + \lambda \|\theta\|_2^2.$$

For each discourse argument pair $(a_1, a_2)$ holding the specific discourse relation $rel_i$, we score it with $\theta_{rel} = \{W^{rel}, W_s^{rel}, W_t^{rel}, W_e^{rel}, b^{rel}\}_{rel=rel_i}$ as $f^+(a_1, a_2)$ and with $\theta_{rel'} = \{W^{rel'}, W_s^{rel'}, W_t^{rel'}, W_e^{rel'}, b^{rel'}\}_{rel'\neq rel_i}$ as $f^-(a_1, a_2)$. The objective function $L_2$ favors a higher score for training pair $(a_1, a_2)$ with a series of parameters corresponding to its class $rel_i$ rather than with any other series of parameters corresponding to the class $rel'$ which is not $rel_i$.

In the testing phase, for each argument pair $(a_1, a_2)$, we score it using a series of parameters for all relations, among which the relation $rel$ with the highest score is held.

## 4. MULTILEVEL DISTRIBUTED REPRESENTATIONS OF THE ARGUMENTS

It is crucial to effectively represent the arguments before recognizing the relations using discourse relation scoring modes. Prior work mostly devotes more effort to exploring various surface features or reducing the sparsity of these surface features, which cannot capture the features at the sentence level. In recent years, researchers resort to a recursive neural network or to a convolution neural network to obtain segment level information, but the word level information is ignored. Furthermore, syntactic features, that is, production rules, have been proven to be more effective than the lexical features for first-level [Zhou et al. 2010] and second-level [Lin et al. 2009] discourse relation recognition. This motivates us to seek a novel approach that covers not only the multilevel features from token to segments, but also from both lexical and syntactic features.

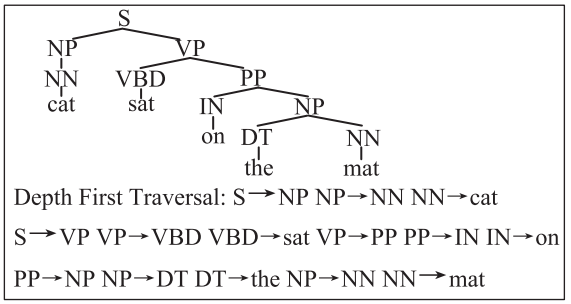



Fig. 2. Linearizations of a parse tree via depth-first traversal.

*Token Level Features*. Pitler et al. [2009] proposes to use the first three and the last words of the argument as features, where connective-like expressions often appear, achieving the best performance among lexical and linguistically informed features. In the meantime, word embeddings learned by using a large corpus have been extensively applied in neural network based methods for many NLP tasks. In fact, pretrained word embeddings obtained by unsupervised methods have shown superiority over randomly initialized ones in many deep learning frameworks. Thus, we introduce embeddings of tokens that are located at the first three and the last positions of the arguments, which is called First3-Last embedding. To obtain the First3-Last embedding, we can simply concatenate or average the embeddings of the first three and the last tokens. These multilevel representations from tokens to segments are used to initialize our network input layer and the token-level representations can be fine-tuned during training.

*Segment Level Features*. In addition to token level embeddings, segment embeddings are also indispensable. Segment embeddings learned using a small corpus PDTB[1] through supervised methods [Ji and Eisenstein 2015; Zhang et al. 2015] cannot beat the surface features, which is consistent with the conclusions of Braud and Denis [2015]. The Paragraph Vector Model, as the augmentation of the Word2vec model, can learn segment-level representations in an unsupervised way. Based on Word2vec (skip-gram), the sentence embedding is regarded as an additional embedding beyond words to predict the target words within the scope of a sentence. After the optimization of the same objective as Word2vec, we can obtain the final sentence embeddings. Specifically, to obtain the distributed representation of arguments, we assign to them vectors that participate in predicting the target word as sentence vectors do. An example is shown in Figure 1(a). In this way, we obtain sentence and argument embeddings. "Sentence" here means an argument pair (a discourse relation instance), which is the same here after.

*Syntactic Features*. To infer the implicit relation between two arguments, the structure difference between them may provide some clues. Lin et al. [2009] propose to employ the production rules extracted from constituent parse trees as features and since then these features have been widely used in implicit discourse relation recognition. Some production rule examples in Figure 2 are S → NP VP, NN → "cat." Li and Nenkova [2014] propose a "stick" version of production rules by splitting all the children of a (parent, children) production rule into several sticks where each one only contains one child. For instance, S → NP VP is converted to S → NP and S → VP. Through depth-first traversal, we linearize a constituent parser tree to a production stick sequence, which is shown in Figure 2. Then, we can learn multilevel distributed

[1]PDTB contains only 16,053 implicit discourse relation instances.

Table III. Distribution of First-Level Implicit Discourse Relation in PDTB

| | Number of instances | | |
|---|---|---|---|
| Discourse Relations | Train | Test | Dev. |
| Comparison | 1,944 | 152 | 197 |
| Contingency | 3,346 | 279 | 292 |
| Expansion | 7,011 | 574 | 671 |
| Temporal | 760 | 85 | 64 |
| Total | 13,061 | 1,090 | 1,224 |

representations of syntactic features (from production stick tokens to production stick sequences) in the same way as lexical features.

## 5. IMPLEMENT DETAILS

We use L-BFGS-B [Zhu et al. 1997] with Batch Normalization [Ioffe and Szegedy 2015] to optimize our model in which we update the model parameters $\{W, W_s, W_t, W_e, b\}$ and word embeddings. We also tried AdaGrad [Duchi et al. 2011] but find that it does not work well. We apply norm clipping [Pascanu et al. 2013] with a threshold of 5 to overcome the gradient exploding problem and early stopping [Prechelt 1998] with the development set to avoid overfitting.

Regarding hyperparameters, we select the dimensions of sentence, argument, and word embedding $d$ among {25, 50, 100}, learning rate $\eta$ among {0.01, 0.001, 0.0001}, regularization parameter $\lambda$ among {0.01, 0.001, 0.0001}, number of negative samples for binary classification $N$ among {10, 30, 50, 100}, and size of the hidden layer as well as number of slices in tensor $H_2$ among {3, 5, 10, 15}. The optimal configurations are determined according to the performance on the development set. The chosen configurations are $d$=25, $\eta$=0.001, $\lambda$=0.0001, $N$=50. For binary classification, $H_2$=10, while for multiclass classification, $H_2$=3.

