# Investigating the Impact of ASR Errors on Spoken Implicit Discourse Relation Recognition

**Linh The Nguyen** and **Dat Quoc Nguyen**
VinAI Research, Hanoi, Vietnam
{v.linhnt140, v.datnq9}@vinai.io

## Abstract

We present an empirical study investigating the influence of automatic speech recognition (ASR) errors on the spoken implicit discourse relation recognition (IDRR) task. We construct a spoken dataset for this task based on the Penn Discourse Treebank 2.0 (Prasad et al., 2008). On this dataset, we conduct "Cascaded" experiments employing state-of-the-art ASR and text-based IDRR models and find that the ASR errors significantly decrease the IDRR performance. In addition, the "Cascaded" approach does remarkably better than an "End-to-End" one that directly predicts a relation label for each input argument speech pair.

## 1 Introduction

Discourse parsing is one of the key research areas in NLP (Marcu, 2000; Li et al., 2022). One important problem in discourse parsing is the implicit discourse relation recognition (IDRR) task (Marcu and Echihabi, 2002), which aims to identify the relation between two discourse arguments (e.g. clauses, sentences or paragraphs in the document) without explicit discourse connectives (e.g., *but*, *and*, *because* and the like). This IDRR task has attracted many research works (Lin et al., 2009; Zhou et al., 2010; Ji and Eisenstein, 2015; Bai and Zhao, 2018; Nguyen et al., 2019; Kim et al., 2020; Dou et al., 2021; Jiang et al., 2021), and it is very useful for many downstream NLP tasks such as machine translation (Joty et al., 2017; Guzmán et al., 2014), text summarization (Li and Rafi, 2019; Gerani et al., 2014) and question answering (Chai and Jin, 2004; Jansen et al., 2014).

Implicit discourse relations also play essential roles in spoken language understanding tasks (Aubin et al., 2019; Ma et al., 2019). Thus, it is worth investigating the IDRR task in spoken form. Research works have been performed for IDRR from the manual speech transcripts (Pettibone and Pon-Barry, 2003; Tonelli et al., 2010; Rehbein et al., 2016). However, to the best of our knowledge, no study has investigated the effect of automatic speech recognition (ASR) errors on the spoken IDRR task. Table 1 shows an example of ASR errors that might affect the IDRR result.

| | |
|---|---|
| **original** | Argument 1: computer-generated videos help<br>Argument 2: the average american watches seven hours of tv a day |
| **transcript** | Argument 1: computer generated **vidio's health**<br>Argument 2: the average american watches seven hours of **tevia** day |

Table 1: An example of ASR errors (highlighted in bold). A prediction model needs to identify the discourse relation "Contingency.Cause.Reason" between Argument 1 and Argument 2, without the discourse marker (here, "since"), which is already challenging. It would be more challenging if the model is required to work on transcript data with potential ASR errors which might change the meanings of input arguments.

In this paper, we present a study that investigates the influence of ASR errors on the downstream spoken IDRR task. As there is no public benchmark dataset for this spoken IDRR task, we construct a dataset for this task based on the Penn Discourse Treebank (PDTB) 2.0 (Prasad et al., 2008). Following previous works (Lee et al., 2018; You et al., 2020; Song et al., 2022) that construct spoken derivatives of text-based question answering and text-to-SQL datasets, we use the Google text-to-speech system to produce a spoken variant of the PDTB 2.0 dataset. In our "Cascaded" experiments combining state-of-the-art ASR and text-based IDRR models, we find that the ASR errors significantly decrease the performance of the downstream IDRR task. We also experiment with an "End-to-End" approach that directly predicts a relation label for each input argument speech pair, and find that the "End-to-End" obtains remarkably lower performances than the "Cascaded".

| Statistic | #Pair | #Hour | WER |
|---|---|---|---|
| **Training** | 12632 | 58.37 | 28.42 |
| **Validation** | 1183 | 5.42 | 27.28 |
| **Test** | 1046 | 4.59 | 30.27 |

Table 2: Our dataset statistics. "#Pair", "#Hour" and "WER" denote the number of spoken pairs, the number of speech audio hours and the word error rate, respectively. Here the word error rate is computed for the automatic transcripts predicted by Wav2Vec 2.0 w.r.t. the original text arguments.

