Let's be explicit about that: Distant supervision for implicit discourse relation classification via connective prediction

Murathan Kurfalı
Department of Linguistics
Stockholm University
Stockholm, Sweden
murathan.kurfali@ling.su.se

Robert Östling
Department of Linguistics
Stockholm University
Stockholm, Sweden
robert@ling.su.se

Abstract

In implicit discourse relation classification, we want to predict the relation between adjacent sentences in the absence of any overt discourse connectives. This is challenging even for humans, leading to shortage of annotated data, a fact that makes the task even more difficult for supervised machine learning approaches. In the current study, we perform implicit discourse relation classification without relying on any labeled implicit relation. We sidestep the lack of data through explicitation of implicit relations to reduce the task to two subproblems: language modeling and explicit discourse relation classification, a much easier problem. Our experimental results show that this method can even marginally outperform the state-of-the-art, in spite of being much simpler than alternative models of comparable performance. Moreover, we show that the achieved performance is robust across domains as suggested by the zero-shot experiments on a completely different domain. This indicates that recent advances in language modeling have made language models sufficiently good at capturing inter-sentence relations without the help of explicit discourse markers.

1 Introduction

Discourse relations describe the relationship between discourse units, e.g. clauses or sentences. These relations are either signalled explicitly with a discourse connective (e.g. *because*, *and*) or expressed implicitly and are inferred by sequential reading (Example 1 below).

(1) A figure above 50 indicates the economy is likely to expand. **[While]** One below 50 indicates a contraction may be ahead. *(Comparison - wsj_0233)*

The relations in the latter category are called *implicit discourse relations* and they are of special significance because their lack of an explicit signal makes them challenging to annotate for even humans, suggested by the lower inter-annotator agreements on implicit relations (Zeyrek and Kurfalı, 2017; Zikánová et al., 2019), let alone classify automatically.

Resources for implicit discourse relations, therefore, are very limited. Even the Penn Discourse Tree Bank 2.0 (PDTB 2.0) (Prasad et al., 2008), which is the most popular resource, includes merely 16K implicit discourse relations, all annotated on the same domain. Explicit discourse relations, on the other hand, are proven to be simple enough to be obtained both manually and automatically. Previous work shows that explicit relations in English have a low level of ambiguity, so the discourse relation can be classified with more than 94% accuracy from the discourse connective alone (Pitler and Nenkova, 2009). This has inspired others to predict connectives for the implicit discourse relations and add them as additional features to existing supervised classifiers (Zhou et al., 2010; Xu et al., 2012).

Our work takes this idea one step further by reducing the amount of supervision required. Instead of training a separate connective classifier, we generate a set of *candidate explicit relations* that are obtained by inserting explicit discourse markers between sentences and score the resulting segments using a large pre-trained language model.[1] The candidates are then classified with an accurate explicit discourse relation classifier, and the final implicit relation prediction can be obtained by either using the candidate with the highest-scoring connective, or marginalizing over the whole distribution of explicit connectives.

The main contributions of our papers are as follows:

[1] In the reminder of the paper, these candidate explicit relations are simply referred as *candidates*.

Proceedings of the 1st Workshop on Understanding Implicit and Underspecified Language (UnImplicit 2021), pages 1–10
Bangkok, Thailand (online), August 5, 2021.

- We show that this simple approach is very effective and even marginally outperforms the current state-of-the-art method that does not use labeled implicit discourse relation data, even though that method uses a significantly more complex adversarial domain adaptation model (Huang and Li, 2019).

- To the best of our knowledge, this is the first study to go beyond the default four-way classification under the low-resource scenario assumption where no labeled implicit discourse relation is available. We show that the proposed pipeline maintains its performance (relative to the baselines) in a more challenging 11-way classification as well as across domains (i.e., biomedical texts (Prasad et al., 2011)).

- We offer explicitation of implicit discourse relations as a probing task to evaluate language models. Despite their relevancy, discourse relations are mostly overlooked in the assessments of language models' understanding of context. As a secondary aim, we investigate a wide range of pre-trained language models' understanding of inter-sentential relations.

We hope that the proposed pipeline will be another step in overcoming the data-bottleneck problem in discourse studies.

2 Background

2.1 Implicit Discourse Relations

PDTB 2.0 adopts a lexicalized approach where each relation consists of a discourse connective (e.g. "but", "and") which acts as a predicate taking two arguments. For each relation, annotators were asked to annotate the connective, the two text spans that hold the relation and the sense it conveys based on the PDTB sense hierarchy (Prasad et al., 2008). The text span which is syntactically bound to the connective is called the second argument (arg2) whereas the other is the first argument (arg1). "Additionally, implicit relations are annotated with that explicit connective which according to judgements best expresses the sense of the relation."

