# Entity Enhancement for Implicit Discourse Relation Classification in the Biomedical Domain

**Wei Shi**[§,†] and **Vera Demberg**[†,‡]
[§] Alibaba Group, Hangzhou, China
[†] Dept. of Language Science and Technology
[‡] Dept. of Mathematics and Computer Science, Saarland University
Saarland Informatics Campus, Saarbrücken, Germany
{w.shi, vera}@coli.uni-saarland.de

## Abstract

Implicit discourse relation classification is a challenging task, in particular when the text domain is different from the standard Penn Discourse Treebank (PDTB; Prasad et al., 2008) training corpus domain (Wall Street Journal in 1990s). We here tackle the task of implicit discourse relation classification on the biomedical domain, for which the Biomedical Discourse Relation Bank (BioDRB; Prasad et al., 2011) is available. We show that entity information can be used to improve discourse relational argument representation. In a first step, we show that explicitly marked instances that are content-wise similar to the target relations can be used to achieve good performance in the cross-domain setting using a simple unsupervised voting pipeline. As a further step, we show that with the linked entity information from the first step, a transformer which is augmented with entity-related information (K-BERT; Liu et al., 2020) sets the new state of the art performance on the dataset, outperforming the large pre-trained BioBERT (Lee et al., 2020) model by 2% points.

## 1 Introduction

Discourse relation classification (DRC) involves automatically inferring the logical link between different text segments (such as causal, contrastive, temporal etc.). It has been shown to be a valuable preprocessing step to many downstream natural language processing tasks such as machine translation (Guzmán et al., 2014; Meyer et al., 2015), text summarization (Gerani et al., 2014) and question-answering (Jansen et al., 2014). A main obstacle to a wider usage of automatic DR classifiers however lies in getting the classifiers to work reliably on domains other than the WSJ, that discourse relation parsers are usually trained on PDTB (Prasad et al., 2008) and RST (Carlson et al., 2003).

Moving to a different domain is particularly challenging in DRC because the overall distribution of relations typically differs between domains, and because many of the content words that classifiers may rely on are very different between domains. We here focus on the most challenging subtask of *implicit discourse relation classification*, which involves classifying those relations that are not linked by any explicit connectives like "because" or "but". In order to correctly recognize implicit relations, the classifier needs to recognize subtle surface cues (which may differ between domains) and learn about typical content-related relations. For instance, from the example "it's hot outside, therefore I'd like to eat an icecream", the words "hot outside" and "icecream" are relevant cues for the relation. An overview of typical cues for determining a coherence relation is provided in Das and Taboada (2018).

The key to improving automatic DRC on a new domain hence consists of better encoding of the discourse relational arguments. As we will show below (in line with earlier findings by Shi and Demberg, 2019b), it makes a big difference to have at least a small amount of in-domain discourse annotated data.

We here explore DRC on the biomedical domain, which seems particularly suitable because a discourse-annotated corpus is available (BioDRB; Prasad et al., 2011), which we can use for evaluation, as well as a setting with a small amount of in-domain training data. Furthermore, the biomedical domain does have large raw text corpora available. An example instance from BioDRB (Prasad et al., 2011) is shown below:

1. *[These abnormalities in active RA are thought to be induced mainly after chronic exposure to high concentrations of IL-6.]*$_{Arg1}$ *(**Implicit=thus**) [The limited efficacy of IL-10*



*Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Short Papers)*, pages 925–931
August 1–6, 2021. 

*treatment of RA patients may be explained in part by the unresponsiveness to IL-10 of inflammatory cells, including T cells .]*$_{Arg2}$

—Implicit, Contingency.Cause

Scientific texts such as those from the biomedical domain are well known to express much of the content in nominal phrases, and less in verb phrases (Halliday, 2006). Concretely, for the above example, understanding the relation between the RA (Rheumatoid Arthritis) and inflammatory cells (including T cells) is important to correctly understanding the relation. The high importance of entities in these texts is a crucial insight on which we base our approach.

In this paper, we first propose an unsupervised method using information retrieval and knowledge graph techniques for identifying text passages that are similar content-wise to the coherence relation we want to label. The underlying assumption here is that if two instances share the same entities in both the relational arguments, it is possible that they have the same or a similar discourse relation. This part of the method is applicable to any domain for which large amounts of in-domain text are available, but no in-domain discourse relation annotations. We find that this method helps to improve results substantially compared to a Bi-LSTM baseline model, but doesn't reach state of the art performance (which is set by transformer models).

