# Bilingually-constrained Synthetic Data for Implicit Discourse Relation Recognition

**Changxing Wu**[1,2], **Xiaodong Shi**[1,2]*, **Yidong Cheng**[1,2], **Yanzhou Huang**[1,2], **Jinsong Su**[3]
Fujian Key Lab of the Brain-like Intelligent Systems, Xiamen University, China[1]
School of Information Science and Technology, Xiamen University, China[2]
School of Software, Xiamen University, China[3]
{wcxnlp, huangyanzhou163}@163.com
{mandel, ydchen, jssu}@xmu.edu.cn

## Abstract

To alleviate the shortage of labeled data, we propose to use bilingually-constrained synthetic implicit data for implicit discourse relation recognition. These data are extracted from a bilingual sentence-aligned corpus according to the implicit/explicit mismatch between different languages. Incorporating these data via a multi-task neural network model achieves significant improvements over baselines, on both the English PDTB and Chinese CDTB data sets.

## 1 Introduction

Discovering the discourse relation between two sentences is crucial to understanding the meaning of a coherent text, and also beneficial to many downstream NLP applications, such as question answering and machine translation. Implicit discourse relation recognition ($DRR_{imp}$) remains a challenging task due to the absence of strong surface clues like discourse connectives (e.g. $but$). Most work resorts to large amounts of manually designed features (Soricut and Marcu, 2003; Pitler et al., 2009; Lin et al., 2009; Louis et al., 2010; Rutherford and Xue, 2014), or distributed features learned via neural network models (Braud and Denis, 2015; Zhang et al., 2015; Ji and Eisenstein, 2015). The above methods usually suffer from limited labeled data.

Marcu and Echihabi (2002) attempt to create labeled implicit data automatically by removing connectives from explicit instances, as additional training data. These data are usually called as *synthetic implicit data* (hereafter $SynData$). However, Sporleder and Lascarides (2008) argue that $SynData$ has two drawbacks: 1) meaning shifts in some cases when removing connectives, and 2) a different word distribution with the *real implicit data*. They also show that using $SynData$ directly degrades the performance. Recent work seeks to derive valuable information from $SynData$ while filtering noise, via domain adaptation (Braud and Denis, 2014; Ji et al., 2015), classifying connectives (Rutherford and Xue, 2015) or multi-task learning (Lan et al., 2013; Liu et al., 2016), and shows promising results.

*ch:* [社会 认为 有 青少年 问题,]$_{Arg1}$
society reckon existence youth problems,
**implicit=但是** [很多 青少年 认为 自己 没 问题.]$_{Arg2}$
but many young people think themselves no problems.

*en:* [society reckons the existence of youth problems, ]$_{Arg1}$
**but** [many young people do not think there is anything wrong with them.]$_{Arg2}$

**Figure 1:** An example illustrating the implicit/explicit mismatch between Chinese ($ch$) and English ($en$). A Chinese implicit instance is translated into an English explicit one. In the PDTB, a discourse instance is defined as a connective (e.g. $but$) taking two arguments ($Arg1$ and $Arg2$).

Different from previous work, we propose to construct *bilingually-constrained synthetic implicit data* (called $BiSynData$) for $DRR_{imp}$, which can alleviate the drawbacks of $SynData$. Our method is inspired by the findings that a discourse instance expressed implicitly in one language may be expressed explicitly in another. For example, Zhou and Xue

*Corresponding author.



*Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2306–2312,
Austin, Texas, November 1-5, 2016. 

(2012) show that the connectives in Chinese omit much more frequently than those in English with about 82.0% vs. 54.5%. Li et al. (2014a) further argue that there are about 23.3% implicit/explicit mismatchs between Chinese/English instances. As illustrated in Figure 1, a Chinese implicit instance where the connective 但是 is absent, is translated into an English explicit one with the connective $but$. Intuitively, the Chinese instance is a *real* implicit one which can be signaled by $but$. Hence, it could potentially serve as additional training data for the Chinese $DRR_{imp}$, avoiding the different word distribution problem of $SynData$. Meanwhile, for the English explicit instance, it is very likely that removing $but$ would not lose any information since its Chinese counterpart 但是 can be omitted. Therefore it could be used for the English $DRR_{imp}$, alleviating the meaning shift problem of $SynData$.