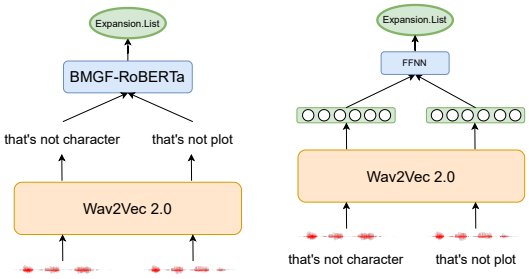


Figure 1: Illustrations of our empirical approaches: "Cascaded" in the left-hand side subfigure and "End-to-End" in the right-hand side subfigure.

## 2 Dataset construction

This section presents the dataset construction process for our spoken IDRR task. We construct our dataset in the spoken form based on the PDTB 2.0 dataset (Prasad et al., 2008), which is one of the largest benchmark datasets used for IDRR research. We employ the Google text-to-speech system to generate spoken variants of the original text arguments from the PDTB 2.0 dataset. We thus obtain speech pairs and the gold relation label for each speech pair (i.e. the label of the original argument pair). We also employ the standard PDTB 2.0 data split (Ji and Eisenstein, 2015) that uses sections 2–20, 0–1 and 21–22 for training, validation and test, respectively. Table 2 shows the statistics of our dataset.

## 3 Empirical approach

On our spoken dataset, we compare two implicit discourse relation recognition approaches: *Cascaded* vs. *End-to-End*.

### 3.1 Cascaded

The "Cascaded" approach combines two main components of automatic speech recognition (ASR) and text-based IDRR, as illustrated in Figure 1.

For the ASR component, we employ the base version of Wav2Vec 2.0 (Baevski et al., 2020)—which is pre-trained and fine-tuned on the 960-hour Librispeech dataset (Panayotov et al., 2015). In particular, we feed the spoken argument audios into Wav2Vec 2.0 to generate the corresponding automatic speech recognition (ASR) transcripts. For each argument speech pair, we thus obtain a corresponding transcript pair generated by Wav2Vec 2.0. Table 1 shows an example of ASR transcription errors from our training set. Table 2 also presents the word error rate of Wav2Vec 2.0 on our dataset.

The text-based IDRR component takes each speech transcript pair produced by the ASR component as input and predicts the discourse relation label for the transcript pair. For IDRR, we employ BMGF-RoBERTa (Liu et al., 2020) with its officially public implementation, which still maintains its state-of-the-art performance level up to date on the PDTB 2.0 dataset. BMGF-RoBERTa employs RoBERTa (Liu et al., 2019) to obtain contextualized representations for word tokens in each argument and also uses the following modules:

- Trainable segment embeddings (**SE**): the trainable segment embeddings are originally used in BERT (Devlin et al., 2019), but removed in RoBERTa. BMGF-RoBERTa employs these embeddings because they are shown to be helpful for the IDRR task (Shi and Demberg, 2019).
- Bilateral Matching (**BM**): comparing each word token of one argument against all tokens of the other one and vice versa.
- Gated Fusion (**GF**): assigning different importance to each word token in arguments, and then aggregating importance results and encoding each argument into a vector representation.
- Prediction: Two arguments' vectors are concatenated into a single one that is fed into a two-layer feed-forward neural network (FFNN) followed by a `softmax` for relation classification.

### 3.2 End-to-End

For the "End-to-End" approach, we propose a speech-based discourse identification model that takes each argument speech pair as input and directly predicts the relation label for the input speech pair. In particular, the model employs Wav2Vec 2.0 to extract a feature vector representation from each speech. The model uses a similar prediction

layer as in BMGF-RoBERTa, which concatenates two audios' vectors into a single vector and then feeds this vector into a two-layer FFNN followed by a `softmax` for relation classification. Figure 1 also illustrates the "End-to-End" architecture.

