TwiConv: A Coreference-annotated Corpus of Twitter Conversations

Berfin Aktaş
SFB1287
Research Focus Cognitive Sciences
University of Potsdam, Germany
berfinaktas@uni-potsdam.de

Annalena Kohnert
Department of Language Science
and Technology
Saarland University, Germany
akohnert@coli.uni-saarland.de

Abstract

This article introduces TwiConv, an English coreference-annotated corpus of microblog conversations from Twitter. We describe the corpus compilation process and the annotation scheme, and release the corpus publicly, along with this paper. We manually annotated nominal coreference in 1756 tweets arranged in 185 conversation threads. The annotation achieves satisfactory annotation agreement results. We also present a new method for mapping the tweet contents with distributed stand-off annotations, which can easily be adapted to different annotation tasks.

1 Introduction and Related Work

Microblog texts from Twitter present a discourse genre that carries non-standard language characteristics (e.g., noisy or informal language with abbreviations, purposeful typos, use of non-alphanumerical symbols such as #- and @-characters, misspellings, etc.) and is therefore challenging for NLP applications (Ritter et al., 2011; Sikdar and Gambäck, 2016). There exist a number of Twitter datasets annotated at different linguistic layers for investigating a variety of NLP tasks on this genre, including sentiment analysis (Cieliebak et al., 2017), named entity recognition (Derczynski et al., 2016), and event coreference resolution (Chao et al., 2019). Aktaş et al. (2018) tested an out-of-the-box nominal coreference resolution system trained on OntoNotes (Hovy et al., 2006; Weischedel et al., 2011) on Twitter data and showed that the system performs with much lower scores than the original reported values on that data. Hence, tweets are a complicated genre also for the task of nominal coreference resolution.

We introduce TwiConv, a nominal coreference-annotated corpus of English-language Twitter posts with the intent to explore the coreference features in conversational Twitter texts. Our annotation scheme is based on (Grishina and Stede, 2016), yet with some domain-driven adaptations. Twitter's Developer Policy[1] does not allow publishing the tweet contents. Therefore, most of the tweet datasets distribute the unique tweet IDs and annotations without the tweet text. However, if the tokenization of the corpus in concern is realized through a relatively complicated procedure or contains manual corrections, stand off annotation layers may not match with the text content in the compiled corpus. We thus present a distribution method for mapping the original tweet texts with our annotations. To our knowledge, TwiConv is the first tweet corpus for nominal coreference.

The remainder of paper is organized as follows. We describe the corpus compilation process in Section 2. In Section 3, we present the annotation principles along with a description of quality assurance methods. The main statistics of our corpus are presented in Section 4. Format of the distributed corpus and data sharing methodology are described in Section 5. Section 6 summarizes the presented work.

[1] developer.twitter.com/en/developer-terms/policy

Proceedings of the 3rd Workshop on Computational Models of Reference, Anaphora and Coreference (CRAC 2020), pages 47–54,
Barcelona, Spain (online), December 12, 2020.

2 Corpus Compilation

2.1 Data collection

We used *twarc*[2] to collect English-language tweets from the Twitter stream on several (non-adjacent) days in December, 2017. We did not filter for topics in any way, since that is not a concern for this corpus. Instead, our aim was to collect threads (conversations) by recursively retrieving parent tweets, whose IDs are taken from the `in_reply_to_id` field of the tweet object returned by the Twitter API. We then used a script from (Scheffler, 2017), which constructs the full conversational tree structure for any tweet that generated replies. A single *thread* (in our terminology) is a path from the root to a leaf node of that tree. For the purposes of this study, we are not interested in alternative replies and other aspects of the tree structure; so we kept only one of the longest threads (paths) from each tree and discarded everything else. Therefore, the data set does not contain any overlaps in tweet sequences. A sample thread structure with one example coreference chain annotation is illustrated in Appendix A.

2.2 Tokenization

It is well known that tokenization is a crucial preparatory step for doing any kind of NLP on texts. We experimented with two different tokenizers: the Stanford *PTBTokenizer* (Manning et al., 2014) and *Twokenizer* (Gimpel et al., 2011). It turned out that these systems have different strengths in handling challenging cases. For instance, only PTBTokenizer can handle the apostrophes (e.g., contracted verb forms and possessive markers). On the other hand, Twokenizer is stronger in recognizing the punctuation symbols even if they are not surrounded by whitespace. These cases are illustrated in Appendix B.