However, in certain cases, a relation holds between the adjacent sentences despite the lack of an overt connective (see Example 1). PDTB 2.0 recognizes such relations as *implicit discourse relations*. Additionally, implicit relations are annotated with an explicit connective which best expresses the sense of the relation is according to annotators. The connective inserted by the annotators is termed as "*implicit connective*" (e.g. "while" in Example 1). Unlike explicit relations where there is an explicit textual cue (the connective), implicit relations can only be inferred which makes them more challenging to spot and annotate.

2.2 Related Work

The research on implicit discourse relation classification is overwhelmingly supervised (Pitler et al., 2009; Rutherford and Xue, 2015; Lan et al., 2017; Nie et al., 2019; Kim et al., 2020). Although unsupervised methods were present in the earliest attempts (Marcu and Echihabi, 2002), they haven't received serious attention and much research concentrated on increasing the available supervision to deal with the data; most prominently, either by automatically generating artificial data (Sporleder and Lascarides, 2008; Braud and Denis, 2014; Rutherford and Xue, 2015; Wu et al., 2016; Shi et al., 2017) or through introducing auxiliary but similar tasks to the training routine to leverage additional information (Zhou et al., 2010; Xu et al., 2012; Liu et al., 2016; Lan et al., 2017; Qin et al., 2017; Shi and Demberg, 2019a; Nie et al., 2019). Zhou et al. (2010) and Xu et al. (2012) constitute the earliest examples where the classification of implicit relations are assisted via connective prediction. Both studies employ language models to predict suitable connectives for implicit relations which are, then, either used as additional features or classified directly.

Ji et al. (2015) is one of the few recent distantly supervised[2] studies which tackle implicit relation classification as a domain adaptation problem where the labeled explicit relations are regarded as the source domain and the unlabeled implicit relations as the target. Huang and Li (2019) improves upon Ji et al. (2015) by employing adversarial domain adaption with a novel reconstruction component.

2.3 Pre-trained Language Models

BERT Bidirectional Encoder Representations for Transformers (BERT) is a multi-layer Transformer encoder based language model (Devlin

[2]Previous work uses the term *unsupervised* (domain adaptation). Although we use the same amount of supervision with earlier work (no labeled implicit relation are utilized), we believe *distant supervision* describes the method better.

et al., 2019). As opposed to directional models where the input is processed from one direction to another, the transformer encoder reads its input at once; hence, BERT learns word representations in full context (both from left and from right). BERT is trained with two pre-training objectives on large-scale unlabeled text: (i) Masked Language Modelling and (ii) Next Sentence Prediction.

A number of BERT variants are available that differ in terms of (i) their architecture, e.g. BERT-base (*12-layer, 110M parameters*) and BERT-large (*24-layer, 340M parameters*); (ii) whether the letter casing in its input is preserved (-cased) or not (-uncased); (iii) their masking strategy, e.g. word pieces (*default)* or whole words (-whole-word-masking).

RoBERTa RoBERTa (Liu et al., 2019) shares the same architecture as BERT but improves upon it via introducing a number of refinements to the training procedure, such as using more data with larger batch sizes, adopting a larger vocabulary, removal of the NSP objective and dynamic masking.

DistilBERT DistilBERT was introduced by (Sanh et al., 2019). It is created by applying knowledge distillation to BERT which is a compression technique in which a small model learns to mimic the full output distribution of the target model (in this case: BERT). DistilBERT is claimed to retain 97% of BERT performance despite being 40% smaller and 60% faster, as suggested by its performance on Question Answering task.

GPT-2 Generatively Pre-trained Transformer (GPT-2) is a unidirectional transformer based language model trained on a dataset of 40 GB of web crawling data (Radford et al., 2019). Unlike BERT, GPT-2 works like a traditional language model where each token can only attend to its previous context. GPT-2 has four variants which differ from each other in the number of layers, ranging from 12 (small) to 48 (XL).

3 Model

The proposed method consists of three main components: (i) a candidate generator that generates sentence pairs connected by each of a set of discourse connectives, (ii) a language model that estimates the likelihood of each candidate, and (iii) an explicit discourse relation classifier to be used on the candidates. Whole pipeline is shown in Figure 1. The proposed methodology does not require even a single implicit discourse relation annotation and is only distantly supervised where the supervision comes from the explicit discourse relations used in training the classifier.