We therefore proceed to enrich a transformer model with the knowledge extracted from the unlabelled texts, using the K-BERT model (Liu et al., 2020). The model is fine-tuned on the discourse-annotated in-domain BioDRB data. We show that this setting sets the new state of the art on discourse relation classification on the biomedical domain, achieving an accuracy of 69.57%.

## 2 Related Work

Early approaches on BioDRB use probabilistic classifiers such as Naïve Bayes, Maximum Entropy, etc. to predict the relation (Xu et al., 2012). Bai and Zhao (2018) combine representations from different types of embeddings including contextualized word vectors from ELMo (Peters et al., 2018) and achieve 55.9% accuracy on BioDRB for in-domain training, and 29.52% in the cross-domain setting (reported in Shi and Demberg (2019b)).

Shi and Demberg (2019b) also explore the performance of BERT (Devlin et al., 2019) models on the DRC task on BioDRB using cross-domain (fine-tuning on PDTB, testing on BioDRB) as well as in-domain (fine-tuning on BioDRB and testing on BioDRB) settings. They find a very good performance of the BERT model, which they attribute to its "next sentence prediction" task in pre-training. Comparing the original BERT model to BioBERT (Lee et al., 2020), which was trained on biomedical text, they however find that BioBERT has only a limited ability for learning domain specific representations: Cross-domain performance is no better than for the BERT model, and in-domain performance improvements are moderate at only 1.5% points. Given that the entities play an important role in inferring implicit discourse relation in scientific texts, putting an emphasis on entities seems vital for achieving further improvements.

In contrast with previous studies that (largely unsuccessfully) attempted to train on explicit discourse relations for learning to classify implicit classifiers in supervised ways, such as Marcu and Echihabi (2002); Sporleder and Lascarides (2008); Biran and McKeown (2013); Qin et al. (2017); Shi et al. (2017) etc., we here propose an unsupervised voting pipeline and achieve good performance even comparing with supervised models like BERT and BioBERT. We believe that the key difference lies in the fact that previous methods tried to learn *surface cues* from explicit relations and tried to use them for implicits (which does not work, because these features differ between explicits and implicits, see e.g., Sporleder and Lascarides (2008); Asr and Demberg (2012)), while our method focuses on the content of the discourse relational arguments.

## 3 Unsupervised Method with Information Retrieval System

The successful usage of a memory network in Shi and Demberg (2019a) showed that instances that share the same relation have close representations. We believe that for sparse data like BioDRB, which has only around 2,000 labeled implicit instances in total, it is essential to use similar explicit instances to help find the latent patterns they share. In this section, we introduce an unsupervised method for implicit DRC, which is inspired by a recent information retrieval method.

The core idea is as follows: we use information retrieval methods to identify explicitly marked coherence relations from the corpus which are content-wise similar to the relation we want to la-

bel. We then automatically label these explicitly marked instances (relying on the high DRC accuracy of ca. 96% for explicit relations) and assign the majority label from the explicit instances to the implicit instance from our test set.

### 3.1 Retrieval of similar instances from a large corpus

Figure 1 illustrates the overall pipeline of the proposed method. First, each instance from BioDRB (Prasad et al., 2011) is seen as a query and fed into the PubMed[1] and PMC[2] databases.

**PubMed** and **PMC** are free full-text archives of biomedical and life sciences journal literature at NIH National Library of Medicine. The database we use here is a corpus created from a subset of the whole PubMed and PMC collections, consisting of 7,079 documents in total (1,376 for pubMed and 5,703 for PMC).

With the query and candidate documents, we employ TF-IDF to extract the top 10 relevant documents. The candidate documents are then fed into a discourse parser; we here use the PDTB-style end-to-end parser by Lin et al. (2014). The outputs of the parser contain the two arguments, the explicit discourse connective and a discourse relation label.

The Quasi Knowledge Graphs System, proposed by Lu et al. (2019), is designed to answer complex questions. It is a novel method that computes answers by dynamically building up a knowledge graph that fits the query. It consists of several steps including the extraction of subject-predicate-object (SPO) triples, knowledge graph construction, and a graph algorithm. We here only use the first step from this pipeline, extracting SPO triples, and actually only use the subject and object, not the predicate, to match with the noun phrases in the query. For example, from the relation instance in Example 1 above, the system would extract SPO triples *(NETosis, enhanced in, RA)* and *(autoantibodies, known risk factors for, RA)*, from which we further employ only *NETosis, RA; autoantibodies, RA*.