### 3.3 Implementation details and Setup

For the "Cascaded" approach, we train BMGF-RoBERTa for 40 epochs on the speech transcript pairs from the training set. We employ optimal hyper-parameters from Liu et al. (2020), which are 0.001, 32 and 0.005 for the Adam learning rate, the batch size and the weight decay, respectively. In each training epoch, we compute the model's accuracy two times on the validation set of speech transcript pairs to select the best checkpoint. The selected checkpoint is then applied to the test set of speech transcript pairs to report final results.

For the "End-to-End", Wav2Vec 2.0 is employed as a feature extractor, frozen during training, while the remaining prediction layer is learned. We train the proposed model for 10 epochs on the speech pairs from the training set, using the Adam learning rate grid-searched at 1e-5 with a batch size of 1 (as the audios are long) and 8 gradient accumulation steps. We evaluate the model two times on the validation set of speech pairs in each training epoch, to select the best checkpoint to apply to the test set.

Note that PDTB 2.0 has a hierarchical annotation scheme of 3 implicit relation levels. Most works using PDTB 2.0 report *accuracy* (Acc.) and *macro-averaged F1* scores for the classification of all **4** labels from the top level (**L1**), including Comparison (*Comp.*), Contingency (*Cont.*), Expansion (*Exp.*) and Temporal (*Temp.*). Recent works (Ji and Eisenstein, 2015; Bai and Zhao, 2018; Dai and Huang, 2019; Shi and Demberg, 2019; Liu et al., 2020) additionally report *accuracy* (Acc.) scores for the classification of the top **11** frequent labels from the second level (**L2**). We follow the recent works to report obtained results on both setups.

## 4 Experimental results

### 4.1 Main results

Table 3 reports multi-class classification results obtained on the test set at the top (L1) and second (L2) levels. When it comes to the effect of ASR errors propagation, all performance scores are significantly decreased: 69.06% → 66.63% and 58.13% → 50.24%, which are classification accuracies for the top- and second-level labels, respectively; and 63.39% → 56.00%, which are $F_1$ scores for the top-level label prediction. Table 4 shows the one-vs-rest binary classification F1 score for each label from the top level. ASR errors also remarkably reduce the performance. In particular, scores are decreased about 3% on the Expansion (77.66% → 74.15%), Comparison (59.44% → 56.78%) and Contingency (60.98% → 57.28%) labels, and about 7% on the Temporal label (50.26% → 43.64%).

| Model | 4-way L1 (Acc. \| F1) | 11-way L2 (Acc.) |
|---|---|---|
| **Liu et al.** | 69.06 \| 63.39 | 58.13 |
| **Cascaded** | 66.63 \| 56.00 | 50.24 |
| **End-to-End** | 51.34 \| 38.29 | 37.92 |

Table 3: Multi-class classification results (in %) on the test set. "Liu et al." denotes results of BMGF-RoBERTa with the original text arguments as its input (i.e. equivalent to a perfect ASR of 0% WER). Each score difference between two models is significant with p-value < 0.01.

| Model | Exp. | Comp. | Cont. | Temp. |
|---|---|---|---|---|
| **Liu et al.** | 77.66 | 59.44 | 60.98 | 50.26 |
| **Cascaded** | 74.15 | 56.78 | 57.28 | 43.64 |
| **End-to-End** | 58.15 | 38.39 | 37.64 | 28.32 |

Table 4: Binary classification F1 score (in %) for each L1 label on the test set. Each score difference between two models is significant with p-value < 0.01.

Tables 3 and 4 also show that the performance of the "End-to-End" approach is far behind the "Cascaded" one's. For example, the accuracy and F1 scores obtained for "End-to-End" on the top-level labels are about 15+% lower than those of "Cascaded". This is not surprising because: (1) our speech dataset is small for this difficult language understanding task of spoken IDRR, and (2) the "Cascaded" approach gets to utilize the powerful pre-trained RoBERTa model while the "End-to-End" one is limited to a simple two-layer FFNN.