The main motivation behind the proposed pipeline is the finding that discourse relations are easily classifiable if they are explicitly marked (Pitler and Nenkova, 2009). We further verify this finding via a preliminary experiment which showed that four-way classification could be performed with an F-score of 88.74 when the implicit discourse relations are "explicitated" with the *gold implicit connectives* they are annotated with (see Table 2). This finding is significant not only because it justifies our motivation but also shows the potential of the current approach. Secondarily, the task requires a high level understanding of the context which allows us to investigate the pre-trained language models capabilities in detecting inter-sentential relations.

3.1 Candidate Generation

Recall Example 1, which contains an implicit relation between argument 1 ("*A figure above . . . to expand.*") and argument 2 ("*one below . . . be ahead.*").

Given a list of English connectives (*and*, *because*, *but*, etc.), we generate the following explicit relation candidates for a given implicit relation:

A_1	C	A_2
. . . to expand	**and**	[o]ne below . . .
. . . to expand	**because**	[o]ne below . . .
. . . to expand	**but**	[o]ne below . . .

The list of connectives are chosen among the lexical items PDTB 2.0 annotation guideline recognizes as discourse connectives (Prasad et al., 2008). Of the listed 100 connectives,[3] we limit ourselves to 65 one-word connectives to generate the candidates due to masked language models' inability to predict multiple tokens simultaneously.

3.2 Prediction of Implicit Connectives

Our next task is to produce a distribution over connectives C conditioned on the context (arguments A_1 and A_2). For unidirectional language models (in our case: GPT-2 variants), we estimate this by computing the language model likelihood of the entire candidates and normalizing over the connec-

[3]The modified connectives such as "partly because" are not counted as distinct types.

Model	4-Way	11-Way
PDTB train (Explicit)	13639	12695
PDTB test	1046	1040
BioDRB(test)	247	140
BioDRB(full)	3001	1755

Table 1: Number of instances in the respective datasets. For the BioDRB test and full distinctions, please refer to Section 5.3.

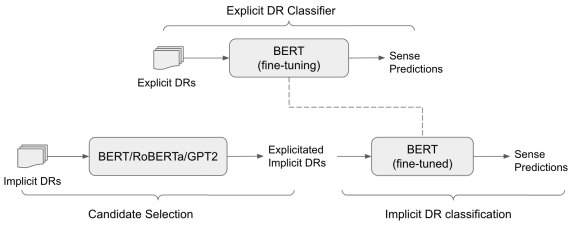


Figure 1: A high level visualization of the proposed pipeline.

tives:

$$P_{Conn}(C|A_1, A_2) \propto P_{LM}(A_1\ C\ A_2)$$

With bidirectional masked language models (in our case: DistilBERT, BERT and RoBERTa) we need to instead provide a candidate template by inserting the special sentence separation ([SEP]) and masking ([MASK]) tokens. Then it is simply a matter of normalizing over the model's estimated probability of the connective being inserted at the position of the masking token:

$$P_{Conn}(C|A_1, A_2) \propto P_{LM}(C|A_1\ \text{[SEP]}\ \text{[MASK]}\ A_2\ \text{[SEP]})$$

3.3 Explicit Discourse Relation Classifier

We regard discourse relation classification as a sentence pair classification task and build a classifier on top of the pre-trained BERT model from Devlin et al. (2019) using the recommended fine-tuning strategy. Specifically, the first and second arguments are separated via the special separator token ([SEP]) with the connective on the second argument and the [CLS] token is used for classification through a fully connected layer with softmax activation. This classifier gives us a model for the distribution $P_{Exp}(l|C, A_1, A_2)$ of relation labels l conditioned on the connective C and its arguments A_1 and A_2.

The annotation of explicit and implicit relations in the PDTB 2.0 differ in several aspects. In the case of implicit relations, PDTB 2.0 annotates arguments in the order they appear in the text, hence implicit relations can only manifest one configuration (i.e. arg1, [conn], arg2). On the other hand, the relative argument order of the explicit relations can vary to the extent that sometimes the arguments may interrupt each other (e.g. *Of course*, if **the film contained dialogue**, *Mr. Lane's Artist would be called a homeless person.* [from wsj-0039]). In order to remedy for this disparity to some extent, we only use the explicit relations which share the same relative argument order with implicit relations (i.e. arg1, conn, arg2) in training the classifier so that there is not any discrepancy in terms of the relation structure between training and inference phases. In total, 2558 (13.85%) explicit relations that do not follow the (arg1,conn,arg2) order are left out.