## 6. EXPERIMENTS

We test our method on both English and Chinese including three tasks. The first task is PDTB first-level discourse relation binary classification. The second is PDTB second-level discourse relation multiclass classification. The third is CDTB discourse relation multiclass classification, which is used to prove the generalization of our model for different languages.

### 6.1. PDTB First-Level Implicit Discourse Relation Recognition

*6.1.1. Experimental Settings.* The task for PDTB first-level implicit discourse relation binary classification is to construct a "one-versus-rest" model for each first-level discourse relation. Following the previous work [Pitler et al. 2009; Zhou et al. 2010; Rutherford and Xue 2015], we use sections 2-20 of PDTB as the training set, sections 0-1 as the development set, and sections 21-22 as the test set. Note that the data preparation for the EXPANSION relation follows the work of Zhou et al. [2010] and Rutherford and Xue [2015]. It is different from the work of Pitler et al. [2009] and Ji and Eisenstein [2015] in which they regard the EntRel relation as a part of EXPANSION. The distribution of first-level implicit discourse relations in PDTB is shown in Table III.

To evaluate the effect of syntactic features on real world data, we do not use the gold standard parse results provided by the Penn Treebank. Our constituent parse results are obtained by using the Stanford Parser [Klein and Manning 2003]. We also employ lowercasing and tokenization. To enlarge the data scale for Paragraph Vector Model training, we use large-scale unlabeled monolingual data from Reuters. From the raw Reuters data, we choose only the sentences in which all the words should appear in PDTB to avoid noise. The selected Reuters corpus contains 1.7 billion tokens and 67.2

Table IV. The Performance (F1-score/%) on Recognizing English First-Level Implicit Discourse Relation with Different Features and Models

| | SL Model | Tensor Model | Trans Model | TTNN Model |
|---|---|---|---|---|
| | | COMPARISON | | |
| Lexical Features | 36.31 | 40.01 | 38.29 | **41.82** |
| Syntactic Features | 36.75 | 39.69 | 39.34 | 41.39 |
| | | CONTINGENCY | | |
| Lexical Features | 48.49 | 51.30 | 50.10 | 52.31 |
| Syntactic Features | 48.53 | 52.35 | 51.29 | **54.17** |
| | | EXPANSION | | |
| Lexical Features | 65.69 | 70.11 | 71.07 | 71.03 |
| Syntactic Features | 66.83 | 70.90 | 70.91 | **71.08** |
| | | TEMPORAL | | |
| Lexical Features | 28.81 | 31.64 | 30.09 | 32.75 |
| Syntactic Features | 29.12 | 31.81 | 31.57 | **34.04** |

Table V. Performance (F1-score/%) Comparison of Different Systems for English First-Level Discourse Relation Binary Classification

| | COMPARISON | CONTINGENCY | EXPANSION | TEMPORAL |
|---|---|---|---|---|
| Pitler et al. [2009] | 21.96 | 47.13 | — | 16.76 |
| Zhou et al. [2010] | 31.79 | 47.16 | 65.95 | 20.3 |
| Rutherford and Xue [2014] | 39.70 | 54.42 | 70.23 | 28.69 |
| Ji and Eisenstein [2015] | 35.93 | 52.78 | — | 27.63 |
| Braud and Denis [2015] | 36.36 | 55.76 | 67.42 | 29.3 |
| Zhang et al. [2015] | 34.22 | 52.04 | 69.59 | 30.54 |
| Liu et al. [2016]* | 37.91 | **55.88** | 69.97 | **37.17** |
| TTNN (ours) | **41.91** | 54.72 | **71.54** | 34.78 |

*Note*: We do not include the result of Pitler et al. and Ji and Eisenstein for EXPANSION because they regard EntRel relation as a part of EXPANSION, which is different from other works including ours. Liu et al. [2016] uses three kinds of additional labeled discourse datasets to jointly train with PDTB implicit discourse relation instances.

million sentences. We obtain First3-Last embeddings via an averaging operation and fix them during training. A detailed comparison of different First3-Last embedding will be given on the second-level classification task.

With different lexical and syntactic features, that is, production rules, we test SL, Tensor, Trans, and hybrid TTNN models. The results are reported in Table IV. Finally, we integrate lexical and syntactic information by summing up the confidence scores obtained from the models with these two features for each instance. Table V presents the final performance of our model compared with that of other competitive methods.

*6.1.2. Experimental Results.* In this section, we try to answer three questions: (1) Which model for discourse relation scoring performs better? (2) Which kinds of distributed features are more effective? (3) Can our method surpass the best reported results?

The detailed experimental results listed in Table IV can answer the first two questions. Overall, among all four models, the hybrid TTNN is superior to others. Among the three single models, the Tensor model has similar performances to the Trans model, which is obviously better than the SL model. Regarding the features, the syntactic features perform better than the lexical features and achieve the best performance in most cases over the four relations.

The results shown in Table V can answer the last question. The experimental results in Table V tell us that our method can achieve the best performance in COMPARISON and EXPANSION relation recognition tasks when compared to the state-of-the-art

Table VI. Distribution of Second-Level Implicit Discourse Relation in PDTB

| Discourse Relations | | Number of Instances | | |
|---|---|---|---|---|
| First-level | Second-level | Train | Test | Dev |
| Comparison | Concession | 184 | 17 | 15 |
| | Contrast | 1,610 | 134 | 171 |
| | Pragmatic concession | 1 | 0 | 0 |
| | Pragmatic contrast | 4 | 0 | 0 |
| Contingency | Cause | 3,277 | 272 | 284 |
| | Pragmatic cause | 64 | 7 | 7 |
| | Condition | 1 | 0 | 0 |
| | Pragmatic condition | 1 | 0 | 0 |
| Expansion | List | 338 | 12 | 10 |
| | Conjunction | 2,882 | 209 | 264 |
| | Instantiation | 1,102 | 122 | 108 |
| | Alternative | 152 | 9 | 10 |
| | Restatement | 2,458 | 216 | 271 |
| | Exception | 2 | 0 | 0 |
| Temporal | Asynchronous | 555 | 57 | 50 |
| | Synchrony | 204 | 28 | 13 |
| Total | | 12,835 | 1,083 | 1,203 |

approaches. It achieves F-score improvements over the state-of-the-art by 2.21% and 1.31%, which is significantly better than the state-of-the-art (McNemar's Chisquared test, $p < 0.05$). We obtain competitive results for the CONTINGENCY and TEMPORAL relations. These results demonstrate that our method is promising for PDTB first-level implicit discourse relation inference.

