# Closing the Gap: Domain Adaptation from Explicit to Implicit Discourse Relations

**Yangfeng Ji  Gongbo Zhang  Jacob Eisenstein**
School of Interactive Computing
Georgia Institute of Technology
{jiyfeng,gzhang64,jacobe}@gatech.edu

## Abstract

Many discourse relations are explicitly marked with discourse connectives, and these examples could potentially serve as a plentiful source of training data for recognizing implicit discourse relations. However, there are important linguistic differences between explicit and implicit discourse relations, which limit the accuracy of such an approach. We account for these differences by applying techniques from domain adaptation, treating implicitly and explicitly-marked discourse relations as separate domains. The distribution of surface features varies across these two domains, so we apply a marginalized denoising autoencoder to induce a dense, domain-general representation. The label distribution is also domain-specific, so we apply a resampling technique that is similar to instance weighting. In combination with a set of automatically-labeled data, these improvements eliminate more than 80% of the transfer loss incurred by training an implicit discourse relation classifier on explicitly-marked discourse relations.

## 1 Introduction

Discourse relations reveal the structural organization of text, potentially supporting applications such as summarization (Louis et al., 2010; Yoshida et al., 2014), sentiment analysis (Somasundaran et al., 2009), and coherence evaluation (Lin et al., 2011). While some relations are signaled **explicitly** with connectives such as *however* (Pitler et al., 2008), many more are **implicit**. Expert-annotated datasets of implicit discourse relations are expensive to produce, so it would be preferable to use weak supervision, by automatically labeling instances with explicit connectives (Marcu and Echihabi, 2003).

However, Sporleder and Lascarides (2008) show that models trained on explicitly marked examples generalize poorly to implicit relation identification. They argued that explicit and implicit examples may be linguistically dissimilar, as writers tend to avoid discourse connectives if the discourse relation could be inferred from context (Grice, 1975). Similar observations are made by Rutherford and Xue (2015), who attempt to add automatically-labeled instances to improve supervised classification of implicit discourse relations.

In this paper, we approach this problem from the perspective of domain adaptation. Specifically, we argue that the reason that automatically-labeled examples generalize poorly is due to domain mismatch from the explicit relations (**source** domain) to the implicit relations (**target** domain). We propose to close the gap by using two simple methods from domain adaptation: (1) feature representation learning: mapping the source domain and target domain to a shared latent feature space; (2) resampling: modifying the relation distribution in the explicit relations to match the distribution over implicit relations. Our results on the Penn Discourse Treebank (Prasad et al., 2008) show that these two methods improve the performance on unsupervised discourse relation identification by more than $8.4\%$ on average $F_1$ score across all relation types, an $82\%$ reduction on the transfer loss incurred by training on explicitly-marked discourse relations.

## 2 Related Work

Marcu and Echihabi (2003) train a classifier for implicit intra-sentence discourse relations from

explicitly-marked examples in the rhetorical structure theory (RST) treebank, where the relations are automatically labeled by their discourse connectives: for example, labeling the relation as CONTRAST if the connective is *but*. However, Sporleder and Lascarides (2008) argue that explicitly marked relations are too different from implicit relations to serve as an adequate supervision signal, obtaining negative results in segmented discourse representation theory (SDRT) relations.

More recent work has focused on the Penn Discourse Treebank (PDTB), using explicitly-marked relations to supplement, rather than replace, a labeled corpus of implicit relations. For example, Biran and McKeown (2013) collect word pairs from arguments of explicit examples to help the supervised learning on implicit relation identification. Lan et al. (2013) present a multi-task learning framework, using explicit relation identification as auxiliary tasks to help main task on implicit relation identification. Rutherford and Xue (2015) explore several selection heuristics for adding automatically-labeled examples from Gigaword to their system for implicit relation detection, obtaining a 2% improvement in Macro-$F_1$. Our work differs from these previous efforts in that we focus exclusively on training from automatically-labeled explicit instances, rather than supplementing a training set of manually-labeled implicit examples.

Learning good feature representations (Ben-David et al., 2007) and reducing mismatched label distributions (Joshi et al., 2012) are two main ways to make a domain adaptation task successful. Structural correspondence learning is an early example of representation learning for domain adaptation (Blitzer et al., 2006); we build on the more computationally tractable approach of marginalized denoising autoencoders (Chen et al., 2012). Instance weighting is an approach for correcting label distribution mismatch (Jiang and Zhai, 2007); we apply a simpler approach of resampling the source domain according to an estimate of the target domain label distribution.

