Annotating Shallow Discourse Relations in Twitter Conversations

Tatjana Scheffler[1], **Berfin Aktaş**[1], **Debopam Das**[2], **and Manfred Stede**[1]

[1]Department of Linguistics / SFB1287, University of Potsdam, Germany
[2]Department of English and American Studies, Humboldt University of Berlin, Germany
tatjana.scheffler@uni-potsdam.de

Abstract

We introduce our pilot study applying PDTB-style annotation to Twitter conversations. Lexically grounded coherence annotation for Twitter threads will enable detailed investigations of the discourse structure of conversations on social media. Here, we present our corpus of 185 threads and annotation, including an inter-annotator agreement study. We discuss our observations as to how Twitter discourses differ from written news text wrt. discourse connectives and relations. We confirm our hypothesis that discourse relations in written social media conversations are expressed differently than in (news) text. We find that in Twitter, connective arguments frequently are not full syntactic clauses, and that a few general connectives expressing EXPANSION and CONTINGENCY make up the majority of the explicit relations in our data.

1 Introduction

The PDTB corpus (Prasad et al., 2008) is a well-known resource of discourse-level annotations, and the general idea of lexically signalled discourse structure annotation has over the years been applied to a variety of languages. A shallow approach to discourse structure in the PDTB style can also be adapted to different genres. In this paper, we consider English conversations on Twitter, and describe the first phase of our annotation, viz. that of explicit connectives whose arguments are within a single tweet. We explain the collection of the data and our annotation procedure, and the results of an inter-annotator-agreement study. We present our analysis of the specific features of this genre of conversation wrt. discourse structure, as well as corpus statistics, which we compare to the distributions in the original PDTB corpus.

We show that explicit discourse relations are frequent in English Twitter conversations, and that the distribution of connectives and relations differs markedly from the distribution in PDTB text. In particular, the Twitter threads contain many more CONTINGENCY relations (in particular, conditional and causal relations). In addition, the connective's arguments in the Twitter data are often elliptical phrases standing in for propositional content.

The upcoming second phase of the project will target connectives whose Arg1 is located in a previous tweet, as well as AltLex realizations and implicit relations. We regard this effort as complementary to approaches that applied Rhetorical Structure Theory (Mann and Thompson, 1988) and Segmented Discourse Representation Theory (Asher and Lascarides, 2003) to dialogue; the important difference being that other than RST and SDRT, PDTB does not make strong commitments as to an overarching structure of the discourse. Overall, we see this as an advantage for studying relatively uncharted territory: The structural peculiarities of social media conversations have not yet been explored in depth, and a PDTB-style annotation is one way of laying empirical groundwork for that endeavour.

2 Twitter and Discourse Relations

Recent studies indicate significant differences in the use of discourse connectives and discourse relations between written and spoken data (Rehbein et al., 2016; Crible and Cuenca, 2017). Though the PDTB approach has been applied to different text types, conversational data has not been systematically analysed yet. There is recent work on annotating spoken TED talks in several languages (Zeyrek et al., 2018), but these planned monologues do not exhibit spontaneous interaction. To our knowledge, only (Tonelli et al., 2010; Riccardi et al., 2016) have constructed PDTB annotations

Proceedings of Discourse Relation Parsing and Treebanking (DISRPT2019), pages 50–55
Minneapolis, MN, June 6, 2019.

for spoken conversations, and they work on Italian dialogs. Riccardi et al. (2016) focus on the detection of discourse connectives from lexical and acoustic features in help desk dialogs. In contrast, we investigate open topic spontaneous conversations in computer mediated communication, to abstract away from the speech mode, but retain the conversational properties.

Twitter[1] is a social media platform that publishes short "microposts" by registered users. In addition to textual content, these posts may contain embedded images or videos. Twitter users can interact by directly replying to each other's messages. Such replies are quite frequent and the resulting conversations often contain discourse connectives (Scheffler, 2014). There is some evidence that the types of relations and connectives found in Twitter conversations differs markedly from edited news text and reflects some features of spoken conversations (Scheffler, 2014; Scheffler and Stede, 2016). Here, we introduce an annotated corpus of explicit connectives in English tweets, which allows us to test genre differences between the discourse structure of newspaper texts (PDTB) and conversational writing (our Twitter corpus).