### 6.2. PDTB Second-Level Discourse Relation Recognition

*6.2.1. Experimental Settings.* This task belongs to multiclass classification. Following the work of Ji and Eisenstein [2015], sections 2-20 of PDTB are used as the training data, sections 0-1, as the development set, and sections 21-22, as the test set. For this task, we only implement our TTNN model because it performs best with respect to binary classification. The distribution of second-level implicit discourse relations in PDTB is shown in Table VI where we can see that there are totally 16 second-level relations. Whereas, there are too few instances for the relations of Pragmatic concession, Pragmatic contrast, Condition, Pragmatic condition, and Exception. Some examples in PDTB are only annotated with the first-level discourse relations. These two kinds of instances are beyond the consideration of our experiment.

To our knowledge, this is the first work to use the first three and the last tokens, that is, words and production rules, embeddings to infer discourse relations. Thus, in this task, we implement extra experiments to evaluate the validity of these token-level embeddings of lexical and syntactic features.

We represent First3-Last by concatenating or averaging its token embeddings and compare First3-Last embedding with sentence and argument embeddings independently. Then we concatenate First3-Last embedding with sentence and argument embeddings. Moreover, we evaluate whether it is necessary to update the token-level embeddings in the training process.

Finally, to compare our model more precisely with other neural network based methods, we choose the best model with distributed representations and add standard surface features as they did. Following Lin et al. [2009], we apply feature selection to obtain 500 word pair features, 100 production rule features, 100 dependency rule features, and 600 Brown cluster features. The difference lies in the fact that we use Information Gain (IG) instead of Mutual Information (MI) as selection criteria because of its better performance [Yang and Pedersen 1997]. The objective function is designed

Table VII. Experimental Results (Accuracy/%) for English Second-Level Implicit Discourse Relation Classification Using Our TTNN Model with Different Embedding Layers

| | | Lexical features | Syntactic features |
|---|---|---|---|
| Static | Sentence | 31.85 | 32.76 |
| | Argument | 37.38 | 38.09 |
| | Sentence + Argument | 38.19 | 38.59 |
| | First3-Last (con.) | 32.76 | 29.44 |
| | First3-Last (ave.) | 34.57 | 31.15 |
| | Sentence + Argument + First3-Last (con.) | 38.09 | 39.20 |
| | Sentence + Argument + First3-Last (ave.) | **40.90** | 39.89 |
| Dynamic | Sentence + Argument + First3-Last (con.) | 35.97 | 36.20 |
| | Sentence + Argument + First3-Last (ave.) | 40.70 | 39.49 |

*Note*: "Sentence," "Argument," and "First3-Last" denote Sentence, Argument, and First3-Last token embeddings, respectively. "con." and "ave." denote the concatenating and averaging operations, respectively. "Static" denotes keeping the token-level embeddings unchanged during training, while "Dynamic" denotes updating them.

Table VIII. Performance Comparison of Different Models for English Second-Level Implicit Discourse Relation Classification

| Models | Features | Accuracy (%) |
|---|---|---|
| Lin et al. [2009] | 1. surface features only | 40.20 |
| DISCO2 (Ji and Eisenstein [2015]) | 2. surface features only | 40.66 |
| | 3. distributed words features | 36.98 |
| | 4. + entity semantics | 37.63 |
| | 5. + surface features | 44.59 |
| TTNN (ours) | 6. surface features only | 40.52 |
| | 7. lexical features | 40.90 |
| | 8. syntactic features | 39.94 |
| | 9. lexical and syntactic features | 41.39 |
| | 10. lexical and syntactic + surface features | **44.75** |

as follows:

$$f(a_1, a_2) = U^T g\big(W_s[a_1; a_2] + {a_1}^T W_t^{[1:H_2]} a_2 + W_e(a_1 + r - a_2) + W_{sur} v + b\big),$$

where $W_{sur} \in \mathbb{R}^{H_1 \times d}$ and $v \in \mathbb{R}^d$ denotes the selected surface feature vector.

*6.2.2. Experimental Results.* Table VII can answer three questions about the embeddings input to our discourse relation scoring model: (1) What type of embeddings are more effective? (2) What type of First3-Last representation is better? (3) Do we need to update the token embeddings? As shown in the first five lines of Table VII, argument embeddings are the most effective, while sentence embeddings are the worst. From the remaining lines in Table VII, we can conclude that averaging is better than concatenation for First3-Last embedding composition. Although concatenation can introduce the word order information, it may lead to the sparsity problem due to separate treatments being used for each word located at the first three and last positions of the arguments. Updating the token-level embeddings during training does not contribute to the classification accuracy perhaps because of the overfitting problem in this model.

Table VIII shows the final results of our model compared with other competitive systems including the model of Lin et al. [2009] and the DISCO2 model proposed by Ji and Eisenstein [2015]. For surface features, our model achieves a performance similar to that of the other two systems. When excluding discrete surface features, the classification accuracy of our model is significantly better than that of DISCO2, with a 3.92% improvement using only word level features, which corresponds to line 3 and 7, and with a 3.76% improvement using all distributed features, which corresponds to

Table IX. The Distribution of Samples of Implicit Discourse Relations in CDTB 0.5

| Relations | Number of Instances | | |
|---|---|---|---|
| | Train | Test | Dev. |
| Causation | 88 | 18 | 5 |
| Conditional | 14 | 2 | 1 |
| Conjunction | 2,541 | 244 | 174 |
| Contrast | 21 | 6 | 1 |
| Expansion | 640 | 72 | 55 |
| Progression | 4 | 0 | 0 |
| Purpose | 39 | 2 | 1 |
| Temporal | 7 | 0 | 0 |
| Total | 3,354 | 344 | 237 |

line 4s and 9. Note that DISCO2 does not beat Lin's purely surface feature model with a gap in accuracy of 2.57%, corresponding to lines 1 and 4. Our model outperforms the surface feature based model (statistically significant, $p < 0.05$; t-test, corresponding to lines 9 and 6). Finally, when surface features are added to our model, we can achieve the best accuracy of 44.75%. Note that the score function of the DISCO2 model can be regarded as a special case of the one-way Tensor model without a hidden layer. Therefore, our model has more expressive power.