## 3 Domain Adaptation for Implicit Relation Identification

We employ two domain adaptation techniques: learning feature representations, and resampling to match the target label distribution.

### 3.1 Learning feature representation: Marginalized denoising autoencoders

The goal of feature representation learning is to obtain dense features that capture feature correlations between the source and target domains. Denoising autoencoders (Glorot et al., 2011) do this by first "corrupting" the original data, $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ into $\tilde{\boldsymbol{x}}_1, \ldots, \tilde{\boldsymbol{x}}_n$, either by adding Gaussian noise (in the case of real-valued data) or by randomly zeroing out features (in the case of binary data). We can then learn a function to reconstruct the original data, thereby capturing feature correlations and improving resilience to domain shift.

Chen et al. (2012) propose a particularly simple and elegant form of denoising autencoder, by marginalizing over the noising process. Their single-layer marginalized denoising autoencoder (mDA) solves the following problem:

$$\min_{\mathbf{W}} E_{\tilde{\boldsymbol{x}}_i|\boldsymbol{x}_i}[\|\boldsymbol{x}_i - \mathbf{W}\tilde{\boldsymbol{x}}_i\|^2] \quad (1)$$

where the parameter $\mathbf{W} \in \mathbb{R}^{d\times d}$ is a projection matrix. After learning the projection matrix, we use $\tanh(\mathbf{W}\boldsymbol{x})$ as the representation for our relation identification task.

Usually, $\boldsymbol{x}_i \in \mathbb{R}^d$ is a sparse vector with more than $10^5$ dimensions. Solving the optimization problem defined in equation 1 will produce a $d \times d$ dense matrix $\mathbf{W}$, and is prohibitively expensive. We employ the trick proposed by Blitzer et al. (2006) to select $\kappa$ *pivot* features to be reconstructed. We then split all features into non-overlapping subsets of size $\leq K$. Then, a set of projection matrices are learned, so as to transform each feature subset to the pivot feature set. The final projection matrix $\mathbf{W}$ is the stack of all projection matrices learned from the feature subsets.

### 3.2 Handling mismatched label distributions: Resampling with minimal supervision

There is a notable mismatch between the relation distributions for implicit and explicitly-marked discourse relations in the Penn Discourse Treebank: as shown in Figure 1, the EXPANSION and CONTINGENCY relations comprise a greater share of the implicit relations, while the TEMPORAL and COMPARISON relations comprise a greater share of the explicitly-marked discourse relations. Such label distribution mismatches can be a major source of transfer loss across domains, and therefore, reducing this mismatch can be an easy

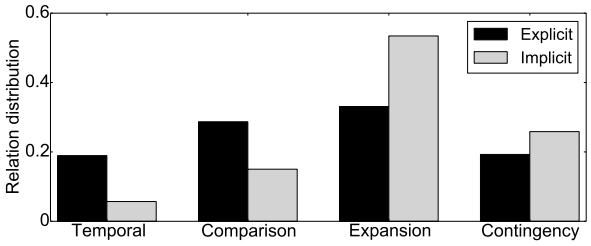


Figure 1: The relation distributions of training examples from the source domain (explicitly-marked relations) and target domain (implicit relations) in the PDTB.

way to obtain performance gains in domain adaptation (Joshi et al., 2012). Specifically, our goal is to modify the relation distribution in the source domain (explicitly-marked relations) and make it as similar as possible to the target domain (implicit relations). Given the label distribution from the target domain, we resample training examples from the source domain with replacement, in order to match the label distribution in the target domain. As this requires the label distribution from the target domain, it is no longer purely unsupervised domain adaptation; instead, we call it *resampling with minimal supervision*.

It may also be desirable to ensure that the source and target training instances are similar in terms of their observed features; this is the idea behind the *instance weighting* approach to domain adaptation (Jiang and Zhai, 2007). Motivated by this idea, we require that sampled instances from the source domain have a cosine similarity of at least $\tau$ with at least one target domain instance (Rutherford and Xue, 2015).

## 4 Experiments

Our experiments test the utility of the two domain adaptation methods, using the Penn Discourse Treebank (Prasad et al., 2008) and some extra-training data collected from a external resource.