3.4 Final Model

In our experiments we combine the models in two ways. The simplest way is a straightforward pipeline approach, where the single most likely implicit connective is predicted, and then fed to the explicit relation classifier:

$$P(l|A_1, A_2) = P_{Exp}(l|\arg\max_C P_{Conn}(C|A_1, A_2), A_1, A_2)$$

Even though the level of ambiguity in English discourse connectives is relatively low, we also try to account for this ambiguity by marginalizing over all connectives:

$$P(l|A_1, A_2) = \sum_C P_{Exp}(l|C, A_1, A_2) \times P_{Conn}(C|A_1, A_2)$$

4 Experiments

We follow the experimental setting of Huang and Li (2019) which is originally adopted by (Ji et al., 2015). The implicit relations in the PDTB 2.0 sections 21-22 are allocated as the test set whereas the explicit relations in sections 2-20;23-24 are used as the training and 0-1 as the development set of the explicit relation classifier. The evaluation is performed for both the four first-level and the most common 11 second-level senses. For the former, we report both per-class and the macro-average F1-scores similar to Huang and Li (2019) whereas the accuracy is also reported on the second level

Model	Temp	Cont	Comp	Exp	4-way	11-way	
	F-score						Acc
Supervised	45.85	57.74	58.35	75.01	59.24	39.33	55.42
Gold Connective	77.29	96.23	88.01	93.42	88.74	57.56	78.87
Most Common Conn (but)	0.00	0.00	24.45	0.07	6.13	2.68	11.60
Most Common Sense	0.00	0.00	0.00	69.41	17.35	3.74	25.89
(Ji et al., 2015)	19.26	41.39	25.74	**68.08**	38.62	-	-
(Huang and Li, 2019)	**31.25**	48.04	25.15	59.15	40.90	-	-
BERT-base-uncased	14.97	29.06	32.05	59.45	33.88	13.88	23.16
+ Margin	19.49	36.54	32.48	52.99	35.37	14.12	24.45
BERT-large-cased	9.33	27.06	36.89	68.58	35.47	12.29	24.55
+ Margin	9.76	35.62	38.80	67.97	38.04	13.24	26.86
BERT-large-cased-wwm	10.89	36.29	42.69	62.38	38.06	16.79	27.23
+ Margin	15.02	41.97	41.81	60.80	39.90	17.50	28.74
BERT-large-uncased	25.69	30.55	30.10	62.50	37.21	15.57	25.25
+ Margin	27.32	41.01	32.28	59.07	39.92	15.67	28.01
BERT-large-uncased-wwm	18.35	35.20	40.19	58.88	38.15	16.47	26.80
+ Margin	17.27	42.93	41.16	55.61	39.24	17.26	29.17
DistilBERT-base-cased	16.77	46.19	23.19	39.07	31.31	15.87	22.71
+ Margin	21.09	**48.05**	29.25	37.04	33.86	16.68	26.38
RoBERTa-base	9.65	19.64	36.57	66.72	33.14	11.67	23.54
+ Margin	9.18	22.77	35.90	66.36	33.55	13.01	25.32
RoBERTa-large	10.79	30.32	48.35	68.44	39.48	16.15	27.66
+ Margin	13.30	33.19	**49.52**	67.90	40.98	17.63	29.57
GPT2	16.60	31.96	35.79	62.62	36.74	11.68	24.04
+ Margin	18.27	37.31	35.93	61.70	38.30	13.07	26.02
GPT2-large	19.91	35.27	40.38	59.17	38.68	15.18	25.63
+ Margin	23.17	40.55	40.30	60.39	**41.10**	16.03	27.50
GPT2-XL	21.59	30.88	40.18	63.01	38.92	16.98	26.01
+ Margin	23.06	34.49	42.66	63.98	41.05	18.50	28.32

Table 2: The results of the proposed methodology with various pre-trained language models. The average performance over four runs is reported (numbers within parentheses indicate the standard deviation). L stands for 'large' and wwm stands for 'whole-word-masking'. "+ Margin" refers to the second inference strategy explained in Section 3.4. Best scores are presented in bold, second bests are in italics (excluding the baselines).

senses following the standard in the literature. The statistics of the used datasets are provided in Table 1.

The classifiers are implemented using the Transformers library by Huggingface (Wolf et al., 2020). We use the uncased BERT large model for the explicit relation classifier (Section 3.3). The model is fine-tuned for ten epochs with a batch size of 16, learning rate of 5×10^{-6}. To optimize the loss function, we use Adam with fixed weight decay (Loshchilov and Hutter, 2018) and warm-up linearly for the first 1K steps. The model is evaluated with the step size of 500 and the one with the best development performance is used as the final model.