We extract our $BiSynData$ from a Chinese-English sentence-aligned corpus (Section 2). Then we design a multi-task neural network model to incorporate the $BiSynData$ (Section 3). Experimental results, on both the English PDTB (Prasad et al., 2008) and Chinese CDTB (Li et al., 2014b), show that $BiSynData$ is more effective than $SynData$ used in previous work (Section 4). Finally, we review the related work (Section 5) and draw conclusions (Section 6).

## 2 BiSynData

Formally, given a Chinese-English sentence pair $(S_{ch}, S_{en})$, we try to find an English explicit instance $(Arg1_{en}, Arg2_{en}, Conn_{en})$ in $S_{en}$[1], and a Chinese implicit instance $(Arg1_{ch}, Arg2_{ch})$ in $S_{ch}$, where $(Arg1_{en}, Arg2_{en}, Conn_{en})$ is the translation of $(Arg1_{ch}, Arg2_{ch})$. In most cases, discourse relations should be preserved during translating, so the connective $Conn_{en}$ is potentially a strong indicator of the discourse relation between not only $Arg1_{en}$ and $Arg2_{en}$, but also $Arg1_{ch}$ and $Arg2_{ch}$. Therefore, we can construct two synthetic implicit instances labeled by $Conn_{en}$, denoted as $\langle (Arg1_{en}, Arg2_{en}), Conn_{en} \rangle$ and $\langle (Arg1_{ch}, Arg2_{ch}), Conn_{en} \rangle$, respectively. We refer to these synthetic instances as $BiSynData$ because they are constructed according to the bilingual implicit/explicit mismatch.

| Conn. | Freq. | Conn. | Freq. |
|---|---|---|---|
| $and$ | 14294 | $while$ | 1031 |
| $if$ | 2580 | $before$ | 822 |
| $as$ | 1951 | $also$ | 552 |
| $when$ | 1521 | $since$ | 511 |
| $but$ | 1122 | $because$ | 503 |

**Table 1:** Top 10 most frequent connectives in our $BiSynData$.

In our experiments, we extract our $BiSynData$ from a combined corpus (FBIS and HongKong Law), with about 2.38 million Chinese-English sentence pairs. We generate 30,032 synthetic English instances and the same number of Chinese instances, with 80 connectives, as our $BiSynData$. Table 1 lists the top 10 most frequent connectives in our $BiSynData$, which are roughly consistent with the statistics of Chinese/English implicit/explicit mismatches in (Li et al., 2014a). According to connectives and their related relations in the PDTB, in most cases, $and$ and $also$ indicate the $Expansion$ relation, $if$ and $because$ the $Contigency$ relation, $before$ the $Temporal$ relation, and $but$ the $Comparison$ relation. Connectives $as$, $when$, $while$ and $since$ are ambiguous. For example, $while$ can indicate the $Comparison$ or $Temporal$ relation. Overall, our constructed $BiSynData$ covers all four main discourse relations defined in the PDTB.

With our $BiSynData$, we define two connective classification tasks: 1) given $(Arg1_{en}, Arg2_{en})$ to predict the connective $Conn_{en}$, and 2) given $(Arg1_{ch}, Arg2_{ch})$ to predict $Conn_{en}$. We incorporate the first task to help the English $DRR_{imp}$, and the second for the Chinese $DRR_{imp}$. It is worthy to note that we use English connectives themselves as classification labels rather than mapping them to relations in both tasks.