### 4.1 Experimental setup

**Datasets** The test examples are implicit relation instances from section 21-22 in the PDTB. For the domain adaptation setting, the training set consists of the explicitly-marked examples extracted from sections 02-20 and 23-24, and the development set consists of the explicit relations from sections 21-22. All relations in the explicit examples are automatically labeled by using the connective-to-relation mapping from Table 2 in (Prasad et al., 2007), where we only keep the majority relation type for every connective. For each identified connective, we use its annotated arguments in the PDTB. As an upper bound, we also train a supervised discourse relation classifier, using the implicit examples in sections 02-20 and 23-24 as the training set, and using sections 00-01 as the development set. Following prior work (Pitler et al., 2009; Park and Cardie, 2012; Biran and McKeown, 2013), we consider the first-level discourse relations in the PDTB — Temporal (TEMP.), Comparison (COMP.), Expansion (EXP.) and Contingency (CONT.). We train binary classifiers and report $F_1$ score on each binary classification task. Extension of this approach to multi-class classification is important, but since this is not the setting considered in most of the prior research, we leave it for future work.

The true power of learning from automatically labeled examples is that we could leverage much larger datasets than hand-annotated corpora such as the Penn Discourse Treebank. To test this idea, we collected 1,000 news articles from CNN.com as extra training data. Explicitly-marked discourse relations from this data are automatically extracted by matching the PDTB discourse connectives (Prasad et al., 2007). For this data, we also need to extract the arguments of the identified connectives: for every identified connective, the sentence following this connective is labeled as Arg2 and the preceding sentence is labeled as Arg1, as suggested by Biran and McKeown (2013). In a pilot study we found that larger amounts of additional training data yielded no further improvements, which is consistent with the recent results of Rutherford and Xue (2015).

**Model selection** We use a linear support vector machine (Fan et al., 2008) as the classification model. Our model includes five tunable parameters: the number of pivot features $\kappa$, the size of the feature subset $K$, the noise level for the denoising autoencoder $q$, the cosine similarity threshold for resampling $\tau$, and the penalty parameter $C$ for the SVM classifier. We consider $\kappa \in \{1000, 2000, 3000\}$ for pivot features and $C \in \{0.001, 0.01, 0.1, 1.0, 10.0\}$ for penalty parameters, $q \in \{0.90, 0.95, 0.99\}$ for noise levels. To reduce the free parameters, we set $K = 5\kappa$ and simply fix the cosine similarity threshold $\tau =$

| Surface Features | +Rep. Learning | +Resampling | Relations | | | | Average $F_1$ |
|---|---|---|---|---|---|---|---|
| | | | TEMP. | COMP. | EXP. | CONT. | |
| *Implicit → Implicit* | | | | | | | |
| 1. FULL | | | 24.15 | 28.87 | 68.84 | 43.45 | 41.32 |
| *Explicit [PDTB] → Implicit* | | | | | | | |
| 2. FULL | No | No | 17.13 | 20.54 | 50.55 | 36.14 | 31.04 |
| 3. FULL | No | Yes | 15.38 | 23.88 | 62.04 | 35.29 | 34.14 |
| 4. FULL | Yes | No | 17.53 | 22.77 | 50.85 | 36.43 | 31.90 |
| 5. FULL | Yes | Yes | 17.05 | 22.00 | 63.51 | 38.23 | 35.20 |
| 6. PIVOT | No | No | 17.33 | 23.89 | 53.53 | 36.22 | 32.74 |
| 7. PIVOT | No | Yes | 17.73 | 25.39 | 62.65 | 36.02 | 35.44 |
| 8. PIVOT | Yes | No | 18.66 | 25.86 | 63.37 | 38.87 | 36.69 |
| 9. PIVOT | Yes | Yes | 19.26 | 25.74 | 68.08 | 41.39 | 38.62 |
| *Explicit [PDTB + CNN] → Implicit* | | | | | | | |
| 10. PIVOT | Yes | Yes | 20.35 | 26.32 | 68.92 | 42.25 | 39.46 |

Table 1: Performance of cross-domain learning for implicit discourse relation identification.

$0.85$; pilot studies found that results are not sensitive to the value of $\tau$ across a range of values.

**Features** All features are motivated by prior work on implicit discourse relation classification: from each training example with two arguments, we extract (1) Lexical features, including word pairs, the first and last words from both arguments (Pitler et al., 2009); (2) Syntactic features, including production rules from each argument, and the shared production rules between two arguments (Lin et al., 2009); (3) Other features, including modality, Inquirer tags, Levin verb classes, and argument polarity (Park and Cardie, 2012). We re-implement these features as closely as possible to the cited works, using the Stanford CoreNLP Toolkit to obtain syntactic annotations (Manning et al., 2014).