3 Collecting and Annotating the Corpus

We collected English language tweets from the Twitter stream on several (non-adjacent) days in December, 2017 and January 2018 without filtering for hashtags or topics in any way. In order to obtain conversations that are linked to each other via the *reply-to* relation, and which altogether then form a tree structure, we recursively retrieved parent tweets of those gathered via the initial search. A single *thread* in our terminology is a path from the root to a leaf node of that tree. For the purposes of the present experiment, we then selected only one of the longest threads (paths) from each tree and discarded everything else in this dataset. See (Aktaş et al., 2018) for details on the data collection. The resulting corpus consists of 1756 tweets arranged in 185 threads, and the average length of a tweet is 153 characters.[2]

So far, we only annotated explicit connectives whose two arguments are contained within the same tweet (whether a source or reply tweet).[3] We primarily used the list of 100 explicit connectives from the PDTB corpus (Prasad et al., 2008) to identify connectives. Additionally, we found a few new connectives in our corpus, such as *by the way, plus, so long as*, and *when-then*. In practice, we annotate an explicit connective, identify its two arguments in the same tweet in which the connective occurs, and finally, label the connective sense according to the PDTB-3 relational taxonomy (Webber et al., 2018)[4]. In the event we find an ambiguous connective or interpret more than one relational reading, we assign multiple senses to the connective. The annotation was conducted using the PDTB annotator tool.

Inter-Annotator Agreement. After one author of this paper labeled the dataset in the way just described, we conducted an Inter-Annotator Agreement (IAA) study on 50 threads selected randomly. This sub-corpus consists of 683 tweets whose average length is 188 characters, and was re-annotated by a research assistant. We calculated the percent agreement for connective detection (i.e. the percentage of connectives marked by both of the annotators), Arg1 and Arg2 span selection, and all levels of sense assignment. Arg1, Arg2 and sense agreements are calculated for the relations annotated by both of the annotators. Table 1 shows the percent agreement for *exact match* and *partial match* of the selected text spans. We consider one character difference in the begin & end indices of text spans as an instance of *exact match* to eliminate disagreements because of the involvement of punctuation at the end or beginning of the texts in marked spans. In *partial match* statistics, in addition to exact matches, the argument spans having any overlapping tokens are also considered matching. We manually inspected all cases of *partial match* and observed that in all cases, one annotator's argument span is fully included in the other annotator's span.

The agreement is generally good, except for exact argument spans for Arg1. The main reason for this is the difficulty in Twitter to determine utterance and clause breaks. There was major disagreement with respect to social media specific items like hashtags and emoji (should they be included in the argument span or not?; see also Section 4).

[1] www.twitter.com

[2] URL strings are excluded, but user names are included in tweet length statistics throughout the paper.

[3] The only exception to this rule is when one message by a single author is split over subsequent adjacent tweets. When the continuation is explicitly marked (e.g., with a '+' symbol at the end of an incomplete tweet), all connectives are annotated (even though the arguments may span across tweets).

[4] We do not annotate other information such as attribution features or supplementary spans for connectives.

Social media text is genuinely more difficult to annotate than news text in this regard, and we will adapt the annotation guidelines accordingly to develop clear instructions for these cases.

Table 2 shows IAA statistics for sense levels[5].

Type	% Exact	%Partial
Connective Detection	70%	-
Arg1 Span	62%	90%
Arg2 Span	89%	92%

Table 1: IAA for text spans.

Sense Level	%
Level-1	88%
Level-2	82%
Level-3	76%

Table 2: IAA for sense annotations.

4 Analysis: Twitter versus WSJ

Qualitative Analysis. As said earlier, Twitter posts, although they are written, are part of interactive conversations. While annotating these posts, we also encountered a number of phenomena that are typically found in spoken registers, and not in written texts. For example, we identified higher numbers of a small set of connectives, such as *and, but* and *when*, that frequently occur in conversations. We rarely annotated (if any) connectives like *since, therefore* or *nevertheless*, which are typically found in formal writing (e.g., newspaper, scientific genre). The most frequent connectives in our corpus and in newspaper text are shown in Tables 3 and 4, respectively.

Twitter texts, like conversations, most often represent spontaneous use of language, and thus contain instances of fragmented or incomplete utterances. In our annotation, we often encounter constructions that comprise only nouns or noun phrases, but nevertheless, are often seen to stand for a complete proposition. These phrases can function as the arguments of a connective. (e.g., "NO PROB *BUT* WHERE THE HELL DID U", or "*If* he could work on that, good prospect"). We

[5]Level-1 specifies four sense classes, TEMPORAL, CONTINGENCY, COMPARISON, and EXPANSION. Level-2 provides 17 sense types, whereas Level-3 encodes only the *directionality* of the sense in the PDTB-3 schema (e.g., REASON vs. RESULT as subtypes of the Level-2 sense type CAUSE).