### 6.3. CDTB Implicit Discourse Relation Recognition

*6.3.1. Experimental Settings.* For CDTB 0.5, we set section 0001-0700 as the training set, section 0701-0760 as the test set, and section 0761-0803 as the development set. The distribution of discourse relations in CDTB is shown in Table IX.

We evaluate three categories of surface features including the lexical and syntactic features, which have been used in English implicit discourse relation recognition of the prior work and the punctuation features that are first proposed specifically for Chinese.

For the lexical features, we use word pair feature [Marcu and Echihabi 2002], Brown cluster pair feature [Rutherford and Xue 2014], and First3-Last word feature [Pitler et al. 2009]. In addition, we evaluate character based lexical features including character pair and First3-Last character feature. We use the Brown 3200 clusters.[2] For syntactic features, we employ production rule features and dependency rule features proposed by Lin et al. [2009]. But the difference lies in the fact that we use the "stick" [Li and Nenkova 2014] version of them. Our constituent and dependency parse results are obtained using the Stanford Parser [Klein and Manning 2003].

Punctuations in Chinese have some discourse functions, for example:

(a) [中国吸引外资]$_{Arg1}$、[引进技术]$_{Arg2}$

(China attracts foreign capital and introduces technology.)

(b) [青海 油田 新年 再 传 捷报]$_{Arg1}$:[截至九七年 十二月 末, 油气 产量达 一百六十万吨]$_{Arg2}$

(Qinghai oil field spreads good news in the New Year, stating that the oil and gas production reached 1.6 million tons by late December 1997.)

Example (a) is an argument pair holding a CONJUNCTION relation. The pause mark expresses the parallel relationship between *Arg*1 and *Arg*2, corresponding to "and" in English. This usage can also be found for the semicolon. Example (b) is an argument pair holding an EXPANSION relation, which can be inferred by the colon between the two discourse arguments.

[2]http://www.cs.brandeis.edu/clp/conll16st/data/gigaword-zh-c3200.txt.

Table X. The Statistics of Punctuations Located Between Two Arguments in the Training Set and the Test Set of CDTB

| Training Set | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | Pause Mark | Semi-colon | Colon | Period | Comma | Ellipsis | Exclamation Point | Question Mark |
| Expansion | 0 | 0 | 16 | 455 | 166 | 0 | 0 | 0 |
| Conjunction | 122 | 114 | 0 | 858 | 1446 | 1 | 0 | 0 |
| Progression | 0 | 0 | 0 | 3 | 1 | 0 | 0 | 0 |
| Causation | 0 | 0 | 0 | 23 | 65 | 0 | 0 | 0 |
| Conditional | 0 | 0 | 0 | 1 | 13 | 0 | 0 | 0 |
| Contrast | 0 | 2 | 0 | 11 | 0 | 0 | 0 | 0 |
| Purpose | 0 | 0 | 0 | 0 | 39 | 0 | 0 | 0 |
| Temporal | 0 | 0 | 0 | 0 | 7 | 0 | 0 | 0 |
| Contrast | 0 | 0 | 0 | 0 | 8 | 2 | 0 | 0 |
| Test Set | | | | | | | | |
| Expansion | 0 | 0 | 6 | 53 | 12 | 0 | 0 | 1 |
| Conjunction | 9 | 2 | 0 | 109 | 123 | 0 | 1 | 0 |
| Progression | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Causation | 0 | 0 | 1 | 6 | 11 | 0 | 0 | 0 |
| Conditional | 0 | 0 | 0 | 0 | 2 | 0 | 0 | 0 |
| Contrast | 0 | 0 | 0 | 5 | 0 | 0 | 0 | 0 |
| Purpose | 0 | 0 | 0 | 0 | 2 | 0 | 0 | 0 |
| Temporal | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Contrast | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |

Therefore, we regard the punctuation located between two arguments as features. There are eight types of punctuations including comma, period, question mark, semicolon, pause mark, exclamation point, ellipsis, and colon. The detailed statistics for each punctuation is shown in Table X. Except for the last three punctuations in Table X, that is, ellipsis, exclamation point, and question mark, which appear no more than three times in both the training and the test set of CDTB, the other punctuations can be good indicators for differentiating discourse relations. If there is a pause mark or semicolon connecting two arguments, the discourse relation is almost surely CONJUNCTION because these two punctuations are strong indicators for two equal-status statements. If there is a colon connecting two arguments, the discourse relation is very likely to be EXPANSION where one argument is an elaboration or restatement of another. For period and comma, although the discourse relations cannot be directly inferred, they have different tendencies for the most frequent relations, that is, EXPANSION and CONJUNCTION. In the training set, EXPANSION and CONJUNCTION account for 19.1% and 75.8%, respectively. 33.7% of argument pairs connected with a period are EXPANSION relations and 63.5% are CONJUNCTION relations. Whereas, 9.5% and 82.9% of two arguments connected with a comma are EXPANSION and CONJUNCTION relations, respectively. We can conclude that if there is a period between two arguments, the discourse relation tends to be EXPANSION, while CONJUNCTION for comma, which is consistent with intuition that comma connected arguments are more coherent.

For surface features, we discard rare features that appear less than five times in the CDTB training set as Lin et al. [2009] does.

To enlarge the dataset for Paragraph Vector Model training, we use large-scale unlabeled monolingual data from the Sogou Chinese corpus [Liu et al. 2012]. From the raw data, we choose only the sentences in which all the words appear in CDTB to avoid noise. The chosen corpus contains 1.38 million tokens and 251 thousand sentences.