The FULL feature set for domain adaptation is constructed by collecting all features from the training set, and then removing features that occur fewer than ten times. The PIVOT feature set includes $\kappa$ high-frequency features from the FULL feature set. To focus on testing the domain adaptation techniques, we use the same FULL and PIVOT set for all four relations, and leave feature set optimization for each relation as a future work (Park and Cardie, 2012). To obtain the upper bound, we employ the same feature categories and frequency threshold to extract features from the in-domain data, hand-annotated implicit discourse relations. To include the representations from the marginalized denoising autoencoder for relation identification, we concatenate them with the original surface feature representations of the same examples.

### 4.2 Experimental results

In experiments, we start with surface feature representations as baselines, then incorporate the two domain adaptation techniques incrementally. As shown in line 2 of Table 1, the performance is poor if directly applying a model trained on the explicit examples with the FULL feature set, which is consistent with the observations of Sporleder and Lascarides (2008): there is a $10.28\%$ absolute reduction on average $F_1$ score from the upper bound obtained with in-domain supervision (line 1). With mDA, the overall performance increases by $0.86\%$ (line 4); resampling gives a further $4.16\%$ improvement mainly because of the performance gain on the EXP. relation (line 5). The resampling method itself (line 3) also gives a better overall performance then mDA (line 4). However, the $F_1$ scores on the TEMP. and CONT. are even worse than the baseline (line 2).

Surface representations with the FULL feature set were found to cause serious overfitting in the experiments. To deal with this problem, we propose to use only $\kappa$ pivot features, which gives a stronger baseline of the cross-domain relation identification, as shown in line 6. Then, by incorporating resampling and feature representation

learning individually, the average $F_1$ increases from $32.74\%$ to $35.44\%$ (line 7) and $36.69\%$ (line 8) respectively. The combination of these two domain adaptation techniques boosts the average $F_1$ further to $38.62\%$ (line 9). The additional CNN training data further improves performance to $39.46\%$ (line 10). This represents an $8.42\%$ improvement of average $F_1$ from the original result (line 2), for more than $80\%$ reduction on the transfer loss incurred by training on explicit discourse relations.

An additional experiment is to use automatic argument extraction in both the PDTB and the CNN data, which would correspond to more truly unsupervised domain adaptation. (Recall that in the CNN data, we used adjacent sentences as argument spans, while in the PDTB data, we use expert annotations.) When using adjacent sentences as argument spans in both datasets, the average $F_1$ is $38.52\%$ for the combination of representation learning and resampling. Compared to line 10, this is a $0.94\%$ performance drop, indicating the importance of argument identification in the PDTB data. In future work we may consider better heuristics for argument extraction, such as obtaining automatically-labeled examples only from those connectors for whom the arguments usually are the adjacent sentences; for example, the connector *nonetheless* usually connects adjacent spans (e.g., *Bob was hungry. Nonetheless he gave Tina the burger.*), while the connector *even though* may connect two spans that follow the connector in the same sentence (e.g., *Even though Bob was hungry, he gave Tina the burger.*).

## 5 Conclusion

We have presented two methods — feature representation learning and resampling — from domain adaptation to close the gap of using explicit examples for unsupervised implicit discourse relation identification. Experiments on the PDTB show the combination of these two methods eliminates more than $80\%$ of the transfer loss caused by training on explicit examples, increasing average $F_1$ from $31\%$ to $39.5\%$, against a supervised upper bound of $41.3\%$. Future work will explore the combination of this approach with more sophisticated techniques for instance selection (Rutherford and Xue, 2015) and feature selection (Park and Cardie, 2012; Biran and McKeown, 2013), while also tackling the more difficult problems of multi-class relation classification and fine-grained level-2 discourse relations.

**Acknowledgments** This research was supported by a Google Faculty Research Award to the third author. The following members of the Georgia Tech Computational Linguistics Laboratory offered feedback throughout the research process: Parminder Bhatia, Naman Goyal, Vinodh Krishan, Umashanthi Pavalanathan, Ana Smith, Yijie Wang, and especially Yi Yang. Thanks to the reviewers for their constructive and helpful suggestions.