[6]The instances of "&" are also counted in this category.

[7]The instances of "b4" are also counted in this category.

Connective	% in Twitter
and[6]	30.0%
but	16.2%
if	7.3%
when	6.5%
so	6.0%
because	4.5%
or	2.9%
as	2.2%
also	2.0%
before[7]	1.3%

Table 3: Top ten connectives in the Twitter corpus.

Connective	% in PDTB
and	26.4%
but	15.4%
also	7.2%
if	4.8%
when	4.4%
as	3.5%
because	3.5%
while	3.3%
after	2.4%
however	2.0%

Table 4: Top ten connectives in the PDTB.

accordingly use more flexible argument selection criteria in order to accommodate such (elliptical) structures, in addition to clauses and other constructions (nominalizations, VP-conjuncts, etc.) that typically constitute arguments in the PDTB. Furthermore, similar to the genre of instant messaging, the Twitter texts contain a wide range of acronyms for sentences/clauses that act like fixed expressions. Examples include: "idc" = I don't care; "idk" = I don't know; "idrk" = I don't really know. In our annotation, we pay special attention to these acronyms as to whether they constitute (part of) the argument of a connective or not. For example, *idc* in "idc *if* u do or not" is annotated as an argument (of *if*), while *idk* is not part of the argument in "i get your point... *but* idk the k-exol who he was talking to was comforted...".

Our Twitter data also exhibits different spellings for the same connective, for example, 'wen' = *when*; 'cos', 'cus', 'cuz' = *because*; 'btw' = *by the way*; '&', '&', 'an' = *and*. We considered these alternative forms as orthographical variants of the same connective.

Finally, we find that the parity between Twit-

ter texts and spoken conversations also operates at the level of annotating senses for the connectives. For examples, the additive connective *and* is frequently used in conversations (or spoken texts in general) to link upcoming utterances with the present one, even though there is no strong semantic relation between them (comparable to the Joint relation in RST). We observe similar uses of *and* in our Twitter corpus, too (e.g., "Happy new year *and* cant wait for you to come back to the UK next year").

Quantitative Analysis. We performed a quantitative analysis of our annotations. In general, we observe that explicit discourse relations are a frequent occurrence in our Twitter data. Out of 1756 tweets, over 40% contain at least one tweet-internal explicit discourse relation. Table 3 shows the relative frequency distribution of the top 10 connectives in our annotations. The calculated percentages are case insensitive (e.g., "and" and "And" are considered as different instances of the same connective). Some basic statistics of our annotation:

- # of relations: 1237
- # of tweets with a single connective: 406
- # of tweets with multiple connectives: 329
- average Arg1 length (chars): 43
- average Arg2 length (chars): 41
- average length of tweets with a single connective (chars):181
- average length of text **not** part of a discourse relation (Arg1 or Arg2) in tweets with a single connective (chars): 103

As for sense distributions, Table 5 shows the relative frequency distribution of Level-1 sense tags (i.e. *Class* level tags) in our corpus. We also calculated the relative frequencies for each *Class* level tag in the PDTB 2.0 according to frequencies of *Explicit* connectives presented in Table 4 in (Prasad et al., 2008). The second column in Table 5 shows the calculated relative frequencies in the PDTB corpus.[8] It can be seen from the distribution that there are a lot more CONTINGENCY relations in our Twitter data than in the PDTB, while there are fewer COMPARISON and TEMPORAL relations. Considering that not all explicit relations have been annotated in our Twitter corpus yet (only relations contained entirely within one tweet), no final conclusions can be drawn yet, but it appears that narrative (temporal) and comparative or contrastive relations are more typical of newspaper writing than spontaneous social media conversations. This is also reflected in the lists of frequent connectives (Tables 3, 4), which show that connectives expressing CONTINGENCY relations like *if*, *when*, and *so* occur relatively more frequently on Twitter.

[8]The class frequencies for PDTB column presented here come from the PDTB 2.0 sense hierarchy and we are using the relations in PDTB 3.0. Since there is no change defined in (Webber et al., 2018) regarding the Class level sense tags, we consider the columns in Table 5 as comparable.