### 4.2 Ablation study

We conduct an ablation study to investigate the contribution of each main module of the BMGF-RoBERTa model to the final results of the "Cascaded" approach. Table 5 shows the results obtained on the validation set. Each of the main modules, including the trainable segment embeddings, the Bilateral Matching and Gated Fusion, plays an essential role in BMGF-RoBERTa (See Section 3.1 for brief descriptions of these modules). Removing

| Model | 4-way L1 (Acc. \| F1) | 11-way L2 (Acc.) | Exp. | Comp. | Cont. | Temp. |
|---|---|---|---|---|---|---|
| **Cascaded** | **68.13 \| 58.16** | **54.59** | **77.63** | **58.33** | **57.23** | **40.26** |
| **(1)** w/o SE | 62.64 \| 49.09 | 47.04 | 75.07 | 43.69 | 54.02 | 35.68 |
| **(2)** w/o BM | 66.27 \| 57.66 | 51.59 | 76.46 | 52.80 | 54.96 | 38.98 |
| **(3)** w/o GF | 66.53 \| 55.95 | 51.93 | 75.98 | 55.24 | 55.76 | 33.66 |
| **(1) & (2) & (3)** | 59.59 \| 49.02 | 43.00 | 73.77 | 43.41 | 47.91 | 29.91 |

Table 5: Ablation results on the validation set. **(1)** w/o SE: Without employing the trainable segment embeddings; **(2)** w/o BM: Without the Bilateral Matching module; **(3)** w/o GF: Without the Gated Fusion module. Each score difference between the full cascaded model and its ablated one is significant with p-value < 0.01.

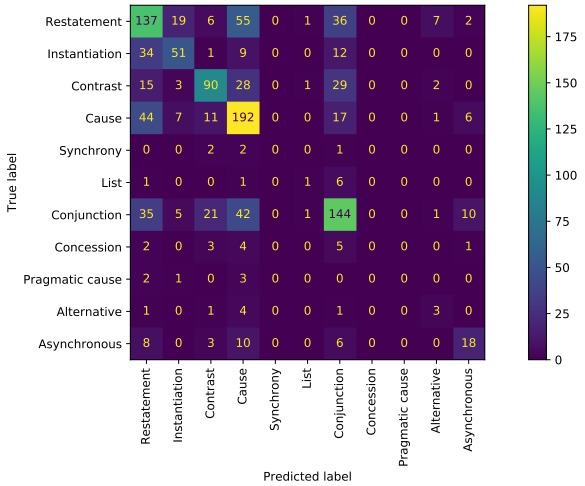


Figure 2: The confusion matrix of the "Cascaded" approach on the validation set w.r.t. the top 11 frequent labels from the second level.

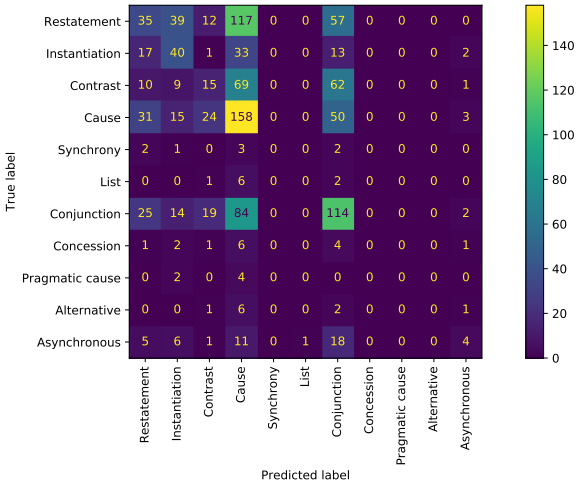


Figure 3: The confusion matrix of the "End-to-End" approach on the validation set w.r.t. the top 11 frequent labels from the second level.

each module significantly reduces the performance. In addition, removing all three modules degrades the obtained results by about 10+% in most cases.