## 3 Multi-Task Neural Network Model

We design a Multi-task Neural Network Model (denoted as $MTN$), which incorporates a connective classification task on $BiSynData$ (auxiliary task) to benefit $DRR_{imp}$ (main task). In general, the more related two tasks are, the more powerful a multi-task learning method will be. In the current problem, the

[1]In our experiments, we use the pdtb-parser toolkit (Lin et al., 2014) to identify English explicit instances.

two tasks are essentially the same, just with different output labels. Therefore, as illustrated in Figure 2, $MTN$ shares parameters in all feature layers ($L_1$-$L_3$) and uses two separate classifiers in the classifier layer ($L_4$). For each task, given an instance $(Arg_1, Arg_2)$, $MTN$ simply averages embeddings of words to represent arguments, as $v_{Arg_1}$ and $v_{Arg_2}$. These two vectors are then concatenated and transformed through two non-linear hidden layers. Finally, the corresponding $softmax$ layer is used to perform classification.

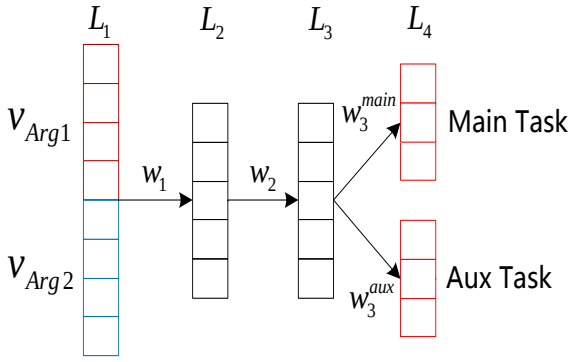


**Figure 2:** $MTN$ with four layers $L_1$-$L_4$.

$MTN$ ignores the word order in arguments and uses two hidden layers to capture the interactions between two arguments. The idea behind $MTN$ is borrowed from (Iyyer et al., 2015), where a deep averaging network achieves close to the state-of-the-art performance on text classification. Though $MTN$ is simple, it is easy to train and efficient on both memory and computational cost. In addition, the simplicity of $MTN$ allows us to focus on measuring the quality of $BiSynData$.

We use the cross-entropy loss function and mini-batch AdaGrad (Duchi et al., 2011) to optimize parameters. Pre-trained word embeddings are fixed. We find that fine-tuning word embeddings during training leads to severe overfitting in our experiments. Following Liu et al. (2016), we alternately use two tasks to train the model, one task per epoch. For tasks on both the PDTB and CDTB, we use the same hyper-parameters. The dimension of word embedding is 100. We set the size of $L_2$ to 200, and $L_3$ to 100. $ReLU$ is used as the non-linear function. Different learning rates 0.005 and 0.001 are used in the main and auxiliary tasks, respectively. To avoid overfitting, we randomly drop out 20% words in each argument following Iyyer et al. (2015). All hyper-parameters are tuned on the development set.

## 4 Experiments

We evaluate our method on both the English PDTB and Chinese CDTB data sets. We tokenize English data and segment Chinese data using the Stanford CoreNLP toolkit (Manning et al., 2014). The English/Chinese Gigaword corpus (3rd edition) is used to train the English/Chinese word embeddings via $word2vec$ (Mikolov et al., 2013), respectively. Due to the skewed class distribution of test data (see Section 4.1), we use the macro-averaged $F_1$ for performance evaluation.

### 4.1 On the PDTB

Following Rutherford and Xue (2015), we perform a 4-way classification on the top-level discourse relations: $Temporal$ ($Temp$), $Comparison$ ($Comp$), $Contigency$ ($Cont$) and $Expansion$ ($Expa$). Sections 2-20 are used as training set, sections 0-1 as development set and sections 21-22 as test set. The training/test set contains 582/55 instances for $Temp$, 1855/145 for $Comp$, 3235/273 for $Cont$ and 6673/538 for $Expa$. The top 20 most frequent connectives in our $BiSynData$ are considered in the auxiliary task, with 28,013 synthetic English instances in total.