Class	% in Twitter	% in PDTB
EXPANSION	33.4%	33.5%
CONTINGENCY	28.0%	18.7%
COMPARISON	24.3%	28.8%
TEMPORAL	14.3%	19.0%

Table 5: Distribution of class level sense tags.

We also allowed the annotator to select more than one sense if both were deemed relevant (see Rohde et al., 2018, for a discussion of multiple concurrent relations in text); this option was chosen in 9 cases, listed in Table 6.

Connective	#	Senses
when	3	SYNCH. + CONDITION
and	2	REASON + CONJUNCT.
(and, an)		RESULT + CONJUNCT.
anytime	1	SYNCH. + CONDITION
however	1	CONTR. + CONCESSION
or	1	DETAIL + CONCESSION
while	1	SYNCH. + CONTRAST

Table 6: Connectives with 2 simultaneous senses.

5 Conclusions and Future Work

We presented initial results of our PDTB style annotation of English Twitter conversations. Social media conversations are an interesting domain for such annotation, because despite the written mode, they show many properties typical of spoken interactions. They therefore allow an investigation of the discourse structure of written multi-party interactions. Since this type of annotation is still rare for spoken data, we hope to add to what is known about discourse structure in conversations.

We reported about our annotation of intra-tweet discourse relations in 185 Twitter threads. We conducted an inter-annotator agreement study, which revealed that in particular the selection of argument spans poses new problems in the Twitter domain. We are currently adapting and further specifying our annotation guidelines to cover the phenomena found in our social media data, such as elliptical constituents, hashtags and emoji, abbreviations, missing punctuation, etc. Based on the amended guidelines, we will validate the existing annotations and edit them for consistency.

The current study only reports on intra-tweet relations, where the connective and both arguments are contributed by the same speaker. However, inter-tweet relations are also found in our data. In these cases, a subsequent reply contains a connective and Arg2, but relates to an Arg1 in a previous tweet (typically by a different speaker). We are planning to add annotations for these types of relations, as well as non-explicit discourse relations, in future work.

Finally, we showed results from a basic analysis that demonstrates how explicit discourse relations in Twitter conversations differ from the relations in the PDTB newspaper text. Due to genre differences, Twitter conversations contain more CONTINGENCY and fewer TEMPORAL and COMPARISON relations. The distribution of connectives in Twitter also differs from newspaper text. This corresponds to known differences between connectives used in spoken and written language. Finally the syntactic type and size of arguments we find in the Twitter data differs markedly from the PDTB arguments.

In current work, we are extending the annotations to include inter-tweet and non-explicit relations. We are planning to use the corpus in developing a shallow discourse parser for English social media text.

Acknowledgements

The authors would like to thank Olha Zolotarenko for assisting with the annotation. This research was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under project nr. 317633480 – SFB 1287, project A03.

References

Berfin Aktaş, Tatjana Scheffler, and Manfred Stede. 2018. Anaphora resolution for Twitter conversations: An exploratory study. In *Proceedings of the First Workshop on Computational Models of Reference, Anaphora and Coreference*, pages 1–10.

Nicholas Asher and Alex Lascarides. 2003. *Logics of Conversation*. Cambridge University Press, Cambridge.

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

William Mann and Sandra Thompson. 1988. Rhetorical structure theory: Towards a functional theory of text organization. *TEXT*, 8:243–281.

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.

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 *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6095–6099. IEEE.

Hannah Rohde, Alexander Johnson, Nathan Schneider, and Bonnie Webber. 2018. Discourse coherence: Concurrent explicit and implicit relations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, volume 1, pages 2257–2267.

Tatjana Scheffler. 2014. A German Twitter snapshot. In *Proc. of the 9th International Conference on Language Resources and Evaluation (LREC)*, pages 2284–2289, Reykjavik, Iceland.

Tatjana Scheffler and Manfred Stede. 2016. Realizing argumentative coherence relations in German: A contrastive study of newspaper editorials and Twitter posts. In *Proceedings of the COMMA Workshop "Foundations of the Language of Argumentation"*, Potsdam, Germany.

Sara Tonelli, Giuseppe Riccardi, Rashmi Prasad, and Aravind K Joshi. 2010. Annotation of discourse relations for conversational spoken dialogs. In *Proc. of the 7th International Conference on Language Resources and Evaluation (LREC)*.

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

Deniz Zeyrek, Amália Mendes, and Murathan Kurfalı. 2018. Multilingual extension of PDTB-style annotation: The case of TED multilingual discourse bank. In *Proc. of the 11th International Conference on Language Resources and Evaluation (LREC)*.