Table XI. The Results of Various Features for Chinese Implicit Discourse Relation Classification with SVM as the Classifier

| Features | | Accuracy (%) |
|---|---|---|
| Most Frequent Relation | | 70.93 |
| Lexical Features | Word Pairs | 59.30 |
| | Character Pairs | 60.17 |
| | First3-Last Word | 61.33 |
| | First3-Last Character | 62.21 |
| | Brown Pairs | 59.88 |
| | Lexical Arg Vectors | 77.33** |
| Syntactic Features | Production Rules | 76.74** |
| | Dependency Rules | 60.76 |
| | Syntactic Arg Vectors | 77.10** |
| Punctuation | | 72.38* |
| All | | 78.77** |

*Note*: "Arg vectors" denote multilevel representations of arguments. "*" denotes statistical significantly better than the baseline, $p < 0.05$, t-test. "**" denotes $p < 0.01$.

We fix First3-Last embeddings in the training process with respect to the fact that updating them performs badly in the task of English.

### 6.4. Experimental Results

To evaluate the effectiveness of various surface features and multilevel distributed features, we resort to LIBSVM [Chang and Lin 2011] with default parameter values as classifier as Rutherford and Xue [2014] does. The result is shown in Table XI. As we can see from Table IX, the most frequent relation, that is, CONJUNCTION, accounts for 70.93% of the test set, which can be regarded as the baseline accuracy. From Table XI, we can see that all the surface lexical features and dependency rule features perform worse than the baseline, which indicates that these features that are suitable for English implicit relation recognition do not work well for Chinese. Comparing word-based and character-based lexical features, we can conclude that character-based features perform slightly better than word-based features, but both of them are much lower than the baseline. Production rule features achieve an accuracy of 76.74%, which is obviously better than any other surface features. The performance of punctuation features slightly exceeds the baseline. Our proposed multilevel distributed features including lexical and syntactic argument vectors perform better than the surface features, achieving 77.33% and 77.1% accuracy, respectively. When integrating all the features, we can achieve the best recognition accuracy of 78.77%.

The difference in performance of surface lexical and syntactic features accords with the result of Lin et al. [2009], in which the English implicit discourse relation recognition accuracy using production rules is approximately 6% higher than that using dependency rules and word pairs. This gap is much bigger in our experiment of Chinese implicit discourse relation recognition (approximately 16%). On the one hand, we can conclude that production rule features can capture more information such as Part of Speech (POS), which are useful for implicit discourse relation recognition for both English and Chinese. On the other hand, lexical features including word pairs, Brown pairs, and First3-Last and dependency rule features are all pairwise features, which comprise words or dependency types, while for production rules, despite the pairwise features, certain words associate with certain types of POS, and the number of POS tags is rather small, thus production rule features can be less sparse than other surface features. For the training set of CDTB, we extract 5,055 production rules,

Table XII. Experimental Results (Accuracy/%) for Chinese Implicit Discourse Relation Classification with TTNN Model with Different Embedding Layers

| | Lexical Features | Syntactic Features |
|---|---|---|
| Sentence | 70.93 | 71.22 |
| Argument | 78.49 | 78.49 |
| First3-Last | 72.38 | 69.77 |
| Sentence + Argument | 75.58 | 76.16 |
| Sentence + First3-Last | 71.80 | 70.05 |
| Argument + First3-Last | 79.06 | 77.03 |
| Sentence + Argument + First3-Last | **79.36** | 78.48 |

*Note*: "Sentence," "Argument," and "First3-Last" denote Sentence, Argument, and First3-Last token embeddings, respectively. Note that the First3-Last vector is generated by averaging operations and we keep the token-level embeddings unchanged during training.

Table XIII. The Results of Chinese Implicit Discourse Relation Classification with TTNN Model

| Features | Accuracy (%) |
|---|---|
| Lexical Arg Vectors | 79.36** |
| Syntactic Arg Vectors | 78.48** |
| Lexical + Syntactic Arg Vectors | 79.65** |
| Lexical + Syntactic Arg Vectors + Surface Features | 82.56** |

*Note*: "Arg vectors" denote multilevel representations of arguments. "*" denotes statistical significantly better than the baseline, $p < 0.05$, t-test. "**" denotes $p < 0.01$.

more than 50,000 lexical features, and 7,693 dependency rules. Meanwhile, the total number of implicit discourse instances in CDTB is much smaller than that of PDTB, thus CDTB may suffer from a more severe sparsity problem.

Table XII shows the performance of different types of embeddings based on lexical and syntactic features with the TTNN model. As shown in the first three lines of Table XII, argument embeddings are the most effective independent embeddings, while sentence embeddings are the worst. From the remaining lines in Table XII, we can conclude that when we combine different types of embeddings, the performance will be better. We get the highest accuracy when we use all kinds of embeddings. We can conclude that our model with multilevel lexical and syntactic argument vectors achieves an accuracy of 79.36% and 78.48%, which is obviously better than any discrete lexical features.

To further evaluate our model, we integrate the relatively effective surface features, that is, production rules and punctuation features, into our model as we do for English. Motivated by Lin et al. [2009], we apply feature selection with information gain to obtain the top 100 production rule features. The results are shown in Table XIII. We obtain an accuracy of 79.65% using both lexical and syntactic argument vectors. Along with production rule and punctuation features, the accuracy of our neural network based model is 82.56%, which outperforms the baseline with a remarkable improvement of 11.63%.

## 7. DISCUSSION

In this section, we perform more detailed analyses to further illustrate the effectiveness of our approach.

To better understand the strength of our multilevel distributed representations, some discourse relation instances that are extracted from the test sets of PDTB and CDTB are given in Table XIV. The recognition results of these instances are incorrect by the model using discrete surface features, while correct using the distributed representations.