| | | $STN$ | $MTN_{bi}$ |
|---|---|---|---|
| $Temp$ | $P$ | 33.33 | **34.48** |
| | $R$ | 14.55 | **18.18** |
| | $F_1$ | 20.25 | **23.81** |
| $Comp$ | $P$ | 38.54 | **42.11** |
| | $R$ | 25.52 | **33.10** |
| | $F_1$ | 30.71 | **37.07** |
| $Cont$ | $P$ | 38.36 | **44.22** |
| | $R$ | **41.03** | 40.66 |
| | $F_1$ | 39.65 | **42.37** |
| $Expa$ | $P$ | 59.60 | **62.56** |
| | $R$ | 66.36 | **71.75** |
| | $F_1$ | 62.80 | **66.84** |
| $macro\ F_1$ | | 38.35 | **42.52** |

**Table 2:** Results of 4-way classification on the PDTB.

Table 2 shows the results of $MTN$ combining our $BiSynData$ (denoted as $MTN_{bi}$) on the PDTB.

$STN$ means we train $MTN$ with only the main task. On the $macro\ F_1$, $MTN_{bi}$ gains an improvement of 4.17% over $STN$. The improvement is significant under one-tailed t-test ($p$<0.05). A closer look into the results shows that $MTN_{bi}$ performs better across all relations, on the precision, recall and $F_1$ score, except a little drop on the recall of $Cont$. The reason for the recall drop of $Cont$ is not clear. The greatest improvement is observed on $Comp$, up to 6.36% $F_1$ score. The possible reason is that only $while$ is ambiguous about $Comp$ and $Temp$, while $as$, $when$ and $since$ are all ambiguous about $Temp$ and $Cont$, among top 10 connectives in our $BiSynData$. Meanwhile the amount of labeled data for $Comp$ is relatively small. Overall, using $BiSynData$ under our multi-task model achieves significant improvements on the English $DRR_{imp}$. We believe the reasons for the improvements are twofold: 1) the added synthetic English instances from our $BiSynData$ can alleviate the meaning shift problem, and 2) a multi-task learning method is helpful for addressing the different word distribution problem between implicit and explicit data.

Considering some of the English connectives (e.g., $while$) are highly ambiguous, we compare our method with ones that uses only unambiguous connectives. Specifically, we first discard $as$, $when$, $while$ and $since$ in top 20 connectives, and get 22,999 synthetic instances. Then, we leverage these instances in two different ways: 1) using them in our multi-task model as above, and 2) using them as additional training data directly after mapping unambiguous connectives into relations. Both methods using only unambiguous connectives do not achieve better performance. One possible reason is that these synthetic instances become more unbalanced after discarding ones with ambiguous connectives.

We also compare $MTN_{bi}$ with recent systems using additional training data. Rutherford and Xue (2015) select explicit instances that are similar to the implicit ones via connective classification, to enrich the training data. Liu et al. (2016) use a multi-task model with three auxiliary tasks: 1) $conn$: connective classification on explicit instances, 2) $exp$: relation classification on the labeled explicit instances in the PDTB, and 3) $rst$: relation classification on the labeled RST corpus (William and Thompson, 1988), which defines different discourse relations with that in the PDTB. The results are shown in Table 3. Although Liu et al. (2016) achieve the state-of-the-art performance (Line 5), they use two additional labeled corpora. We can find that $MTN_{bi}$ (Line 6) yields better results than those systems incorporating $SynData$ (Line 1, 2 and 3), or even the labeled RST (Line 4). These results confirm that $BiSynData$ can indeed alleviate the disadvantages of $SynData$ effectively.

| | System | $macro\ F_1$ |
|---|---|---|
| 1 | Rutherford and Xue (2015) | 40.50 |
| 2 | Liu et al. (2016) $_{conn}$ | 38.09 |
| 3 | Liu et al. (2016) $_{exp}$ | 39.03 |
| 4 | Liu et al. (2016) $_{rst}$ | 40.67 |
| 5 | Liu et al. (2016) $_{conn+exp+rst}$ | **44.98** |
| 6 | $MTN_{bi}$ | 42.52 |

**Table 3:** Comparison with recent systems on the PDTB. $conn+exp+rst$ means using three auxiliary tasks simultaneously.