We thus decided to implement a tokenization pipeline where the output of the Twokenizer is given as input to the PTBTokenizer. The outcome of this pipeline process is compatible with Penn Treebank conventions[3] and, therefore, with the other corpora following the same conventions, such as OntoNotes (Weischedel et al., 2013) and Switchboard (Calhoun et al., 2010). We found that the number of tokens increased in the second step of the pipeline by 4%, and only 5% of newly generated tokens are erroneous over-generated tokens. Therefore, we don't consider over-tokenization as a potential problem for token-based compatibility with other corpora.

2.3 Sentence Segmentation

We followed a semi-automated segmentation procedure to split the tokenized tweets into sentences. We first segmented the text using the SoMaJo sentence splitter for English (Proisl and Uhrig, 2016). SoMaJo deals well with common Twitter tokens such as links, hashtags and abbreviations but fails when sentences in the same tweet start with lowercase letters or hashtags, and when the user does not use any punctuation. Therefore, we manually corrected the boundaries detected by SoMaJo.

3 Annotation

3.1 Annotation Principles

In our scheme, *markables* are phrases with nominal or pronominal heads. All nominal expressions, such as names, definite/indefinite noun phrases, pronouns, and temporal expressions are annotated for coreference. Non-referential pronouns, predicative copula constructions, and appositions are also annotated and distinguished by the attribute values assigned to them. Elements of the web language such as usernames and hashtags are considered as markables as well. Links and emojis are treated according to their grammatical roles. We illustrate these cases in Appendix C. We annotated all chains including singletons. Chains can contain several markables from the same tweet (intra-tweet) or from different replies (inter-tweet), which can lead to 1st, 2nd and 3rd pronouns referring to the same entity within one thread as in Example 1. We do not allow dicontinuous markables, therefore split antecedents and their co-referring mentions are annotated as separate markables (Example 3) unless they occur as compound phrases (Example 2)[4].

[2] `https://github.com/DocNow/twarc`

[3] `https://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html`

[4] The full guideline with examples is shared together with the corpus.

(1) Thanks to [you]$_i$, [I]$_j$ can now understand the whole conversation.
[You]$_j$ are welcome.

(2) [The baby and I]$_i$ are listening to [our]$_i$ favourite music.

(3) [I]$_i$ met [him]$_j$ at [our]$_k$ favourite café.

We used the MMAX2 tool (Müller and Strube, 2006) for annotations and customized its default settings according to our scheme. We defined comprehensive attributes for chains and mentions. All chains should be assigned a representative mention (i.e., the most descriptive mention in the chain), a semantic class (i.e., the semantic category of the entity) and genericity value (i.e., whether the referred entity is specific or generic). Mentions are assigned a nominal form (np_form) and grammatical role.

3.2 Annotation Quality

We applied the following procedures to assess and evaluate the quality of manual annotations.

1. **Automated Checks** We validated the consistency of the annotations by applying a number of automated procedures checking whether the constraints specified in the guideline are applied uniformly.

2. **Review of Annotations** We reviewed the annotations of the first 27 threads (15% of all threads in the corpus). In total, 33 problematic annotation cases were detected during this review, which affected approximately 50 mentions. Most of the problematic cases were due to incorrect selection of mention span or assignment of wrong attributes for different features specified in the guideline. The proportion of detected problems affects only 2% of all mentions in this sub-corpus. Therefore we did not see the necessity to extend the review process to the entire corpus.

3. **Inter-Annotator Agreement** We assessed the inter-annotator agreement (IAA) to evaluate the reliability of our annotation process. In the first version of the TwiConv corpus, we annotated only the coreference chains containing 3rd person pronouns. We conducted the inter-annotator agreement evaluation on this first version of the corpus. The most common annotator errors were different selection of mentions (missing or spurious markables), missing chains if they only contained very few mentions or the splitting of one chain into two, as well as occasional differences in markable span boundaries.

   We then extended the guideline (GL) and annotated all the coreference chains in the second version of the dataset. The changes in the extended GL only concern attributes, which are not addressed in the IAA study. Therefore, we are confident that this agreement study can assess our final scheme in terms of mention detection and chain linking.