Table XIV. Several Implicit Discourse Relation Instances Extracted From Test Set of PDTB and CDTB

| | | |
|---|---|---|
| 1 | Discourse relation | CONTRAST |
| | Arg1 | The common view is that there will be mild economic growth, modest profit expansion, and things are going to be hunky-dory. |
| | Arg2 | Our view is that we may see a profit decline. |
| 2 | Discourse relation | CONTRAST |
| | Arg1 | 二十多年前，广东东莞水口村村民陈忠发是个地地道道的农民，日出而作，日落而息，和妻子一道种田挣工分维持一家七口人的生活。对那时的他而言，将自己落后零散的村庄改造得象城市一样，自己也能象城里人一样生活、工作，实在是个遥不可及的梦。 |
| | Arg2 | 今天，水口村已崛起了座座现代化工业区和厂房，并建起了公园、影剧院、酒家、商场；而陈忠发一家，也和其他村民一样，住上了漂亮的小洋楼，至于电视机、电冰箱、洗衣机、摩托车则早已成了生活必备品。“梦想成真”的陈忠自豪地说：“现在城里人有的，我们也都有。 |
| 3 | Discourse relation | INSTANTIATION |
| | Arg1 | Futures prices declined |
| | Arg2 | The March contract was off 0.32 cent a pound at 13.97 cents |
| 4 | Discourse relation | EXPANSION |
| | Arg1 | “沧海横流，方显出英雄本色” |
| | Arg2 | 面对持续一年多的亚洲金融危机，中国表现出了“大国风范” |

We first explain the necessity of our distributed First3-Last embeddings, and then, we explore the deeper reasons at the sentence level.

Pitler et al. [2009] proposes that connective-like expressions appear at the first three and the last words of the arguments and we find that our distributed First3-Last embeddings have an advantage over their discrete features. We demonstrate this using the first two examples in Table XIV.

In Example 1 of Table XIV, the first three words of *Arg*1 - “`The common view`” - and the first two words of *Arg*2 - “`Our view`” - indicate that it will pose opposite opinions for the two arguments, thus the CONTRAST discourse relation exists between the argument pair. However, this rule does not appear in the training set of PDTB. In other words, it is impossible to detect the discourse relation by using the discrete First3-Last features. “二十多年前...梦” and “今天...有” in example 2 express the CONTRAST discourse relation between the two arguments. But similarly, such a case only occurs in the test set of CDTB. In contrast, our distributed First3-Last representation can capture these connective-like expressions and recognize the discourse relation successfully.

Next, we show the effectiveness of segment-level distributed representations. Lin et al. [2009] explains the difficulties of the English implicit discourse relation recognition and points out four challenges: (1) ambiguity between relations, (2) inference, (3) contextual modeling, and (4) world knowledge. They believe that a deeper semantic representation and a more robust model will be helpful. Distributed representation of words or segments and a neural network model may meet these requirements.

Regarding example 3 in Table XIV, *Arg*2 is an instantiation of *Arg*1. Word pair “`declined, off`” provides a strong indication for this discourse relation, but we find that such a case does not occur in the training set. Thus, it is not surprising that using surface features including word pairs fails to detect the INSTANTIATION discourse relation.

For our distributed representation based model, the recognition result is correct. We seek the most similar argument in the PDTB training set of $Arg1$ and $Arg2$ using cosine similarity in the argument embedding space, yielding "For the first nine months, the trade deficit was 14.933 trillion lire, compared with 10.485 trillion lire in the year-earlier period" (denoted as $Arg1'$) for $Arg1$ and "The stock fell 75 cents" (denoted as $Arg2'$) for $Arg2$. We find that a few words overlap between $Arg1'$ and $Arg1$ as well as between $Arg2'$ and $Arg2$, but there is a relatively high semantic similarity between $Arg1'$ and $Arg1$ as well as between $Arg2'$ and $Arg2$: $Arg1'$ and $Arg1$ both express the meaning of slowdown in terms of the economy; $Arg2'$ and $Arg2$ express that the price of something decreases by a specific number of cents.

This is similar for Chinese example 4 in Table XIV, in which $Arg1$ is extracted from a poem that is a metaphor showing that a person is conscientious even when confronting a crisis, while $Arg2$ is a straightforward expression of a specific case, thus the EXPANSION relation exists between this argument pair.

According to the preceding analyses, we can conclude that our model has the ability to capture deeper semantic meaning, while the discrete surface feature model fails.

## 8. RELATED WORKS

We will introduce the related works about English implicit discourse relation recognition from two hierarchies: modeling and feature engineering.

Regarding modeling, most of the previous work of English implicit discourse relation recognition regards it as a classification task. Typically, SVM, ME, Naïve Bayes classifier, and softmax layer of neural network are applied to determine the discourse relations. Recently, Ji and Eisenstein [2015] employs a neural network based bilinear model to explore the interactions between argument pairs. Compared to the previous work, we further explore relations between discourse argument pairs from another perspective that models the transformation property between two arguments. The study about knowledge base completion [Bordes et al. 2013] has shown extreme effectiveness in modeling the relation transformation between entities in the knowledge base on this point, which motivates us to predict discourse relations from this point of view.

Regarding feature engineering, Marcu and Echihabi [2002] presents an unsupervised approach to use word pair features to identify discourse relations on an artificial corpus. Their training data are generated manually from raw corpora by detecting certain patterns based on cue phrases. For example, if there is a word "but" between two successive sentences, the sentence pairs are extracted as a CONTRAST instance after removing the discourse connectives. They show that lexical features are effective for identifying discourse relations. This work is followed by Saito et al. [2006]. They perform an experiment to combine word pairs and phrasal patterns to recognize discourse relations in Japanese. A phrasal pattern explores the information existing in longer context beyond two sentence pairs. For instance, the pattern " . . . should have done . . ." immediately following " . . . did . . ." can significantly imply that the discourse relation is CONTRAST.