### 4.2 On the CDTB

Four top-level relations are defined in the CDTB, including $Transition$ ($Tran$), $Causality$ ($Caus$), $Explanation$ ($Expl$) and $Coordination$ ($Coor$). We use instances in the first 50 documents as test set, second 50 documents as development set and remaining 400 documents as training set. We conduct a 3-way classification because of only 39 instances for $Tran$. The training/test set contains 682/95 instances for $Caus$, 1143/126 for $Expl$ and 2300/347 for $Coor$. The top 20 most frequent connectives (excluding $and$)[2] in our $BiSynData$ are considered in the auxiliary task, with 13,899 synthetic Chinese instances in total. The results are shown in Table 4. Compared with $STN$, $MTN_{bi}$ raises the $macro\ F_1$ from 55.44% to 58.28%. The improvement is significant under one-tailed t-test ($p$<0.05). Therefore, $BiSynData$ is also helpful for the Chinese $DRR_{imp}$.

Because of no reported results on the CDTB, we use $MTN$ with two different auxiliary tasks as baselines: 1) $exp$: relation classification on the labeled

[2]Including $and$ degrades the performance slightly. A possible reason is that $and$ can be related to both the $Expl$ and $Coor$ relations in the CDTB, and instances marked by $and$ account for about half of our $BiSynData$.

| | | $STN$ | $MTN_{bi}$ |
|---|---|---|---|
| $Caus$ | $P$ | 47.92 | **52.94** |
| | $R$ | 24.21 | **28.42** |
| | $F_1$ | 32.17 | **36.99** |
| $Expl$ | $P$ | **54.62** | 53.47 |
| | $R$ | 56.35 | **61.11** |
| | $F_1$ | 55.47 | **57.04** |
| $Coor$ | $P$ | 74.36 | **78.02** |
| | $R$ | 83.57 | **83.86** |
| | $F_1$ | 78.70 | **80.83** |
| $macro\ F_1$ | | 55.44 | **58.28** |

**Table 4:** Results of 3-way classification on the CDTB.

explicit instances in the CDTB, including 466 instances for $Caus$, 201 for $Expl$ and 974 for $Coor$. 2) $conn$: connective classification on explicit instances from the Xinhua part of the Chinese Gigaword corpus. We collect explicit instances with the top 20 most frequent Chinese connectives and sample 20,000 instances for the experiment. Both $exp$ and $conn$ can be considered as tasks on $SynData$. The results in Table 5 show that $MTN$ incorporating $BiSynData$ (Line 3) performs better than using $SynData$ (Line 1 and 2), for the task on the CDTB.

| | System | $macro\ F_1$ |
|---|---|---|
| 1 | $MTN_{exp}$ | 56.42 |
| 2 | $MTN_{conn}$ | 56.86 |
| 3 | $MTN_{bi}$ | **58.28** |

**Table 5:** $MTN$ with different auxiliary tasks on the CDTB.

## 5 Related Work

One line of research related to $DRR_{imp}$ tries to take advantage of explicit discourse data. Zhou et al. (2010) predict the absent connectives based on a language model. Using these predicted connectives as features is proven to be helpful. Biran and McKeown (2013) aggregate word-pair features that are collected around the same connectives, which can effectively alleviate the feature sparsity problem. More recently, Braud and Denis (2014) and Ji et al. (2015) consider explicit data from a different domain, and use domain adaptation methods to explore the effect of them. Rutherford and Xue (2015) propose to gather weakly labeled data from explicit instances via connective classification, which are used as additional training data directly. Lan et al. (2013) and Liu et al. (2016) combine explicit and implicit data using multi-task learning models and gain improvements. Different from all the above work, we construct additional training data from a bilingual corpus.