   Artstein and Poesio (2008) propose the use of Krippendorff's α (Krippendorff, 1980) for set-based agreement tasks such as coreference annotations. Following their proposal, we used Krippendorff's α to measure the IAA for 12 randomly selected threads. Two linguistics students annotated this sub-corpus. We computed the IAA for mention detection and chain linking. We calculated the Krippendorff's α by following the methodology described in (Passonneau, 2006) and found its value as 0.872 ($\alpha \geq .800$) which indicates reliability of our data annotations for research purposes.

4 Corpus Overview

The resulting TwiConv corpus consists of 1756 tweets in 185 threads, with the average length of a tweet being 153 characters. We present additional descriptive statistics for TwiConv corpus in Table 1 and for annotations in Table 2.

5 Corpus Distribution

5.1 Corpus format

The annotations are stored in a CoNLL format (i.e., tab-separated) with 17 columns in total, one file per Twitter thread. The content of each column is described in Table 3 and an example is presented

# of threads	185
# of tweets	1756
# of tokens	48172
# of sentences	3503
# of clauses	6719
average thread length (token)	260.4
average sentence length (token)	13.6

Table 1: General statistics on the corpus

# of mentions:	12374
# of chains:	7035
# of non-singleton (ns) chains:	1734
# of intra-tweet coref chains (ns):	674
# of inter-tweet coref chains (ns):	1060
# of username mentions:	124
# of mentions including hashtag:	94
Average mention length (in tokens):	1.94

Table 2: Descriptive statistics of the coreference annotations

in Appendix E. The Part-of-Speech tags and parses in column 4 and 5 are automatically created with Stanford Parser (Manning et al., 2014) with no manual correction. Empty lines indicate sentence breaks.

Column	Content	Column	Content
0	Thread ID	9	NP form/reference type
1	Thread No	10	Coreference ID
2	Token No in sentence	11	Clause boundary
3	Token	12	Shortest NP boundary
4	POS tag	13	Longest NP boundary
5	Parse info	14	Grammatical role
6	Speaker/User handle	15	Genericity
7	Representative mentions	16	[Tweet No in thread]_[Sentence
8	Semantic class		No in tweet]_[Token No in sentence]

Table 3: Column content in CoNLL format corpus

It is possible that different mentions start at the same token, e.g. "My Twitter username" marks both the beginning of the pronoun mention "My" as well the full definite noun mention "My Twitter username". In this case, we used pipe symbols ("|") to separate the annotations for different mentions. The order of the annotations separated by the pipe symbol remained the same for the entire line, meaning that the order of annotations in pipe-separated columns is always the same.

Further, some annotations such as NP form and grammatical role have sub-categories, which we express by slashes ("/"): e.g. *ppers/anaphora* marks a personal pronoun that functions as an anaphoric expression. Similarly, the grammatical role *other* can be either appositive, vocative or other (e.g., *other/vocative*), but those sub-categories were only assigned to the *other* type, not to subjects, prepositional phrases etc.

We used the automatically created parses to detect the clause and NP boundaries (both for shortest and longest NP spans) in tweets. We manually corrected the detected boundaries and added boundary information to the data files (i.e., boundary start and end tokens are specified in columns 11-13 in Table 3). The last column in the data files represent the relative order of tokens in the texts.

5.2 Sharing Method

Due to Twitter's Developer Policy, we have to refer to tweets via their ID, through which the message text as well as other tweet-related information can be downloaded.

In order to share the data, we use a method similar to the distribution of the CoNLL-2012 Shared Task Data (Pradhan et al., 2012) and provide skeleton files which include all annotations, but no tokens from the Twitter message and no usernames (instead, they are replaced by underscore characters). For each token, the ID of the tweet from which the token originates is indicated at the end of the corresponding line. As we have tokenized the data, we also provide reference files to recreate our tokenization steps. To create those *diff* files, we compared files with the whitespace tokenized tweets (with one token per

line) to ones with the tweets with our final tokenization (one token per line as well) with the Linux program *diff*. We share only those tokens in the *diff* files that were affected by the tokenization method or other forms of modification such as encoding differences for emoticons. For a sample representation, see Appendix D.

After downloading all still available tweets, they have to be transformed into above described format (whitespace tokenized, one token per line, one file per tweet). We provide an assembly script that will use these tweet files, the skeleton files and *diff* files to create the complete CoNLL files with all annotations and tokens[5]. The script itself contains no information about the content of the annotations and can be re-used for any other tweets, given that the *diff* and skeleton files (following the CoNLL-style format described in Table 3) have been generated correctly. For unavailable tweets, the tokens will remain anonymized (meaning the underscore character remains).