Before the release of PDTB, due to the lack of hand-annotated data of implicit discourse relations, researchers resort to automatically creating implicit examples by removing the connectives in the explicit ones. Thus, it is unclear whether both of the two previous methods would work in natural texts without unambiguous connectives. PDTB provides a benchmark resource for detecting implicit discourse relations and has been widely used by the majority of follow-up work. Pitler et al. [2009] first infer implicit discourse relations using PDTB. They compare the performance of various types of word pair features, and explore the utility of several linguistically informed features, including polarity tags, inquirer tags, verb classes, and modality, showing that these features are useful for recognizing implicit discourse relations. Lin et al. [2009]

is the first to introduce syntactic features, specifically, constituent parse features and dependency parse features. They present an implicit discourse relation classifier on second-level types in the PDTB using syntactic and lexical features. Wang et al. [2010] extends this work by applying the tree kernel method to the syntactic features. Zhou et al. [2010] proposes a method that uses the language models to automatically predict implicit connectives. Xu et al. [2012] extends this work by integrating various linguistically informed features. Park and Cardie [2012] optimizes the combination of lexical and syntactic features and achieved a solid performance. Subsequent works focus on settling the data sparsity problem. Biran and McKeown [2013] aggregates the lexical features by reweighting word pairs. Li et al. [2014] introduces the simplification of the parse tree to relieve the sparsity of syntactic features. Rutherford and Xue [2014] employs Brown cluster to generate more compact word pair features, reducing the feature size from quadratic of the vocabulary size to $3{,}200^2$. Rutherford and Xue [2015] train the model using elaborately collected explicit discourse relation data. Lately, the neural network based approach [Braud and Denis 2015] is proposed to use various types of word representations and representation combination schemes to represent text segments, which also resorts to standard classifier to infer discourse relations. Zhang et al. [2015] presents a shallow convolutional neural network to learn sentence embeddings to classify the discourse relations. Ji and Eisenstein [2015] employs a recursive neural network to learn distributed representations of argument pairs. Their model achieves state-of-the-art performance for second-level implicit discourse relations. Liu et al. [2016] proposes a multitask convolutional neural network that uses various labeled discourse corpora, such as Rhetorical Structure Theory - Discourse Treebank and New York Times Corpus, to enhance the first-level implicit discourse relation recognition performance.

However, without the surface features, the model performance of Ji and Eisenstein [2015] is approximately 3% lower than the purely discrete surface feature model of Lin et al. [2009]. Part of the reason may be that it is difficult to learn sufficiently satisfying representations of sentence pairs with only an extremely small-sized PDTB corpus (there are about 16,000 instances in PDTB). Moreover, production rules extracted from constituent parse trees have been proven more effective [Park and Cardie 2012] than lexical features. Therefore, we believe that it is worth encoding them into distributed representations. We resort to the distributed representations of multilevel lexical and syntactic features learned via an unsupervised approach from a large-scale corpus. In this way, we can not only alleviate the data sparsity problem to a large extent but also take full advantage of both raw text of a word sequence and constituent parser tree in the form of a production rule sequence.

Compared with the implicit discourse relation recognition for English, there are few studies for Chinese. Huang and Chen [2011] builds a corpus of Chinese with human annotated discourse relations and conducts experiments to recognize discourse relations using SVM classifier. Li et al. [2014] presents a discourse relation analysis between Chinese and English, and recognizes discourse relations using the parallel corpus in a semisupervised manner. These two systems do not separate the implicit discourse relations from the explicit ones, while the Chinese implicit discourse relation recognition is a more challenging task.

## 9. CONCLUSIONS AND FUTURE WORK

In this article, we propose a novel method for implicit discourse relation recognition based on a neural network, in which two components, the model hierarchy and the feature hierarchy, are constructed. Regarding the model hierarchy, we propose a max-margin neural network that considers two views, including the relation classification view and the relation transformation view. Regarding the feature hierarchy, we learn

and use distributed representations from multilevels, namely, from words, arguments, and syntactic structures to sentences.

We tested our method for both English and Chinese implicit discourse relation classification. The experimental results demonstrate that our method can achieve new state-of-the-art performance in most cases and substantially outperform the previous competitive approaches. Furthermore, we show for the first time that the distributed features can perform better than the surface discrete features for second-level implicit discourse relation recognition.

In the future, we will devote efforts to encoding more linguistic features into a distributed representation and to exploring more effective methods for representation learning. We will also attempt to detect text spans of discourse arguments and to build an end-to-end discourse parser for English and Chinese texts.

## ACKNOWLEDGMENTS

We thank the three anonymous reviewers for their helpful comments and suggestions.

## REFERENCES

Or Biran and Kathleen McKeown. 2013. Aggregated word pair features for implicit discourse relation disambiguation. In *Proceedings of ACL'13*.

Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. 2013. Translating embeddings for modeling multi-relational data. In *Proceedings of the Advances in Neural Information Processing Systems (NIPS'13)*. 2787–2795.

Chloé Braud and Pascal Denis. 2015. Comparing word representations for implicit discourse relation classification. In *Proceedings of the Empirical Methods in Natural Language Processing (EMNLP'15)*.

Chih-Chung Chang and Chih-Jen Lin. 2011. LIBSVM: A library for support vector machines. *ACM Trans. Intell. Syst. Technol.* 2, 3, Article 27 (May 2011), 27 pages. DOI:http://dx.doi.org/10.1145/1961189.1961199

Philipp Cimiano, Uwe Reyle, and Jasmin Šarić. 2005. Ontology-driven discourse analysis for information extraction. *Data Knowl. Eng.* 55, 1 (2005).

John Duchi, Elad Hazan, and Yoram Singer. 2011. Adaptive subgradient methods for online learning and stochastic optimization. *J. Mach. Learn. Res.* 12 (July 2011), 2121–2159.

Hen-Hsen Huang and Hsin-Hsi Chen. 2011. Chinese discourse relation recognition. In *IJCNLP*. 1442–1446.

Sergey Ioffe and Christian Szegedy. 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167* (2015).

Peter Jansen, Mihai Surdeanu, and Peter Clark. 2014. Discourse complements lexical semantics for non-factoid answer reranking. In *Proceedings of ACL'14*.

Yangfeng Ji and Jacob Eisenstein. 2015. One vector is not enough: Entity-augmented distributed semantics for discourse relations. *Trans. Assoc. Comput. Ling.* 3, 1.

Xiaomian Kang, Haoran Li, Long Zhou, Jiajun Zhang, and Chengqing Zong. 2016. An end-to-end chinese discourse parser with adaptation to explicit and non-explicit relation recognition. In *Proceedings of the 20th Conference on Computational Natural Language Learning: Shared Task (CoNLL)*. 27–32.