After extracting the SPO triples from all the explicit discourse instances, we employ two types of matching strategies to connect them with the query:

[1]PubMed [Internet]. Bethesda (MD): National Library of Medicine (US). [1946]. Available from: https://www.ncbi.nlm.nih.gov/pubmed/

[2]PubMed Central (PMC) [Internet]. Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information; 2000. Available from: https://www.ncbi.nlm.nih.gov/pmc/

| Methods | Cross-domain |
|---|---|
| Majority class | 20.66 |
| Bai and Zhao (2018) | 29.52 |
| Bi-LSTM + Word2Vec | 32.97 |
| BERT | 44.79 |
| BioBERT | 44.33 |
| Hard-matching | 35.29 |
| Soft-matching | 41.95 |

Table 1: Performances on BioDRB across domains. Across domains means that the model is trained on PDTB and tested on BioDRB. Majority class here is the majority relation of explicits.

(i) **Hard matching**, which means that if the subject or object appear in the query, we count it as a vote. (ii) **Soft matching**. We find that with the hard matching, lots of positive samples have been filtered out and very few explicit instances are identified. Therefore, we use the cosine similarity between the subject or object and the noun phrases in the query, to detect similar entities. Cosine similarities are estimated based on the BioBERT encoding of the entities. We define a threshold for deciding when an explicit instance is similar enough to be counted as a valid vote or not. It is seen in the training phase as a hyper-parameter to be fine-tuned on the validation set. This method for detecting similar explicit instances is also used in our second approach described in Section 4.

With the steps described above, eventually each query has been connected to a number of similar explicit instances and the prediction for the query is the majority vote from all of them with their explicit discourse sense labels.

### 3.2 Experiments and results

On average 813.99 explicit instances are extracted for each query. With the hard matching, 7.91 similar entities are matched with the Subject or the Object in the query. For the soft matching, we randomly choose 10% of the total instances acting as validation set in order to help set the threshold for the cosine similarity score.

The experimental results are shown in Table 1. We compare the results with related work by Bai and Zhao (2018) as well as several models reported in Shi and Demberg (2019b).

Our proposed unsupervised method achieves an accuracy of 35.29% with hard-matching and 41.95% with soft-matching. These results outper-

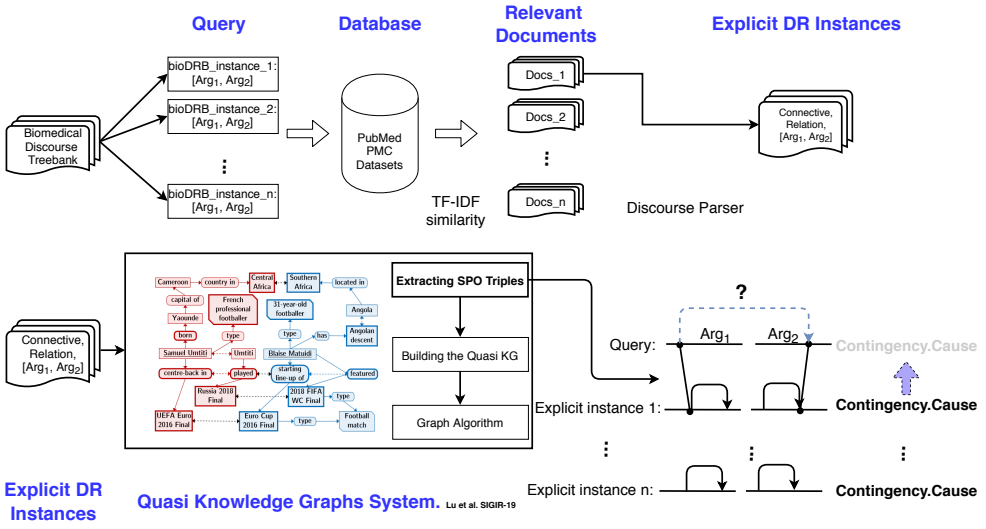


Figure 1: The Pipeline of the Proposed Method.

form other non-transformer approaches by a large margin. Comparing the hard and soft matching variants, our results show that identifying instances with similar entities leads to a larger set of relevant documents, which then help to increase robustness in the majority vote.

The table also shows that the approach almost reaches the performance of recent very strong transformer models: the BERT model achieves a performance of 44.79% accuracy in the cross-domain setting (Shi and Demberg, 2019b).

The approach proposed here could be further refined by using better argument representations than simple matching of subject and object entities, and by learning the classification decisions instead of using simple majority voting, and by moving to transformer architectures. Our second approach addresses these points by employing a transformer architecture which can take the SPO triple information into account for more richly encoding the relational arguments.

## 4 DRC with an entity-augmented transformer

Integrating external domain-specific knowledge into the model is beneficial for this task has been found by Kishimoto et al. (2018), who integrated the ConceptNet relations as additional knowledge into the LSTM network and achieved better performance on the PDTB.