6 Conclusion

We have developed a comprehensive annotation scheme for annotating nominal coreference in English Twitter conversations and fully annotated 1756 tweets arranged in 185 threads. Assessment of annotations and correction of erroneous cases were made via inter-annotator agreement evaluation, partial review, and automated checks. We distribute the corpus without tweet contents and introduce tools for researchers to map the tweet texts, captured using the tweet IDs, with the shared annotations. We hope that the release of the TwiConv corpus will increase the interest in coreference studies on this genre.

Acknowledgements

We thank the anonymous reviewers and Manfred Stede for their helpful observations and suggestions. This work is funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - Projektnummer 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 (CRAC@NAACL 2018)*, pages 1–10, New Orleans, Louisiana, June. Association for Computational Linguistics.

Ron Artstein and Massimo Poesio. 2008. Survey article: Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.

Sasha Calhoun, Jean Carletta, Jason M. Brenier, Neil Mayo, Dan Jurafsky, Mark Steedman, and David Beaver. 2010. The nxt-format switchboard corpus: A rich resource for investigating the syntax, semantics, pragmatics and prosody of dialogue. *Language Resources and Evaluation*, 44(4):387–419, 12.

W. Chao, P. Wei, Z. Luo, X. Liu, and G. Sui. 2019. Selective expression for event coreference resolution on twitter. In *2019 International Joint Conference on Neural Networks (IJCNN)*.

Mark Cieliebak, Jan Milan Deriu, Dominic Egger, and Fatih Uzdilli. 2017. A twitter corpus and benchmark resources for german sentiment analysis. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, pages 45–51.

Leon Derczynski, Kalina Bontcheva, and Ian Roberts. 2016. Broad Twitter corpus: A diverse named entity recognition resource. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1169–1179, Osaka, Japan, December. The COLING 2016 Organizing Committee.

Kevin Gimpel, Nathan Schneider, Brendan O'Connor, Dipanjan Das, Daniel Mills, Jacob Eisenstein, Michael Heilman, Dani Yogatama, Jeffrey Flanigan, and Noah A. Smith. 2011. Part-of-speech tagging for twitter: Annotation, features, and experiments. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Short Papers - Volume 2*, HLT '11, pages 42–47, Stroudsburg, PA, USA. Association for Computational Linguistics.

[5]Scripts and data to reproduce the corpus can be found at `https://github.com/berfingit/TwiConv`

Yulia Grishina and Manfred Stede, 2016. *Parallel coreference annotation guidelines.*, November.

Eduard Hovy, Mitchell Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. 2006. OntoNotes: The 90% solution. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*, pages 57–60, New York City, USA, June. Association for Computational Linguistics.

K. Krippendorff. 1980. *Content Analysis: An Introduction To Its Methodology*. Sage commtext series. Sage Publications.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*, pages 55–60.

Christoph Müller and Michael Strube. 2006. Multi-level annotation of linguistic data with mmax2. In Joybrato Mukherjee Sabine Braun, Kurt Kohn, editor, *Corpus Technology and Language Pedagogy: New Resources, New Tools, New Methods*.

Rebecca Passonneau. 2006. Measuring agreement on set-valued items (MASI) for semantic and pragmatic annotation. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy, May. European Language Resources Association (ELRA).

Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Olga Uryupina, and Yuchen Zhang. 2012. CoNLL-2012 Shared Task: Modeling Multilingual Unrestricted Coreference in OntoNotes. In *Joint Conference on EMNLP and CoNLL-Shared Task*, pages 1–40.

Thomas Proisl and Peter Uhrig. 2016. SoMaJo: State-of-the-art tokenization for German web and social media texts. In *Proceedings of the 10th Web as Corpus Workshop (WAC-X) and the EmpiriST Shared Task*, pages 57–62, Berlin. Association for Computational Linguistics (ACL).

Alan Ritter, Sam Clark, Mausam, and Oren Etzioni. 2011. Named entity recognition in tweets: An experimental study. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, page 1524–1534, USA. Association for Computational Linguistics.

Tatjana Scheffler. 2017. Conversations on twitter. In Darja Fišer and Michael Beißwenger, editors, *Researching computer-mediated communication: Corpus-based approaches to language in the digital world*, pages 124–144. University Press, Ljubljana.