Multi-task neural networks have been successfully used for many NLP tasks. For example, Collobert et al. (2011) jointly train models for the Part-of-Speech tagging, chunking, named entity recognition and semantic role labeling using convolutional network. Liu et al. (2015) successfully combine the tasks of query classification and ranking for web search using a deep multi-task neural network. Luong et al. (2016) explore multi-task sequence to sequence learning for constituency parsing, image caption generation and machine translation.

## 6 Conclusion

In this paper, we introduce bilingually-constrained synthetic implicit data ($BiSynData$), which are generated based on the bilingual implicit/explicit mismatch, into implicit discourse relation recognition for the first time. On both the PDTB and CDTB, using $BiSynData$ as the auxiliary task significantly improves the performance of the main task. We also show that $BiSynData$ is more beneficial than the synthetic implicit data typically used in previous work. Since the lack of labeled data is a major challenge for implicit discourse relation classification, our proposed $BiSynData$ can enrich the training data and then benefit future work.

## Acknowledgments

We would like to thank all the reviewers for their constructive and helpful suggestions on this paper. This work is partially supported by the Natural Science Foundation of China (Grant Nos. 61573294, 61303082, 61672440), the Ph.D. Programs Foundation of Ministry of Education of China (Grant No. 20130121110040), the Fund of Research Project of Tibet Autonomous Region of China (Grant No. Z2014A18G2-13), and the Natural Science Foundation of Fujian Province (Grant No. 2016J05161).

## References


Or Biran and Kathleen McKeown. 2013. Aggregated Word Pair Features for Implicit Discourse Relation Disambiguation. In *Proceedings of ACL (Volume 2: Short Papers)*, pages 69–73, Sofia, Bulgaria.

Chloé Braud and Pascal Denis. 2014. Combining Natural and Artificial Examples to Improve Implicit Discourse Relation Identification. In *Proceedings of COLING*, pages 1694–1705, Dublin, Ireland.

Chloé Braud and Pascal Denis. 2015. Comparing Word Representations for Implicit Discourse Relation Classification. In *Proceedings of EMNLP*, pages 2201–2211, Lisbon, Portugal.

Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural Language Processing (Almost) from Scratch. *Journal of Machine Learning Research*, 12(1):2493–2537.

John Duchi, Elad Hazan, and Yoram Singer. 2011. Adaptive Subgradient Methods for Online Learning and Stochastic Optimization. *The Journal of Machine Learning Research*, 12:2121–2159.

Mohit Iyyer, Varun Manjunatha, Jordan Boyd-Graber, and Hal Daumé III. 2015. Deep Unordered Composition Rivals Syntactic Methods for Text Classification. In *Proceedings of ACL-IJCNLP*, pages 1681–1691, Beijing, China.

Yangfeng Ji and Jacob Eisenstein. 2015. One Vector is Not Enough: Entity-Augmented Distributed Semantics for Discourse Relations. In *Transactions of the Association for Computational Linguistics*, volume 3, pages 329–344.

Yangfeng Ji, Gongbo Zhang, and Jacob Eisenstein. 2015. Closing the Gap: Domain Adaptation from Explicit to Implicit Discourse Relations. In *Proceedings of EMNLP*, pages 2219–2224, Lisbon, Portugal.

Man Lan, Yu Xu, and Zhengyu Niu. 2013. Leveraging Synthetic Discourse Data via Multi-task Learning for Implicit Discourse Relation Recognition. In *Proceedings of ACL*, pages 476–485, Sofia, Bulgaria.

Junyi Jessy Li, Marine Carpuat, and Ani Nenkova. 2014a. Cross-lingual Discourse Relation Analysis: A Corpus Study and a Semi-supervised Classification System. In *Proceedings of COLING : Technical Papers*, pages 577–587, Dublin, Ireland.