We mainly compare our results against the recent unsupervised studies we are aware of (Huang and Li, 2019; Ji et al., 2015). Additionally, we report the performance of a number baselines and upper bounds to put the results into a perspective:

- **Most Common Sense:** The performance when the most common sense of each evaluation level is predicted for every relation in the test set (Expansion for the first level; Contingency.cause for the second).
- **Most Common Connective:** The performance when the candidate with the most common explicit connective (*but*) is selected for

every relation in the test set.

- **Gold Connective:** The performance when the candidate with the gold implicit connective is selected. This baseline also shows the upper bound of the proposed pipeline (see Section 3.
- **Supervised baseline:** This is the results of the BERT classifier fine-tuned on the implicit discourse relations.

5 Results and Discussion

5.1 Evaluation on PDTB

The results are provided in Table 2. Overall, the 4-way classification F-score ranges between 33.86 (DistilBERT) to 41.10 (GPT2-large) where three models outperform the previous state-of-the-art (RoBERTa-large, GPT2-large, GPT2-XL). Moreover, the performance is robust across different sense levels as suggested by its relative performance to the baselines in the more challenging 11-way classification.

In addition to the increase in the overall performance, the most substantial gain is observed in Comparison relations where the unsupervised state-of-the-art is improved by almost 25% points to 49.52%, bringing it closer to the supervised baseline (58.35%). The relatively successful performance in Comparison relations hold for all language models, suggesting that language models are good at detecting the cues for these relations.

Marginalizing over all connectives leads constant improvements with all language models. Marginalization yields average gain of 2.12% when with BERT-variants and 2.04% with GPT2 models. This step alters only a small portion of predictions, on average 10.1% of the predictions change after marginalization. Relation-wise Contingency benefits from this step most with the average increase of 4.20%. In order to have a better insight, we closely inspect the label shifts in RoBERTa-large's predictions which reveals that the most frequent label shift is from Expansion to Contingency relations (41.1%). These changes mostly occur when there is a clear mismatch between the top connective and others following it in terms of their sense. To illustrate, Example 2 presents a relation, label of which was changed from *Expansion* to *Contingency* where the top five selected connectives were: "and","as","because","since","for". Of these connectives, only "and" dominantly conveys Expansion whereas others commonly convey Contingency. Marginalization acts as a corrective step in such cases and saves the model from depending on the top-rank connective by allowing it to consider the connective predictions with lower ranks.

(2) *Experts are predicting a big influx of new shows in 1990, when a service called "automatic number information" will become widely available.* [IMP=because] **This service identifies each caller's phone number, and it can be used to generate instant mailing lists.**

Finally, as for 11-way classification, the same pattern also holds where marginalization leads to the average of 1.07% and 2.27% improvement in F-score and accuracy, respectively.

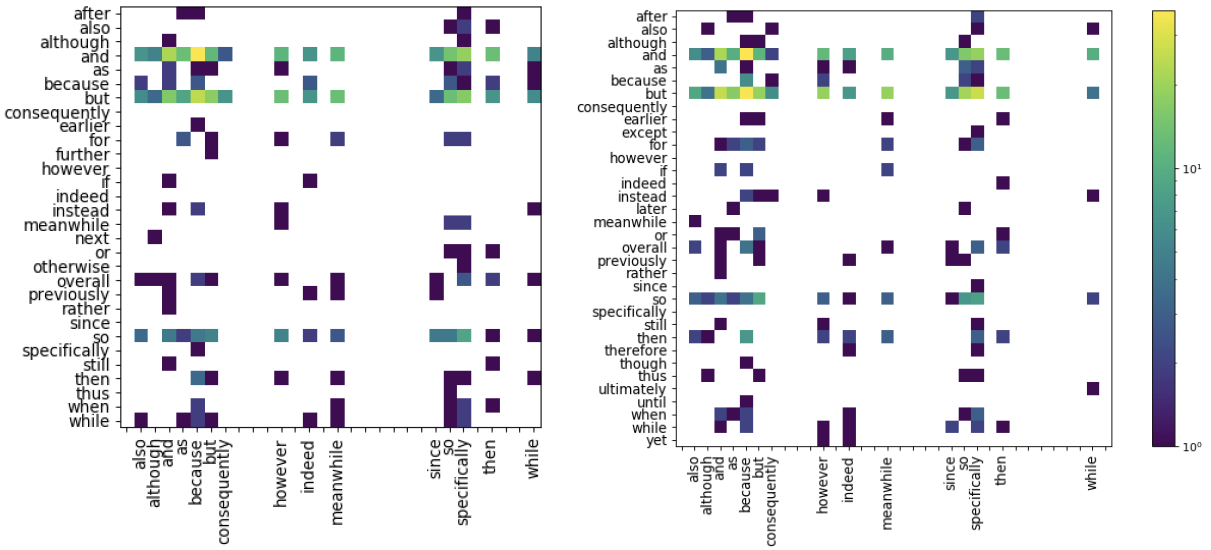
Model	Conn	Sense
always 'and'	9.38	53.15
always 'but'	7.00	13.96
BERT-base-cased	14.43	38.43
BERT-base-uncased	14.85	43.40
BERT-large-cased-wwm	**19.61**	**48.09**
BERT-large-cased	15.69	43.31
BERT-large-uncased-wwm	16.67	45.98
BERT-large-uncased	16.39	46.85
DistilBERT-base-cased	13.45	35.18
RoBERTa-base	14.85	39.39
RoBERTa-large	17.23	46.08
GPT2	13.87	35.37
GPT2-large	14.43	39.58
GPT2-XL	14.85	39.48

Table 3: The agreement in percent of the language models for connective and sense prediction (see text for details). The first two rows show the results when only the respective connectives are predicted for all relations.

5.2 Evaluation of the Language Models via Selected Candidates

In order to investigate how well the language models perform their task, we present in Table 3 the agreement between the human-annotated implicit connective and each model's top-ranked connective[4] (column *Conn*) as well as the agreement between the most frequent sense of that top-ranked connective and the gold sense label (column *Sense*). From the low connective agreement figures, we see

[4]We limit this analysis only to the relations annotated with an one-word gold implicit connective due to our design criteria (see Section 3.1).



(a) RoBERTa-large

(b) GPT2-large

Figure 2: The (truncated) confusion matrices between the predicted and gold connectives of the implicit relations in PDTB 2.0 test set. The matrices are confined to relations with one of the most frequent 10 implicit connectives for readability purposes. The x-axis presents the gold connectives whereas the y-axis shows the predictions.

that the models generally fail to prioritize the connective favored by the annotators; yet, as evidenced by the high sense agreement, they are able to select a connective which suits the given context and thereby helps the explicit relation classifier. We further illustrate the connective predictions of the top language models from each family (RoBERTa-large and GPT2-large) via confusion matrices in Figure 2. As can be seen, the connective predictions are very scattered showing that language models struggle to predict annotators' decisions. However, we would like to note that matching human annotators' performance in connective insertion does not yield informative insights due to ambiguity; that is, for many implicit relations, there are multiple connectives that work as fine. Therefore, we suggest the evaluation focusing on the sense conveyed by the implicit relation and the connective (column *Sense*) as a more reliable way to assess the language models' performance.

too harsh a criteria to assess the language models since in many cases, there are more than one possible connectives that work as fine. Therefore, we would like to note that the second evaluation, matching the sense

Table 3 also suggests that BERT-based models perform better when it comes to selecting a suitable connective than the GPT2 family. We hypothesize that this is because bidirectional gap-filling language models have a training objective that is very close to the type of candidates we use. Finally, despite yielding the worst results, DistilBERT can retain most of BERT-base's performance ($\sim 97\%$), proving that even the smaller models can be utilized for the current task.

5.3 Cross-domain Evaluation

The limited number of the manual annotations does not account for the whole data bottleneck problem in discourse parsing, as the available corpora lack textual variety as much as numbers. Inarguably, PDTB is used as both the training and validation data in the bulk of studies; hence, most research on discourse parsing is confined to one domain. Unfortunately, initial attempts show that sub-tasks of discourse parsing generalize poorly across-domains (Stepanov and Riccardi, 2014).

In order to test how our pipeline generalizes to another domain, we run a set of experiments on the Biomedical Discourse Relation Bank (BioDRB) (Prasad et al., 2011). BioDRB closely follows the PDTB 2.0 annotation framework[5] and is annotated over 24 full-text articles in the biomedical domain which is quite different from that of PDTB. Probably due to this difference and its relatively smaller size, BioDRB is mostly overlooked in computational studies. Consequently, there are only few

[5] Yet, BioDRB uses slightly different sense hierarchy. We follow the instructions on (Prasad et al., 2011) to map the senses back to PDTB 2.0 hierarchy.

	Test set				Full Data			
	4-way		11-way		4-way		11-way	
	Acc	F1	Acc	F1	Acc	F1	Acc	F1
Bi-LSTM baseline	-	-	32.97	-	-	-	-	-
(Bai and Zhao, 2018)	-	-	29.52	-	-	-	-	-
MaxEnt baseline	58.44	26.64	-	-	-	-	-	-
(Shi and Demberg, 2019b)	77.34	43.03	45.19	-	-	-	-	-
BERT-base-uncased	54.15	30.29	36.98	14.59	54.90	36.30	33.80	13.99
+ Margin	52.11	30.15	36.69	15.41	55.15	37.46	35.75	14.59
BERT-large-cased	75.37	26.51	37.12	10.29	72.28	30.11	32.19	8.28
+ Margin	70.57	25.62	34.53	10.74	68.69	31.21	31.82	10.07
BERT-large-cased-wwm	62.36	24.59	32.95	10.87	65.36	33.59	31.81	11.39
+ Margin	56.99	24.79	31.37	11.18	59.83	33.10	31.20	11.90
BERT-large-uncased	58.05	30.43	35.25	12.82	57.32	36.21	34.23	13.85
+ Margin	57.24	31.84	37.99	15.58	57.01	37.73	35.23	14.54
BERT-large-uncased-wwm	61.22	32.24	**38.27**	15.29	60.05	37.49	34.58	14.09
+ Margin	51.95	30.39	36.83	15.62	53.98	37.03	34.40	14.55
DistilBERT-base-cased	39.51	23.62	21.44	11.77	41.21	28.93	21.78	10.47
+ Margin	40.00	27.78	25.32	14.97	38.35	30.43	23.89	11.56
GPT2	59.11	24.36	30.94	11.29	62.85	32.37	29.82	10.44
+ Margin	58.86	24.53	30.36	12.15	62.12	33.69	30.66	11.62
GPT2-large	62.85	29.70	36.69	**19.17**	62.47	36.59	33.08	12.97
+ Margin	60.81	29.48	34.82	15.18	61.91	38.33	33.86	13.61
GPT2-XL	58.86	**33.54**	35.11	16.23	59.19	39.86	34.22	14.75
+ Margin	56.75	33.17	34.53	12.54	59.19	**41.28**	**35.33**	**15.25**
RoBERTa-base	**78.70**	29.70	37.84	12.92	**74.73**	33.52	33.45	10.22
+ Margin	78.05	28.83	37.41	13.55	74.67	34.31	34.13	10.98
RoBERTa-large	71.38	28.44	37.84	13.21	71.26	35.77	32.42	11.25
+ Margin	70.98	28.46	38.13	13.49	71.42	37.71	33.70	12.93

Table 4: The results of the cross-domain experiments on BioDRB set. Test set refers to the results on the designated test set of BioDRB whereas Full data is the whole corpus. All baselines are supervised and their results are taken from (Shi and Demberg, 2019b).

results on BioDRB and unsurprisingly they are all from supervised methods. We compare our results with (Shi and Demberg, 2019b) which reports the state-of-the-art cross-domain results, along with the results from a number of baselines. For the sake of comparability, we follow their experimental settings and report both 4- and 11-way classification results on the BioDRB test set[6].

Additionally, as a more rigorous evaluation, we also report results on the whole BioDRB corpus. That way, we aim to free the evaluation of the generalization abilities of our pipeline from any bias that may rise from using a certain sub-part of the corpus. Finally, it must be noted that the LMs are not fine-tuned in any way on the target corpus (BioDRB) in either setting. The results are provided in Table 4.

The results suggest that our pipeline has strong cross-domain performance despite explicit relation classifier's being trained on only PDTB. In both 4-way and 11-way classification, we are able to outperform the zero-shot performance of even the supervised approaches, including the recent neural approaches (Bai and Zhao, 2018). We hypothesize that our two-step pipeline plays the key role in mitigating the domain-specific problems. Since we are using the "raw" (unfinetuned) language models to rank candidates, we are able to directly leverage the knowledge of these models that they learn from numerous domains thanks to their diverse training

[6]which is originally suggested by (Xu et al., 2012) and consists of the files *GENIA_1421503* and *GENIA_1513057*

data. Once the suitable connectives are highlighted by the language model, the explicit relation classifier can mainly rely on them to make the prediction; hence, less affected by the domain change.

6 Conclusions

In addition to its inherent difficulty, implicit discourse relation classification becomes even more challenging with the lack of sufficient data. In the current study, we focus on the latter problem by assuming the extreme low-resource scenario where there are no labeled implicit discourse relations. The data shortage is mitigated by leveraging the contextual information of the available pre-trained language models through explicitation of the implicit relations. We show that the proposed pipeline, despite its simplicity, is able to outperform the previous attempts. Furthermore, by taking another step, we tested the proposed architecture in the more challenging 11-way setting as well as on a completely different domain. The experimental results confirm that our model is robust and generalizes well, even compared to recent supervised approaches.

Acknowledgments

We would like to thank Dmitry Nikolaev, Johan Sjons, Bernhard Wälchli and Faruk Büyüktekin for their useful comments. The three outstanding reviews from the workshop also helped us greatly. We thank NVIDIA for their GPU grant, and the Swedish National Infrastructure For Computing (SNIC) for providing computational resources under Project 2020/33-26.

References

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, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Chloé Braud and Pascal Denis. 2014. Combining natural and artificial examples to improve implicit discourse relation identification. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1694–1705, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT (1)*.

Hsin-Ping Huang and Junyi Jessy Li. 2019. Unsupervised adversarial domain adaptation for implicit discourse relation classification. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 686–695, Hong Kong, China. Association for Computational Linguistics.

Yangfeng Ji, Gongbo Zhang, and Jacob Eisenstein. 2015. Closing the gap: Domain adaptation from explicit to implicit discourse relations. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2219–2224.

Najoung Kim, Song Feng, Chulaka Gunasekara, and Luis Lastras. 2020. Implicit discourse relation classification: We need to talk about evaluation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5404–5414, Online. Association for Computational Linguistics.

Man Lan, Jianxiang Wang, Yuanbin Wu, Zheng-Yu Niu, and Haifeng Wang. 2017. Multi-task attention-based neural networks for implicit discourse relationship representation and identification. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1299–1308, Copenhagen, Denmark. Association for Computational Linguistics.

Yang Liu, Sujian Li, Xiaodong Zhang, and Zhifang Sui. 2016. Implicit discourse relation classification via multi-task neural networks. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, pages 2750–2756.

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*.

Ilya Loshchilov and Frank Hutter. 2018. Fixing weight decay regularization in adam. *arXiv preprint ArXiv:1711.05101*.

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.

Allen Nie, Erin Bennett, and Noah Goodman. 2019. DisSent: Learning sentence representations from explicit discourse relations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4497–4510, Florence, Italy. Association for Computational Linguistics.

Emily Pitler, Annie Louis, and Ani Nenkova. 2009. Automatic sense prediction for implicit discourse relations in text. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the*

4th International Joint Conference on Natural Language Processing of the AFNLP, pages 683–691.

Emily Pitler and Ani Nenkova. 2009. Using syntax to disambiguate explicit discourse connectives in text. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, pages 13–16, Suntec, Singapore. Association for Computational Linguistics.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind K Joshi, and Bonnie L Webber. 2008. The penn discourse treebank 2.0. In *In Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC)*.

Rashmi Prasad, Susan McRoy, Nadya Frid, Aravind Joshi, and Hong Yu. 2011. The biomedical discourse relation bank. *BMC bioinformatics*, 12(1):1–18.

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)*, pages 1006–1017, Vancouver, Canada. Association for Computational Linguistics.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Attapol Rutherford and Nianwen Xue. 2015. Improving the inference of implicit discourse relations via classifying explicit discourse connectives. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 799–808, Denver, Colorado. Association for Computational Linguistics.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

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 5790–5796, Hong Kong, China. Association for Computational Linguistics.

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.

Evgeny Stepanov and Giuseppe Riccardi. 2014. Towards cross-domain pdtb-style discourse parsing. In *Proceedings of the 5th International Workshop on Health Text Mining and Information Analysis (Louhi)*, pages 30–37.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Changxing Wu, Xiaodong Shi, Yidong Chen, Yanzhou Huang, and Jinsong Su. 2016. Bilingually-constrained synthetic data for implicit discourse relation recognition. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2306–2312, Austin, Texas. Association for Computational Linguistics.

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.

Deniz Zeyrek and Murathan Kurfalı. 2017. TDB 1.1: Extensions on Turkish discourse bank. In *Proceedings of the 11th Linguistic Annotation Workshop*, pages 76–81, Valencia, Spain. Association for Computational Linguistics.

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

Šárka Zikánová, Jiří Mírovskỳ, and Pavlína Synková. 2019. Explicit and implicit discourse relations in the prague discourse treebank. In *International Conference on Text, Speech, and Dialogue*, pages 236–248. Springer.