Utpal Kumar Sikdar and Björn Gambäck. 2016. Feature-rich twitter named entity recognition and classification. In *Proceedings of the 2nd Workshop on Noisy User-generated Text (WNUT)*, pages 164–170, Osaka, Japan, December. The COLING 2016 Organizing Committee.

Ralph Weischedel, Eduard Hovy, Mitchell Marcus, Martha Palmer, Robert Belvin, Sameer Pradan, Lance Ramshaw, and Nianwen Xue. 2011. Ontonotes : A large training corpus for enhanced processing. In Joseph Olive, Caitlin Christianson, and John McCary, editors, *Handbook of Natural Language Processing and Machine Translation*.

Ralph Weischedel, Martha Palmer, Mitchell Marcus, Eduard Hovy, Sameer Pradhan, Lance Ramshaw, Nianwen Xue, Ann Taylor, Jeff Kaufman, Michelle Franchini, Mohammed El-Bachouti, Robert Belvin, and Ann Houston. 2013. Ontonotes release 5.0 ldc2013t19. *Web Download. Linguistic Data Consortium, Philadelphia, PA*.

Appendices

Appendix A: Thread sample

The only Russia collusion occurred when [@HillaryClinton]$_i$ conspired to sell US Uranium to a Russian oligarch while [she]$_i$ was in charge.

Why is the mainstream media so quiet? Probably because [#theSecretaryofState]$_i$ is still powerfull.

Haven't you heard , dear???? [HRC]$_i$ is NOT president!!!

.[She]$_i$ doesn't have to be a President to face crimes [she]$_i$ committed, dear .

Appendix B: Tokenization examples

	String	**Twokenizer**	**PTBTokenizer**	**TwiConv Pipeline**
1	aren't	aren't (1)[6]	are, n't (2)	are, n't (2)
2	you've	you've (1)	you, 've (2)	you, 've (2)
3	London's	London's (1)	London, 's (2)	London, 's (2)
4	here:)Because	here, :), Because (3)	here:)Because (1)	here, :), Because (3)
5	..	.. (1)	., . (2)	., . (2*)

Table 4: Tokenization output

Appendix C: Twitter mention examples

(4) .. [@SomeUser] just said twice that.. *("username" as a mention)*

(5) this doesn't pass [the #smelltest] *("hashtag" as part of a mention)*

(6) [] are fools ... *("emoji" as a mention)*

(7) If crashing, please refer to this: **[https://exampleurl.com]** *("link" as a mention)*

Appendix D: Tokenization differences

```
This
is
just
a
test.
Hi
Twitter!
```

Figure 1: Example Tweet, whitespaced tokenized

```
5,6c5
< test
< .
---
> test.
8,9c7
< Twitter
< !
---
> Twitter!
```

Figure 2: *diff* file example

Appendix E: CoNLL-formatted sample annotation

```
#begin document (001_98765432100000000.branch1.); part 0
001_98765432100000000.branch1.  0  0  This  DT   (ROOT(S(NP*)  SomeUsername  -  -  pds/cataphora  (0)  CL0  NP_S(  NP_L(  -  -  0_0_0
001_98765432100000000.branch1.  0  1  is  VBZ  (VP*  SomeUsername  -  -  -  -  -  )NP_S  -  -  -  0_0_1
001_98765432100000000.branch1.  0  2  just  RB  (ADVP*  SomeUsername  -  -  -  -  -  -  -  -  -  0_0_2
001_98765432100000000.branch1.  0  3  a  DT  (NP*  SomeUsername  representative_men  -  indefnp/none  (1  -  NP_S()  )NP_L  -  -  0_0_3
001_98765432100000000.branch1.  0  4  test  NN  *)))  SomeUsername  -  -  -  1)  -  -  -  -  -  -  -  0_0_4
001_98765432100000000.branch1.  0  5  .  .  *))  SomeUsername  -  -  -  -  CL0  -  -  -  -  -  -  -  0_0_5

001_98765432100000000.branch1.  0  0  Hi  UH  (ROOT(INTJ*  SomeUsername  -  -  -  -  CL1  -  -  -  -  0_1_0
001_98765432100000000.branch1.  0  1  Twitter  NNP  (NP*)  SomeUsername  representative_men  -  ne/none  (2)  -  -  -  -  -  -  0_1_1
001_98765432100000000.branch1.  0  2  !  .  *))  SomeUsername  -  -  -  -  CL1  -  -  -  -  0_1_2

#end document
```

Figure 3: Example CoNLL file