We here aim to explore whether model performance can be further improved by exploiting richer entity representations in specialized texts

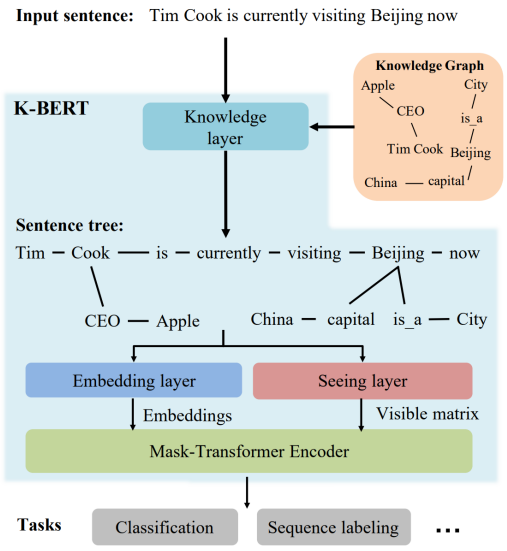


Figure 2: The structure of K-BERT. It is equipped with an editable knowledge graph which can be adapted to its application domain. Picture taken from Liu et al. (2020).

like the biomedical domain. The pipeline with soft-matching proposed in the above section provides us with SPO triples from related documents for each implicit relation instance in the test set. We here employ the recently proposed **Knowledge-enabled Language Representation model** (Liu et al., 2020, K-BERT) to integrate the external entity knowledge into the pre-trained language model for better argument representations.

### 4.1 K-BERT

Due to the domain gap between the pre-training and fine-tuning, unsupervised language models (such

as BERT etc.) do not perform well on knowledge-driven tasks (Liu et al., 2020). Integrating domain specific knowledge into pre-trained model can alleviate this problem. However, the process of knowledge acquisition can be inefficient and expensive.

In order to tackle the heterogeneous embedding space and knowledge noise problems, Liu et al. (2020) proposed a Knowledge-enabled Bidirectional Encoder Representation from Transformers (K-BERT), as illustrated in Figure 2. With the knowledge layer and the external knowledge graph, the input sentence has been expanded into a sentence tree, which is then fed into into the embedding layer and the "seeing" layer. The seeing layer controls when the model has access to the original sentence and when it has access to the additional information.

However, knowledge graphs are not available for all domains. We therefore here replace information from the knowledge graph with the SPO triples extracted from related raw texts. Compared to a general knowledge graph, our extracted SPO triples have attached more importance on the discourse relations since that they are extracted from the explicit instances, and are specifically selected to be on-topic. For each input sentence, we attach the top 2 (default number from the K-BERT) similar SPO triples to the entities and convert it into a sentence tree. We train K-BERT on the BioDRB as a classification task. The input sequence of the Example 1 is shown below, where the words in italics are the linked entities.

2. These abnormalities in active *NETosis enhanced in autoantibodies known risk factors for* RA *result in Neutrophil Chemotaxis* are thought to be induced mainly after chronic exposure to high concentrations of IL-6. The limited efficacy of IL-10 treatment of RA patients *reduced complement activation* may be explained in part by the unresponsiveness to IL-10 of inflammatory cells, including T cells *isolated from CTCL patient*.

The whole sentence tree has been flattened into a sequence with the position index. The visible matrix is generated to keep the interactions of each of the tokens within the original sentence and also inside the knowledge graph triples. The visible matrix controls the self-attention layers in the transformer not to look into tokens other than the corresponding entities.

| Methods | In-domain |
|---|---|
| Bai and Zhao (2018) | 55.90 |
| Bi-LSTM + Word2Vec | 46.49 |
| BERT | 63.02 |
| BioBERT | 67.58 |
| proposed model using K-BERT | **69.57*** |

Table 2: Performances on BioDRB within domain. Within domain here means 5-folds cross validation (see also Shi and Demberg (2017)) on BioDRB. * denotes significant improvement over BioBERT with $p<0.05$.

### 4.2 Experiments and Results

The experimental results are illustrated in Table 2. We compare the results with the previous state of the art on the BioDRB dataset (Shi and Demberg, 2019b). K-BERT, which is initialized with the original BERT parameters, achieves 69.57% accuracy and outperforms BERT without entity augmentation by 6.5% points, and the the gigantic in-domain continuously pre-trained BioBERT by around 2%. In addition, we tried to remove the relevant entities. The model then performed similar to the basic BERT, which is consistent with the results reported in Liu et al. (2020). These results confirm that adding related entities improves argument encoding and help improve the DRC task.