### 4.3 Error analysis

Figure 2 presents the confusion matrix of the "Cascaded" approach on the validation set w.r.t. multi-class classification of the top 11 frequent labels from the second level. We find that correct predictions mainly come from 6 major labels of *Cause*, *Conjunction*, *Restatement*, *Contrast*, *Instantiation* and *Asynchronous*. We also find that main errors come from the confusion between the relations *Restatement* and *Cause*, the relations *Conjunction* and *Cause* and the relations *Contrast* and *Conjunction*. They are difficult to distinguish because the form of the discourse unit in the two relation labels is semantically similar. We observe similar findings for the "End-to-End" as shown in Figure 3.

We provide a qualitative example to demonstrate the challenges of this spoken IDRR task. Given an input speech pair of the original text argument pair ("*After the race, Fortune 500 executives* ***drooled*** *like schoolboys over the cars and drivers*", "*No dummies, the drivers pointed out they still had space on their machines for another sponsor's name or two*"), in the "Cascaded" approach, the original token "**drooled**" from the first argument is incorrectly predicted as **druled** by the ASR component. Both the "Cascaded" and "End-to-End" approaches produce an incorrect label prediction of *Contrast*, while BMGF-RoBERTa takes this original text argument pair as input and produces a correct label of *Cause*.

## 5 Discussion

The method of employing the Google text-to-speech to generate spoken forms of the original text arguments in the PDTB 2.0 dataset produces an artificially generated dataset, thus not fully reflecting

the error types of human speech. In addition, the original raw PDTB 2.0 dataset comes from the Wall Street Journal (WSJ) articles. So our dataset might not cover relevant real-world spoken genres.

We unfortunately were unaware of the availability of the Continuous Speech Recognition (CSR) corpus that consists of human-read speech with texts from the WSJ when conducting our study.[1] There might be an overlap between original texts from the PDTB 2.0 dataset and the CSR corpus, thus the overlap might be used for further evaluation in future work.

## 6 Conclusion

We have presented an empirical study investigating the influence of ASR errors on the spoken IDRR task. We construct a spoken derivative of the PDTB 2.0 dataset and conduct "Cascaded" experiments employing state-of-the-art ASR and text-based IDRR models on this spoken dataset. We find that the ASR errors significantly reduce the IDRR performance. We also find that an "End-to-End" approach that directly predicts a relation label for each input speech pair obtains remarkably lower performances than the "Cascaded" one.

## References


Adèle Aubin, Alessandra Cervone, Oliver Watts, and Simon King. 2019. Improving Speech Synthesis with Discourse Relations. In *Proceedings of INTERSPEECH*, pages 4470–4474.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Proceedings of NeurIPS*, pages 12449–12460.

Hongxiao Bai and Hai Zhao. 2018. Deep Enhanced Representation for Implicit Discourse Relation Recognition. In *Proceedings of COLING*, pages 571–583.

Joyce Y. Chai and Rong Jin. 2004. Discourse Structure for Context Question Answering. In *Proceedings of the Workshop on Pragmatics of Question Answering at HLT-NAACL 2004*, pages 23–30.

Zeyu Dai and Ruihong Huang. 2019. A Regularization Approach for Incorporating Event Knowledge and Coreference Relations into Neural Discourse Parsing. In *Proceedings of EMNLP-IJCNLP*, pages 2976–2987.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of NAACL*, pages 4171–4186.

Zujun Dou, Yu Hong, Yu Sun, and Guodong Zhou. 2021. CVAE-based Re-anchoring for Implicit Discourse Relation Classification. In *Findings of EMNLP*, pages 1275–1283.

Shima Gerani, Yashar Mehdad, Giuseppe Carenini, Raymond T. Ng, and Bita Nejat. 2014. Abstractive Summarization of Product Reviews Using Discourse Structure. In *Proceedings of EMNLP*, pages 1602–1613.

Francisco Guzmán, Shafiq Joty, Lluís Màrquez, and Preslav Nakov. 2014. Using Discourse Structure Improves Machine Translation Evaluation. In *Proceedings of ACL*, pages 687–698.