Yancui Li, Wenhe Feng, Jing Sun, Fang Kong, and Guodong Zhou. 2014b. Building Chinese Discourse Corpus with Connective-driven Dependency Tree Structure. In *Proceedings of EMNLP*, pages 2105–2114, Doha, Qatar.

Ziheng Lin, Min-Yen Kan, and Hwee Tou Ng. 2009. Recognizing Implicit Discourse Relations in the Penn Discourse Treebank. In *Proceedings of EMNLP*, pages 343–351, PA, USA.

Ziheng Lin, Hwee Tou Ng, and Min-Yen Kan. 2014. A PDTB-styled End-to-end Discourse Parser. *Natural Language Engineering*, 20(02):151–184.

Xiaodong Liu, Jianfeng Gao, Xiaodong He, Li Deng, Kevin Duh, and Ye-Yi Wang. 2015. Representation Learning Using Multi-task Deep Neural Networks for Semantic Classification and Information Retrieval. In *Proceedings of NAACL*, pages 912–921, Denver, Colorado.

Yang Liu, Sujian Li, Xiaodong Zhang, and Zhifang Sui. 2016. Implicit Discourse Relation Classification via Multi-Task Neural Networks. In *Proceedings of AAAI*, pages 2750–2756, Arizona, USA.

Annie Louis, Aravind Joshi, Rashmi Prasad, and Ani Nenkova. 2010. Using Entity Features to Classify Implicit Discourse Relations. In *Proceedings of SIGDIAL*, pages 59–62, PA, USA.

Minh-Thang Luong, Quoc V. Le, Ilya Sutskever, Oriol Vinyals, and Lukasz Kaiser. 2016. Multi-task Sequence to Sequence Learning. In *Proceedings of ICLR*, pages 1–10, San Juan, Puerto Rico.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP Natural Language Processing Toolkit. In *Proceedings of ACL (System Demonstrations)*, pages 55–60, Maryland, USA.

Daniel Marcu and Abdessamad Echihabi. 2002. An Unsupervised Approach to Recognizing Discourse Relations. In *Proceedings of ACL*, pages 368–375, PA, USA.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. *CoRR*, abs/1301.3781.

Emily Pitler, Annie Louis, and Ani Nenkova. 2009. Automatic Sense Prediction for Implicit Discourse Relations in Text. In *Proceedings of ACL-IJCNLP*, pages 683–691, PA, USA.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. 2008. The Penn Discourse TreeBank 2.0. In *Proceedings of LREC*, volume 24, pages 2961–2968, Marrakech, Morocco.

Attapol T. Rutherford and Nianwen Xue. 2014. Discovering Implicit Discourse Relations Through Brown Cluster Pair Representation and Coreference Patterns. In *Proceedings of EACL*, pages 645–654, Gothenburg, Sweden.

Attapol Rutherford and Nianwen Xue. 2015. Improving the Inference of Implicit Discourse Relations via Classifying Explicit Discourse Connectives. In *Proceedings of NAACL*, pages 799–808, Denver, Colorado.

Radu Soricut and Daniel Marcu. 2003. Sentence Level Discourse Parsing Using Syntactic and Lexical Information. In *Proceedings of NAACL*, pages 149–156, PA, USA.

Caroline Sporleder and Alex Lascarides. 2008. Using Automatically Labelled Examples to Classify Rhetorical Relations: An Assessment. *Natural Language Engineering*, 14(3):369–416.

Mann William and Sandra Thompson. 1988. Rhetorical structure theory: Towards a Functional Theory of Text Organization. *Text*, 8(3):243–281.

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 EMNLP*, pages 2230–2235, Lisbon, Portugal.

Yuping Zhou and Nianwen Xue. 2012. PDTB-style discourse annotation of Chinese text. In *Proceedings of ACL*, pages 69–77, Jeju Island, Korea.

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 COLING*, pages 1507–1514, PA, USA.