## 5 Conclusion

In this paper, we address the task of implicit discourse relation classification on BioDRB in the biomedical domain. Due to the importance of entities in scientific text, we decided to address this problem by identifying explicitly marked relations containing the same instances, and using a simple majority voting system. While this setting showed good performance in the unsupervised setting, much better results are achieved when at least a small amount of labelled data is available. We show that when a transformer model is augmented with entity information from the domain, the previous state of the art on the task is exceeded by 2% points.

## Acknowledgments

We would like to thank the anonymous reviewers for their valuable and constructive feedback. This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project-ID 232722074 – SFB 1102.

## References


Fatemeh Torabi Asr and Vera Demberg. 2012. Implicitness of discourse relations. In *Proceedings of COLING 2012*, pages 2669–2684.

Hongxiao Bai and Hai Zhao. 2018. Deep enhanced representation for implicit discourse relation recognition. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 571–583.

Or Biran and Kathleen McKeown. 2013. Aggregated word pair features for implicit discourse relation disambiguation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 69–73.

Lynn Carlson, Daniel Marcu, and Mary Ellen Okurowski. 2003. Building a discourse-tagged corpus in the framework of rhetorical structure theory. In *Current and new directions in discourse and dialogue*, pages 85–112. Springer.

Debopam Das and Maite Taboada. 2018. Rst signalling corpus: A corpus of signals of coherence relations. *Language Resources and Evaluation*, 52(1):149–184.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Shima Gerani, Yashar Mehdad, Giuseppe Carenini, Raymond T. Ng, and Bita Nejat. 2014. Abstractive summarization of product reviews using discourse structure. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 1602–1613. Association for Computational Linguistics.

Francisco Guzmán, Shafiq R. Joty, Lluís Màrquez, and Preslav Nakov. 2014. Using discourse structure improves machine translation evaluation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, pages 687–698. Association for Computational Linguistics.

Michael Alexander Kirkwood Halliday. 2006. *Language of science*, volume 5. Bloomsbury Publishing.

Peter Jansen, Mihai Surdeanu, and Peter Clark. 2014. Discourse complements lexical semantics for non-factoid answer reranking. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 977–986.

Yudai Kishimoto, Yugo Murawaki, and Sadao Kurohashi. 2018. A knowledge-augmented neural network model for implicit discourse relation classification. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 584–595.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

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.

Weijie Liu, Peng Zhou, Zhe Zhao, Zhiruo Wang, Qi Ju, Haotang Deng, and Ping Wang. 2020. K-bert: Enabling language representation with knowledge graph. In *AAAI*, pages 2901–2908.

Xiaolu Lu, Soumajit Pramanik, Rishiraj Saha Roy, Abdalghani Abujabal, Yafang Wang, and Gerhard Weikum. 2019. Answering complex questions by joining multi-document evidence with quasi knowledge graphs. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 105–114.

Daniel Marcu and Abdessamad Echihabi. 2002. An unsupervised approach to recognizing discourse relations. In *Proceedings of the 40th annual meeting of the association for computational linguistics*, pages 368–375.

Thomas Meyer, Najeh Hajlaoui, and Andrei Popescu-Belis. 2015. Disambiguating discourse connectives for statistical machine translation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 23(7):1184–1197.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237.

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 the Sixth International Conference on Language Resources and Evaluation (LREC08)*.

Rashmi Prasad, Susan McRoy, Nadya Frid, Aravind Joshi, and Hong Yu. 2011. The biomedical discourse relation bank. *BMC bioinformatics*, 12(1):188.

Lianhui Qin, Zhisong Zhang, Hai Zhao, Zhiting Hu, and Eric Xing. 2017. Adversarial connective-exploiting networks for implicit discourse relation classification. In *Proceedings of the 55th Annual*

*Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1006–1017.

Wei Shi and Vera Demberg. 2017. On the need of cross validation for discourse relation classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 150–156.

Wei Shi and Vera Demberg. 2019a. Learning to explicitate connectives with Seq2Seq network for implicit discourse relation classification. In *Proceedings of the 13th International Conference on Computational Semantics - Long Papers*, pages 188–199, Gothenburg, Sweden. Association for Computational Linguistics.

Wei Shi and Vera Demberg. 2019b. Next sentence prediction helps implicit discourse relation classification within and across domains. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5794–5800.

Wei Shi, Frances Yung, Raphael Rubino, and Vera Demberg. 2017. Using explicit discourse connectives in translation for implicit discourse relation classification. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 484–495.

Caroline Sporleder and Alex Lascarides. 2008. Using automatically labelled examples to classify rhetorical relations: An assessment. *Natural Language Engineering*, 14(3):369.

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 *The 2012 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE.