Shallow Discourse Parsing on Twitter Conversations

Berfin Aktaş, Burak Özmen
UFS Cognitive Science, University of Potsdam, Germany
berfinaktas@uni-potsdam.de,ozmen.brk@gmail.com

Abstract

We present our PDTB-style annotations on conversational Twitter data, which was initially annotated by Scheffler et al. (2019). We introduced 1,043 new annotations to the dataset, nearly doubling the number of previously annotated discourse relations. Subsequently, we applied a neural Shallow Discourse Parsing (SDP) model to the resulting corpus, improving its performance through retraining with in-domain data. The most substantial improvement was observed in the sense identification task (+19%). Our experiments with diverse training data combinations underline the potential benefits of exploring various data combinations in domain adaptation efforts for SDP. To the best of our knowledge, this is the first application of Shallow Discourse Parsing on Twitter data.

Keywords: shallow discourse parsing, Twitter, discourse relations, PDTB

1. Introduction

Discourse parsing, the identification of discourse relations between text spans, has seen substantial advancements in recent years. However, a significant challenge arises when the parsers are tested on a different domain, as recent research (Scholman et al., 2021; Liu and Zeldes, 2023) demonstrates a notable degradation in their performance. Consequently, the need for additional resources for discourse relations in diverse genres becomes increasingly important.

Penn Discourse Tree Bank (PDTB) refers to both the largest corpus, composed of news texts, annotated for shallow discourse relations and to the framework describing the annotation of these relations. The dataset (Prasad et al., 2018) is composed of written news texts (from Wall Street Journal). The main purpose of PDTB-style annotation is to identify two (mostly consecutive) arguments Arg1 and Arg2 which are semantically related. This relation can be constructed via explicitly expressed discourse connectives (i.e., an explicit relation) or can be inferred implicitly (i.e., an implicit relation).

There exist studies applying the PDTB framework to a variety of formal and informal spoken texts (Tonelli et al., 2010; Rehbein et al., 2016; Riccardi et al., 2016; Crible and Cuenca, 2017). These studies show that the use of discourse connectives and relations differs significantly between written and spoken data. Scheffler et al. (2019) conduct a pilot study on conversational Twitter data, where they annotated a corpus of Twitter Conversations (henceforth *TwiConv*) for explicit intra-tweet relations (i.e., the connective and arguments are in the same tweet). Their analysis indicates that Twitter conversations resemble spoken texts in terms of discourse relations. Nevertheless, there is still a noticeable gap in research focusing on interaction on social media, which remains a relatively unexplored area.

Our primary contribution (Sections 2 and 3) is tackling this challenge through the expansion of the initial annotations put forth by Scheffler et al. (2019). In addition to existing explicit intra-tweet annotations (Example 1[1]), we include in our annotations [**A**] explicit inter-tweet relations (i.e., arguments of the relation are located on different tweets, mostly posted by different users, as in Example 2), as well as [**B**] all implicit (Example 3) and [**C**] hypophora relations (i.e., question-answer pairs in the text as in Example 4).[2]

(1) *Black folks in Alabama organized.* And **WON!** [Single Tweet]

(2) Tweet1:*Like I said, you don't know the whole situation to make such a judgement.*
Tweet2: And **until you have raised one yourself, sit down and shut up!**

(3) Tweet1: [..] *Time is short!!!*
Tweet2: **Not as short as your career highlights.** [..]

(4) Tweet1: *Higher than a the office of a Governor?? Or he's talking of the offices when turned upside down?*
Tweet2: **A speaker is higher than the governor**

Our second contribution (Section 4) is the first, to the best of our knowledge, application of shallow discourse parsing on Twitter data. We apply

[1] In the examples given in this paper, first argument (Arg1) in a discourse relation is marked by *italics letters*, second argument (Arg2) by **bold letters** and connectives by underlining.

[2] Annotations are available here: https://github.com/berfingit/TwiConv-discourse-relations

ISA-20 Proceedings @LREC-COLING-2024, pages 60–65

domain adaptation by retraining a state-of-the-art neural shallow discourse parsing model (Knaebel, 2021), using the annotations we generated.

2. Discourse Relations in TwiConv

2.1. Data

The TwiConv corpus contains English language tweets collected from the Twitter stream on several (non-adjacent) days in December 2017 and January 2018 without filtering for hashtags or topics. Conversations are gathered by recursively obtaining parent tweets, whose IDs were derived from the in_reply_to_id field of the tweet objects returned by the former Twitter API. For specifics regarding the data collection, refer to Aktaş and Kohnert (2020).

TwiConv comprises 1756 tweets, posted by 594 distinct users.[3] Tweets are organized into 185 conversation threads[4], with an average tweet length of 153 characters. The threads vary in length from 3 to 78 tweets, with an average length of 10 tweets and a median of 7. There are 48,172 tokens in TwiConv.

2.2. Annotation Procedure

Annotations were conducted by a linguistics undergraduate student. We built upon the guidelines devised by Scheffler et al. (2019), further extending them to encompass the additional relations we annotated. Additionally, we refined the instructions for selecting argument spans to enhance clarity for our annotators. Annotations were marked with the PDTB annotator tool (Lee et al., 2016). We followed the PDTB-3 scheme for annotations.

The PDTB-3 framework uses a 3-level hierarchy for the semantic categorization of relations (i.e., through sense labels), where at the top level is the "class" label, distinguishing between EXPANSION, COMPARISON, CONTINGENCY, and TEMPORAL relations. Level-2 and level-3 in the sense hierarchy represent the fine-grained labels refining the semantics of the class. There are a total of 36 categories available for assignment as sense labels. For more details on the PDTB sense hierarchy, see Webber et al. (2018).

2.3. Inter-annotator Agreement

We conducted an Inter-annotator Agreement (IAA) study on a subcorpus of 20 randomly chosen threads. They comprise 267 tweets with an average length of 187 characters. A second linguistics student annotated them for the IAA computation. Following earlier PDTB studies (e.g., Prasad et al. (2008); Rehbein et al. (2016)), we report percent agreement for explicit relations on the sense assignments, Arg1 and Arg2 span selection, and for implicit relations on their senses.

The agreement on argument spans for **explicit** relations (Table 1) was notably high, surpassing those reported by Scheffler et al. (2019). This improvement is likely due to our less ambiguous span selection guidelines for social media symbols such as hashtags, links, and emoticons.

Type	Exact	Partial
Connective Detection	71%	-
Arg1 Span	79%	93%
Arg2 Span	95%	97%

Table 1: IAA for explicit relation text spans

Only the **implicit** relations annotated by both annotators were examined in this IAA study. We defined an implicit relation as shared between the two if both annotators identified an implicit relation with exactly matching argument spans. As a result, the argument spans (Arg1 and Arg2) for the implicit relations we analyzed always aligned. Therefore, our agreement analysis focused solely on the sense assignments for these shared implicit relations. Specifically, the first annotator identified 169 implicit relations, of which 126 shared argument spans with those identified by the second annotator. Hence, our agreement analysis is based on these 126 common implicit relations.

Table 2 presents the sense agreement statistics. IAA for implicit relations is generally lower compared to explicit relations, as found in existing literature (Prasad et al., 2008; Zeyrek and Kurfalı, 2017; Zikánová et al., 2019; Hoek et al., 2021). Our statistics confirm the acknowledged difficulty in annotating implicit relations. Additionally, we argue that annotating implicit relations is particularly challenging in Twitter conversations due to the text ambiguity resulting from Twitter's character limit (280 characters during data collection) and the non-standard items (e.g., hashtags, abbreviations, and images) in tweets.

Sense Level	Explicit	Implicit
Level-1	88%	68%
Level-2	82%	45%
Level-3	76%	41%

Table 2: IAA for sense annotations

In Table 3 we present the most common disagreements in implicit relation senses between the annotators. Scholman et al. (2022) allow annotation of multiple senses and then determine the senses that

[3] In the conversations, it is possible for a single user to respond multiple times.

[4] A set of tweets consisting of one or more users replying to each other is called a *thread* in our terminology.

frequently occur together (p. 3287). We observe that the pair exhibiting the highest co-occurrence frequency in their study (*Conjunction* and *Result*) is identical to the one found in our disagreement matrix. Additionally, the pairing of *Arg2-as-detail* and *Conjunction* is another prevalent combination in both statistics. This suggests that our disagreements might correspond with the observations by Scholman et al. (2022), highlighting the inherent ambiguity of implicit relations and the necessity for implementing multi-sense annotation.

Sense1	Sense2	Percentage
Conjunction	Result	9.7%
Belief.Reason	Reason	6.9%
Conjunction	Arg2-as-detail	5.6%
Contrast	Arg2-as-denier	5.6%
Conjunction	Reason	4.2%
Conjunction	Arg2-as-subst	4.2%

Table 3: Most common disagreements in sense assignments in implicit relation annotations

2.4. Quantitative Analysis

The annotations comprise a total of 2281 discourse relations, with 1237 originating from the prior annotations of Scheffler et al. (2019). Within the full set, 1433 are explicit relations, 732 are implicit relations, and the remaining 116 are hypophora relations.

We observe that explicit discourse relations are a frequent occurrence in our Twitter data. Out of 1756 tweets, 47% contain at least one discourse connective, and 22% contain more than one (up to 6). A tweet with 6 connectives is given in Example 5.

(5) Yes, but if it were true and she has decided to run in 2020, it gives more people something to rally behind, a reason to get out and vote this year, a Democratic Congress when she arrives! I'm all in, and think an Oprah run would greatly help in 2018 Mid Terms! #Oprah2020

Table 4 shows the distribution of intra- and inter-tweet relations. The majority of Explicit and Implicit relations occur within a single tweet, whereas Hypophora relations are typically inter-tweet relations. 98.5% of the inter-tweet relations span into two tweets, as illustrated in examples 3 and 4 for an implicit relation and an hypophora relation, respectively; but there also exist relation instances that span into three tweets (1.5%). Inter-tweet relations typically occur between tweets posted by different users (81%) but they also exist between tweets posted by the same user (19%).[5]

[5]A comparison of relations established by the same user and by different users is left to future work.

Relation Type	intra-tweet	inter-tweet
Explicit	90%	10%
Implicit	88%	12%
Hypophora	4%	96%

Table 4: Intra- and inter-tweet relation distributions (All relations except intra-tweet Explicit relations have been annotated by our team.)

3. TwiConv vs PDTB 3.0

Table 6 shows the distribution of the level-1 relation senses in our corpus and in the PDTB corpus (Prasad et al., 2019). Our Twitter data has substantially more CONTINGENCY relations than the PDTB. In line with this observation, connectives expressing CONTINGENCY relations like *if*, *when*, *because* and *so* occur relatively more frequently on Twitter as shown in Table 5. During our annotation process, we noticed that longer threads often represent argumentative discussions, and the prevalence of CONTINGENCY connectives can serve as evidence for this: Users provide substantiation for their arguments. In contrast, news texts in PDTB use more narrative (TEMPORAL) and EXPANSION relations.

Connective	TwiConv	Connective	PDTB
and	27.6%	and	26.3%
but	15.9%	but	15.2%
if	7.9%	also	7.1%
so	6.6%	if	4.7%
when	6.2%	when	4.3%
because	5.7%	while	3.3%
or	2.8%	as	3.3%
also	2.8%	because	3.1%
as	2.2%	after	2.1%
then	1.8%	however	2%

Table 5: Top ten connectives in the TwiConv and PDTB-3 explicit relations

Regarding the implicit/explicit difference, in the TwiConv corpus, CONTINGENCY relations are more often realized implicitly, whereas TEMPORAL relations are more often explicit (like in PDTB). In PDTB, COMPARISON relations are much more often explicit (25% vs 11%) whereas in the TwiConv data, both relation types have similar proportion.

Finally, we briefly look at patterns regarding spoken vs. written differences. Crible and Cuenca (2017) argue that discourse markers in spoken genres are more multi-functional than in written genres, which indicates greater diversity within spoken genres, particularly in the sense distributions of certain connectives. Here, we compared the

Class	Relation	TwiConv	PDTB
EXPANSION	All	32%	44%
EXPANSION	Explicit	33%	42%
EXPANSION	Implicit	30%	46%
CONTINGENCY	All	34%	25%
CONTINGENCY	Explicit	29%	16%
CONTINGENCY	Implicit	43%	35%
COMPARISON	All	24%	18%
COMPARISON	Explicit	25%	25%
COMPARISON	Implicit	23%	11%
TEMPORAL	All	10%	13%
TEMPORAL	Explicit	13%	17%
TEMPORAL	Implicit	4%	8%

Table 6: Level-1 sense distributions for TwiConv and PDTB 3.

level-1 sense annotations for "and" which is the most frequent connective in both corpora. Table 7 reveals that it is used to establish TEMPORAL relations (as illustrated in Example 6) in 8.2% of explicit relations in TwiConv, but is not used for that purpose in PDTB. Tonelli et al. (2010) had observed a similar pattern in their dialog annotation in Italian, where the connective "e" ("and") can express TEMPORAL as well as EXPANSION relations. Furthermore, in TwiConv, the COMPARISON relations established by "and" are much more common than in PDTB (5.7% vs 0.03%). This supports the idea that TwiConv represents patterns of spoken language in terms of connective functionality, which we plan to study further in future work.

Class	TwiConv	PDTB
COMPARISON	5.7%	0.3%
CONTINGENCY	4.0%	2.7%
EXPANSION	82.2%	97%
TEMPORAL	8.2%	-

Table 7: Level-1 sense distributions for "and" (case insensitive)

(6) [..] *I'm going to create a totally new arbitrary number* and **assign meaning to it**.

4. Shallow Discourse Parsing (SDP) on Twitter Conversations

Experiments. Our experiments utilize the neural shallow discourse parser "*discopy*", which was introduced by Knaebel (2021). The *discopy* model achieves state-of-the-art results in connective identification, and also demonstrates competitive performance in other SDP tasks, notably in Arg1 identification. The experimental design was the one proposed at the CoNLL Shared Task 2016 (Xue et al., 2016), and the reported results conform to that.

The main goal of our work is to assess whether incorporation of Twitter Conversation data into the training data of *discopy* affects the performance of the model when tested on TwiConv. To accomplish this, we segment our TwiConv data into training, testing, and validation sets with the distribution of 80%, 10%, and 10% of data, respectively.

We then combine the TwiConv training set with different portions of PDTB data from the CoNLL 2016 Shared Task (Xue et al., 2016), which consists of 930k tokens and has been employed to train the original *discopy* model. These combinations encompass varied token quantities from the PDTB data, allowing us to manipulate the proportion of TwiConv data in the training set. We establish four distinct setups:

- setup 1 (only PDTB)
- setup 2 (30k tokens PDTB + TwiConv)
- setup 3 (465k tokens PDTB + TwiConv)
- setup 4 (complete PDTB + Twiconv)

We conduct experiments in these setups with both RoBERTa- and BERT-base embeddings, and we show the results in Table 8. (We only present the best scores for the sake of simplicity.)

We also implemented preprocessing steps on TwiConv, which involve eliminating URLs, poster handles, mentions, and transforming hashtags into complete words. For instance, '#ClintonFoundation' was changed to 'Clinton Foundation'. The results for the same setups with the preprocessed data are also provided in Table 8.

Results. Our baseline consists of parsing our test set with the *discopy* model trained solely on PDTB data (i.e., setup 1). We achieved our best results with RoBERTa-base for that setting, so we have adopted it as our baseline. It shows a substantial drop when run on the Twitter data, losing almost 50% of the results reported by Knaebel (2021) for PDTB parsing.

We obtained the best results for most of the metrics with BERT-base with setup 4, which improves over the baseline in almost all cases, including a 6% increase in connective identification. With the preprocessed data, we obtained the best results in setup 4 for most of the metrics with RoBERTa-base.

Discussion. Incorporating Twitter data into the training set generally proves useful; however, there is no universal configuration that consistently outperforms the other setups across all metrics. In most cases, an increase in the volume of PDTB training data leads to metric enhancements, although exceptions exist. For instance, the most

Setup	$F1_{conn}$	$F1_{Arg1}$	$F1_{Arg2}$	$F1_{Sense}$
Baseline-rb	0.46	0.25	0.38	0.32
setup 3-rb	**0.52**	0.24	0.37	0.37
setup 4-rb	0.51	0.25	0.33	0.39
setup 3-bb	**0.52**	**0.28**	0.34	**0.49**
setup 4-bb	**0.52**	0.27	**0.39**	**0.49**
setup 2-rb^p	**0.52**	0.17	0.29	0.33
setup 4-rb^p	0.49	**0.3**	**0.37**	**0.51**
setup 3-bb^p	0.51	0.29	**0.37**	0.41
setup 4-bb^p	0.51	0.28	**0.37**	0.38

Table 8: Performance of *discopy* on the TwiConv test set, with RoBERTa-base (rb) and BERT-base (bb). We use strict measuring according to (Knaebel, 2021), i.e., a 0.9 threshold for overlap. The "p" superscript signifies experiments conducted on preprocessed data.

favorable result for connective identification on preprocessed data (0.52) emerges when TwiConv is integrated with a relatively small portion (30K) of PDTB data. This highlights the significance of experimenting with various data combinations in domain adaptation efforts, depending on the SDP subtask that is most relevant for a downstream purpose.

When evaluating the optimal outcomes, it is evident that connective (+6%) and Arg1 identification (+5%) shows notable improvements through retraining. Sense identification exhibits improvements across nearly all configurations compared to the baseline, with a remarkable (19%) improvement when the data is preprocessed. On the other hand, Arg2 identification shows minimal benefits and, in most cases, becomes worse, with the best scenario yielding only a modest (1%) improvement. The average improvement in preprocessed results is only marginally superior to the outcomes attained using BERT-base on non-preprocessed data.

5. Conclusions

We introduced non-explicit (implicit and hypophora) and inter-tweet explicit relations to the TwiConv corpus, which was initially annotated by Scheffler et al. (2019) for intra-tweet explicit relations, almost doubling the amount of original annotations. Subsequently, we applied a neural Shallow Discourse Parsing model to the dataset, enhancing the model's performance on TwiConv data through retraining. We conducted experiments utilizing both BERT and RoBERTa embeddings, and the best results were obtained using BERT on the unprocessed data. This resulted in improvements across all tasks, except for Arg2 identification, which presents an interesting case requiring further investigation. Extensive preprocessing of the Twitter data results in only marginal improvements.

6. Acknowledgements

We thank the anonymous reviewers and Manfred Stede for their valuable observations and suggestions. This work is funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project-ID 317633480 – SFB 1287.

7. Bibliographical References

Berfin Aktaş and Annalena Kohnert. 2020. TwiConv: A coreference-annotated corpus of Twitter conversations. In *Proceedings of the Third Workshop on Computational Models of Reference, Anaphora and Coreference*, pages 47–54, Barcelona, Spain (online). Association for Computational Linguistics.

Ludivine Crible and Maria Josep Cuenca. 2017. Discourse markers in speech: characteristics and challenges for corpus annotation. *Dialogue and Discourse*, 8(2):149–166.

Jet Hoek, Merel C.J. Scholman, and Ted J.M. Sanders. 2021. Is there less agreement when the discourse is underspecified? In *Proceedings of the DiscAnn Workshop*.

René Knaebel. 2021. discopy: A neural system for shallow discourse parsing. In *Proceedings of the 2nd Workshop on Computational Approaches to Discourse*, pages 128–133, Punta Cana, Dominican Republic and Online. Association for Computational Linguistics.

Alan Lee, Rashmi Prasad, Bonnie Webber, and Aravind K. Joshi. 2016. Annotating discourse relations with the PDTB annotator. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: System Demonstrations*, pages 121–125, Osaka, Japan. The COLING 2016 Organizing Committee.

Yang Janet Liu and Amir Zeldes. 2023. Why can't discourse parsing generalize? a thorough investigation of the impact of data diversity. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3112–3130, Dubrovnik, Croatia. Association for Computational Linguistics.

R. Prasad, N. Dinesh, A. Lee, E. Miltsakaki, L. Robaldo, A. Joshi, and B. Webber. 2008. The

Penn Discourse Treebank 2.0. In *Proc. of the 6th International Conference on Language Resources and Evaluation (LREC)*, Marrakech, Morocco.

Rashmi Prasad, Bonnie Webber, and Alan Lee. 2018. Discourse annotation in the PDTB: The next generation. In *Proceedings 14th Joint ACL - ISO Workshop on Interoperable Semantic Annotation*, pages 87–97, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Ines Rehbein, Merel Scholman, and Vera Demberg. 2016. Annotating discourse relations in spoken language: A comparison of the PDTB and CCR frameworks. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC)*.

Giuseppe Riccardi, Evgeny A. Stepanov, and Shammur Absar Chowdhury. 2016. Discourse connective detection in spoken conversations. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6095–6099.

Tatjana Scheffler, Berfin Aktaş, Debopam Das, and Manfred Stede. 2019. Annotating shallow discourse relations in Twitter conversations. In *Proceedings of the Workshop on Discourse Relation Parsing and Treebanking 2019*, pages 50–55, Minneapolis, MN. Association for Computational Linguistics.

Merel Scholman, Tianai Dong, Frances Yung, and Vera Demberg. 2021. Comparison of methods for explicit discourse connective identification across various domains. In *Proceedings of the 2nd Workshop on Computational Approaches to Discourse*, pages 95–106, Punta Cana, Dominican Republic and Online. Association for Computational Linguistics.

Merel Scholman, Tianai Dong, Frances Yung, and Vera Demberg. 2022. DiscoGeM: A crowdsourced corpus of genre-mixed implicit discourse relations. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3281–3290, Marseille, France. European Language Resources Association.

Sara Tonelli, Giuseppe Riccardi, Rashmi Prasad, and Aravind Joshi. 2010. Annotation of discourse relations for conversational spoken dialogs. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Bonnie Webber, Rashmi Prasad, Alan Lee, and Aravind Joshi. 2018. The Penn Discourse Treebank 3.0 Annotation Manual. Report, The University of Pennsylvania.

Nianwen Xue, Hwee Tou Ng, Sameer Pradhan, Attapol Rutherford, Bonnie Webber, Chuan Wang, and Hongmin Wang. 2016. CoNLL 2016 shared task on multilingual shallow discourse parsing. In *Proceedings of the CoNLL-16 shared task*, pages 1–19, Berlin, Germany. Association for Computational Linguistics.

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.

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

8. Language Resource References

Prasad et al. 2019. *Penn Discourse Treebank Version 3.0*. University of Pennsylvania. distributed via Linguistic Data Consortium: LDC2019T05, ISLRN 977-491-842-427-0.