## References

Shai Ben-David, John Blitzer, Koby Crammer, Fernando Pereira, et al. 2007. Analysis of representations for domain adaptation. In *Neural Information Processing Systems (NIPS)*, Vancouver.

Or Biran and Kathleen McKeown. 2013. Aggregated word pair features for implicit discourse relation disambiguation. In *Proceedings of the Association for Computational Linguistics (ACL)*, pages 69–73, Sophia, Bulgaria.

John Blitzer, Ryan McDonald, and Fernando Pereira. 2006. Domain adaptation with structural correspondence learning. In *Proceedings of Empirical Methods for Natural Language Processing (EMNLP)*, pages 120–128.

Minmin Chen, Z. Xu, Killian Weinberger, and Fei Sha. 2012. Marginalized denoising autoencoders for domain adaptation. In *Proceedings of the International Conference on Machine Learning (ICML)*.

Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. 2008. LIBLINEAR: A library for large linear classification. *The Journal of Machine Learning Research*, 9:1871–1874.

Xavier Glorot, Antoine Bordes, and Yoshua Bengio. 2011. Domain adaptation for large-scale sentiment classification: A deep learning approach. In *Proceedings of the International Conference on Machine Learning (ICML)*, Seattle, WA.

H Paul Grice. 1975. Logic and Conversation. In P. Cole and J. L. Morgan, editors, *Syntax and Semantics Volume 3: Speech Acts*, pages 41–58. Academic Press.

Jing Jiang and ChengXiang Zhai. 2007. Instance weighting for domain adaptation in nlp. In *Proceedings of the Association for Computational Linguistics (ACL)*, Prague.

Mahesh Joshi, William W Cohen, Mark Dredze, and Carolyn P Rosé. 2012. Multi-domain learning: when do domains matter? In *Proceedings of the*

*2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1302–1312. Association for Computational Linguistics.

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

Ziheng Lin, Min-Yen Kan, and Hwee Tou Ng. 2009. Recognizing implicit discourse relations in the penn discourse treebank. In *Proceedings of Empirical Methods for Natural Language Processing (EMNLP)*, pages 343–351, Singapore.

Ziheng Lin, Hwee Tou Ng, and Min-Yen Kan. 2011. Automatically Evaluating Text Coherence Using Discourse Relations. In *Proceedings of the Association for Computational Linguistics (ACL)*, pages 997–1006, Portland, OR.

Annie Louis, Aravind Joshi, and Ani Nenkova. 2010. Discourse indicators for content selection in summarization. In *Proceedings of the 11th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 147–156. Association for Computational Linguistics.

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 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60.

Daniel Marcu and Abdessamad Echihabi. 2003. An unsupervised approach to recognizing discourse relations. In *Proceedings of the Association for Computational Linguistics (ACL)*, pages 368–375.

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*, pages 108–112, Seoul, South Korea, July. Association for Computational Linguistics.

Emily Pitler, Mridhula Raghupathy, Hena Mehta, Ani Nenkova, Alan Lee, and Aravind Joshi. 2008. Easily identifiable discourse relations. In *Proceedings of the International Conference on Computational Linguistics (COLING)*, pages 87–90, Manchester, UK.

Emily Pitler, Annie Louis, and Ani Nenkova. 2009. Automatic Sense Prediction for Implicit Discourse Relations in Text. In *Proceedings of the Association for Computational Linguistics (ACL)*, Suntec, Singapore.

Rashmi Prasad, Eleni Miltsakaki, Nikhil Dinesh, Alan Lee, Aravind Joshi, Livio Robaldo, and Bonnie L Webber. 2007. The Penn Discourse Treebank 2.0 annotation manual. The PDTB Research Group.

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*.

Attapol Rutherford and Nianwen Xue. 2015. Improving the Inference of Implicit Discourse Relations via Classifying Explicit Discourse Connectives. In *Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pages 799–808, Denver, CO, May–June.

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 Empirical Methods for Natural Language Processing (EMNLP)*, Singapore.

Caroline Sporleder and Alex Lascarides. 2008. Using automatically labelled examples to classify rhetorical relations: An assessment. *Natural Language Engineering*, 14(3):369–416.

Yasuhisa Yoshida, Jun Suzuki, Tsutomu Hirao, and Masaaki Nagata. 2014. Dependency-based Discourse Parser for Single-Document Summarization. In *Proceedings of Empirical Methods for Natural Language Processing (EMNLP)*.