Dan Klein and Christopher D. Manning. 2003. Accurate unlexicalized parsing. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1 (ACL'03)*. Association for Computational Linguistics, Stroudsburg, PA, 423–430. DOI:http://dx.doi.org/10.3115/1075096.1075150

Junyi Jessy Li, Marine Carpuat, and Ani Nenkova. 2014. Cross-lingual discourse relation analysis: A corpus study and a semi-supervised classification system. In *COLING*. 577–587.

Junyi Jessy Li and Ani Nenkova. 2014. Reducing sparsity improves the recognition of implicit discourse relations. In *Proceedings of the 15th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. Citeseer, 199.

Ziheng Lin, Min-Yen Kan, and Hwee Tou Ng. 2009. Recognizing implicit discourse relations in the penn discourse treebank. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1 - Volume 1 (EMNLP'09)*. Association for Computational Linguistics, Stroudsburg, PA, 343–351.

Yiqun Liu, Fei Chen, Weize Kong, Huijia Yu, Min Zhang, Shaoping Ma, and Liyun Ru. 2012. Identifying web spam with the wisdom of the crowds. *ACM Trans. Web* 6, 1, Article 2 (March 2012), 30 pages. DOI:http://dx.doi.org/10.1145/2109205.2109207

Yang Liu, Sujian Li, Xiaodong Zhang, and Zhifang Sui. 2016. Implicit discourse relation classification via multi-task neural networks. *arXiv preprint arXiv:1603.02776* (2016).

Annie Louis, Aravind Joshi, Rashmi Prasad, and Ani Nenkova. 2010. Using entity features to classify implicit discourse relations. In *Proceedings of the 11th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL'10)*. Association for Computational Linguistics, Stroudsburg, PA, 59–62.

Daniel Marcu and Abdessamad Echihabi. 2002. An unsupervised approach to recognizing discourse relations. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics (ACL'02)*. Association for Computational Linguistics, 368–375. DOI:http://dx.doi.org/10.3115/1073083.1073145

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781* (2013).

Joonsuk Park and Claire Cardie. 2012. Improving implicit discourse relation recognition through feature set optimization. In *Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL'12)*. Association for Computational Linguistics, Stroudsburg, PA, 108–112.

Razvan Pascanu, Tomas Mikolov, and Yoshua Bengio. 2013. On the difficulty of training recurrent neural networks. In *Proceedings of The 30th International Conference on Machine Learning*. 1310–1318.

Wenzhe Pei, Tao Ge, and Baobao Chang. 2014. Max-margin tensor neural network for Chinese word segmentation. In *Proceedings of ACL'14*, Vol. 1. 293–303.

Emily Pitler, Annie Louis, and Ani Nenkova. 2009. Automatic sense prediction for implicit discourse relations in text. In *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 (ACL'09)*. Association for Computational Linguistics, Stroudsburg, PA, 683–691.

Rashmi Prasad, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind K. Joshi, Bonnie L. Webber, and Nikhil Dinesh. 2008. The penn discourse treebank 2.0. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2008)*. 2961–2968.

Lutz Prechelt. 1998. Automatic early stopping using cross validation: Quantifying the criteria. *Neural Netw.* 11, 4 (1998), 761–767.

Attapol Rutherford and Nianwen Xue. 2014. Discovering implicit discourse relations through Brown cluster pair representation and coreference patterns. In *Proceedings of the EACL*, Vol. 645. 2014.

Attapol Rutherford and Nianwen Xue. 2015. Improving the inference of implicit discourse relations via classifying explicit discourse connectives. In *Proceedings of the NAACL-HLT*.

Manami Saito, Kazuhide Yamamoto, and Satoshi Sekine. 2006. Using phrasal patterns to identify discourse relations. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers (NAACL-Short'06)*. Association for Computational Linguistics, Stroudsburg, PA, 133–136.

Richard Socher, Danqi Chen, Christopher D. Manning, and Andrew Ng. 2013. Reasoning with neural tensor networks for knowledge base completion. In *Proceedings of the 26th International Conference on Advances in Neural Information Processing Systems*. 926–934.

Swapna Somasundaran, Galileo Namata, Janyce Wiebe, and Lise Getoor. 2009. Supervised and unsupervised methods in employing discourse relations for improving opinion polarity classification. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1 (EMNLP'09)*. Association for Computational Linguistics, Stroudsburg, PA, 170–179.

Mei Tu, Yu Zhou, and Chengqing Zong. 2013. A novel translation framework based on rhetorical structure theory. In *Proceedings of ACL'13*, Vol. 2. 370–374.

WenTing Wang, Jian Su, and Chew Lim Tan. 2010. Kernel based discourse relation recognition with temporal ordering information. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL'10)*. Association for Computational Linguistics, 710–719.

Yu Xu, Man Lan, Yue Lu, Zheng Yu Niu, and Chew Lim Tan. 2012. Connective prediction using machine learning for implicit discourse relation classification. In *Proceedings of the 2012 International Joint Conference on Neural Networks (IJCNN'12)*. IEEE, 1–8.

Yiming Yang and Jan O. Pedersen. 1997. A comparative study on feature selection in text categorization. In *Proceedings of ICML*, Vol. 97. 412–420.

Biao Zhang, Jinsong Su, Deyi Xiong, Yaojie Lu, Hong Duan, and Junfeng Yao. 2015. Shallow convolutional neural network for implicit discourse relation recognition. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. 2230–2235.

Yuping Zhou and Nianwen Xue. 2015. The Chinese discourse treebank: A Chinese corpus annotated with discourse relations. *Lang. Res. Eva.* 49, 2 (2015), 1–35.

Zhi-Min Zhou, Yu Xu, Zheng-Yu Niu, Man Lan, Jian Su, and Chew Lim Tan. 2010. Predicting discourse connectives for implicit discourse relation recognition. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters (COLING'10)*. Association for Computational Linguistics, 1507–1514.

Ciyou Zhu, Richard H. Byrd, Peihuang Lu, and Jorge Nocedal. 1997. Algorithm 778: L-BFGS-B: Fortran subroutines for large-scale bound-constrained optimization. *ACM Trans. Math. Softw.* 23, 4 (Dec. 1997), 550–560. DOI: http://dx.doi.org/10.1145/279232.279236

Chengqing Zong. 2013. *Statistical Natural Language Processing*. Tsinghua University Press.

Received May 2016; revised September 2016; accepted December 2016