Peter Jansen, Mihai Surdeanu, and Peter Clark. 2014. Discourse Complements Lexical Semantics for Non-factoid Answer Reranking. In *Proceedings of ACL*, pages 977–986.

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

Feng Jiang, Yaxin Fan, Xiaomin Chu, Peifeng Li, and Qiaoming Zhu. 2021. Not Just Classification: Recognizing Implicit Discourse Relation on Joint Modeling of Classification and Generation. In *Proceedings of EMNLP*, pages 2418–2431.

Shafiq Joty, Francisco Guzmán, Lluís Màrquez, and Preslav Nakov. 2017. Discourse Structure in Machine Translation Evaluation. *Computational Linguistics*, 43:683–722.

Najoung Kim, Song Feng, Chulaka Gunasekara, and Luis Lastras. 2020. Implicit Discourse Relation Classification: We Need to Talk about Evaluation. In *Proceedings of ACL*, pages 5404–5414.

Chia-Hsuan Lee, Szu-Lin Wu, Chi-Liang Liu, and Hung yi Lee. 2018. Spoken SQuAD: A Study of Mitigating the Impact of Speech Recognition Errors on Listening Comprehension. In *Proceedings of INTERSPEECH*, pages 3459–3463.

J. Li and M. Rafi. 2019. Utilize Discourse Relations to Segment Document for Effective Summarization. In *Proceedings of SKG*, pages 12–15.

Jiaqi Li, Ming Liu, Bing Qin, and Ting Liu. 2022. A survey of discourse parsing. *Frontiers of Computer Science*, 16(5).

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.


[1] https://catalog.ldc.upenn.edu/LDC94S13A

Xin Liu, Jiefu Ou, Yangqiu Song, and Xin Jiang. 2020. On the Importance of Word and Sentence Representation Learning in Implicit Discourse Relation Classification. In *Proceedings of IJCAI*, pages 3830–3836.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv preprint*, arXiv:1907.11692.

Mingyu Derek Ma, Kevin Bowden, Jiaqi Wu, Wen Cui, and Marilyn Walker. 2019. Implicit Discourse Relation Identification for Open-domain Dialogues. In *Proceedings of ACL*, pages 666–672.

Daniel Marcu. 2000. *The Theory and Practice of Discourse Parsing and Summarization*. MIT Press.

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

Linh The Nguyen, Linh Van Ngo, Khoat Than, and Thien Huu Nguyen. 2019. Employing the Correspondence of Relations and Connectives to Identify Implicit Discourse Relations via Label Embeddings. In *Proceedings of ACL*, pages 4201–4207.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: An ASR corpus based on public domain audio books. In *Proceedings of ICASSP*, pages 5206–5210.

Jeanette Pettibone and Heather Pon-Barry. 2003. A Maximum Entropy Approach to Recognizing Discourse Relations in Spoken Language. Technical report, Stanford University.

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*.

Ines Rehbein, Merel Scholman, and Vera Demberg. 2016. Annotating Discourse Relations in Spoken Language: A Comparison of the PDTB and CCR Frameworks. In *Proceedings of LREC*, pages 1039–1046.

Wei Shi and Vera Demberg. 2019. Next Sentence Prediction helps Implicit Discourse Relation Classification within and across Domains. In *Proceedings of EMNLP-IJCNLP*, pages 5790–5796.

Yuanfeng Song, Raymond Chi-Wing Wong, Xuefang Zhao, and Di Jiang. 2022. Speech-to-SQL: Towards Speech-driven SQL Query Generation From Natural Language Question. *ArXiv preprint*, arxiv:2201.01209.

Sara Tonelli, Giuseppe Riccardi, Rashmi Prasad, and Aravind Joshi. 2010. Annotation of Discourse Relations for Conversational Spoken Dialogs. In *Proceedings of LREC*.

Chenyu You, Nuo Chen, Fenglin Liu, Dongchao Yang, and Yuexian Zou. 2020. Towards Data Distillation for End-to-end Spoken Conversational Question Answering. *ArXiv preprint*, arxiv:2010.08923.

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: Posters*, pages 1507–1514.