Resolving Pronouns in Twitter Streams: Context can Help!

Anietie Andy
University of Pennsylvania
andyanietie@gmail.com

Chris Callison-Burch
University of Pennsylvania
ccb@cis.upenn.edu

Derry Tanti Wijaya
Boston University
wijaya@bu.edu

Abstract

Many people live-tweet televised events like Presidential debates and popular TV-shows and discuss people or characters in the event. Naturally, many tweets make pronominal reference to these people/characters. We propose an algorithm for resolving personal pronouns that make reference to people involved in an event, in tweet streams collected during the event.

1 Introduction

Pronoun resolution is an important task in natural language processing (Denis and Baldridge, 2007; Clark and Manning, 2015; Cheri et al., 2016; Yin et al., 2018). However, not a lot of work has been done to address pronoun resolution in Twitter streams related to events. During a televised event such as a Presidential debate or TV-shows, individuals publish tweets about people in the context in which they are being portrayed in the event (Andy et al., 2017; Andy et al., 2019). Some of these tweets make reference to people using third-person singular pronouns like *he*, *him*, *his*, *she*, and *her*. For example, here are some tweets about an episode of the TV-show, Game of Thrones season 7 (GoTS7) that were published during the same minute while the episode was airing: (i) *"she took his face"*, (ii) *"walder frey?! probably just before arya killed him"*, and (iii) *"wait where is arya did she change to his face"*. With short text such as these event-related tweets, although some tweets might mention the referents in the same tweets as the pronouns (e.g., in (ii) where *"him"* makes reference to the character *"Walder Frey"* in the same tweet), some other tweets may not contain the referents in the same tweets (e.g., in (i) and (iii), the pronouns {*"she"*, *"his"*} and *"his"* respectively refer to the characters, {*"Arya"*, *"Walder Frey"*} and *"Walder Frey"* that are not mentioned in (i) and (iii) respectively).

Resolving a pronominal mention in an event-related tweet is a challenging task because a tweet with pronominal mentions either: (i) does not mention any person, (ii) makes reference to a person not mentioned in the tweet, and (iii) mentions more than one person (who may or may not have the same gender as the pronoun).

In this paper, taking advantage of the context in which a pronoun is mentioned in the tweet, the tweet's temporal information, and the context in which other people are mentioned in tweets about the same event published at the same time period, we develop an algorithm to automatically resolve these third-person singular pronouns.

We evaluate our algorithm on tweets collected around two events: (1) a United States (US) Democratic party Presidential debate and (2) an episode of a popular TV-show, GoTS7. We show that our algorithm outperforms baselines. We will make these datasets available to the research community.

2 Related Work

A lot of work on resolving pronouns in text has been done. In Denis and Baldridge (2007), a ranking approach for resolving pronouns in text was proposed. In Clark and Manning (2015), an entity-based coreference model which incrementally learns to resolve coreference was proposed. Yin et al. (2018) proposed a self-attention method to model zero pronouns. In Cheri et al. (2016), the eye movements of participants annotating documents were tracked to gain insights to some of the processes people use in coreference resolution. Lee et al. (2017) proposed an end-to-end coreference resolution model that is

based on a neural model; the proposed model takes into consideration all the spans in a given text and given a span, it determines if there is a previous span which is an antecedent. Abzaliev (2019) proposed a coreference resolution model which uses fine-tuned BERT embeddings. In Kocijan et al. (2019), a large coreference resolution dataset was constructed. In Rudinger et al. (2018), gender bias in coreference resolution was studied and it was shown that gender bias exists in some coreference resolution systems. In Zhao et al. (2018), a coreference resolution dataset focused on gender bias was constructed and similar to Zhao et al. (2018) it was shown that some coreference resolution systems are gender biased; a model was proposed to remove these biases while maintaining the performance on coreference datasets.

Not a lot of prior work has been done to resolve pronouns in Twitter data. The closest related work to our work is Aktaş et al. (2018), which studies pronominal anaphora on conversations in Twitter and constructs a corpus to determine relevant factors for resolving anaphora in Twitter conversation data.

Our algorithm is different from Aktaş et al. (2018) because it focuses on resolving pronouns in tweets that make reference to people and characters portrayed in an event.

3 Dataset and Labeling

Our dataset consist of tweets collected during the airing of an hour-long episode of the popular HBO show, GoTS7 and tweets collected during night 1 of the first 2020 US Democratic party Presidential debate (the first debate was held on 2 nights).

3.1 GoTS7 dataset

Using a Twitter streaming API, we collected 4,223 time-stamped tweets that contained *"#got"* - a popular hashtag for the show, while episode 3 of GoTS7 was airing. From these we identified tweets that mentioned a third-person singular pronoun and we collected the timestamp in which each of these tweets was published. In this episode of GoTS7, 35 characters were portrayed, 24 of which were male and 11 female. From our dataset, we observed that in this episode an average of 93 posts were published per minute.

Labeling: The day after each episode of GoTS7 aired, the New York Times (NYTimes) published a summary of the episode. We collected the NYTimes summary of this episode. We showed this summary to 3 annotators who had watched the episode. Then given a tweet with a third-person singular pronoun mention and the timestamp this tweet was published, we asked the annotators to identify the character that was being referred to in the tweet. We selected a character for each pronoun mention if at least 2 of the annotators identified the same character as the one the pronoun was referring to. Using kappa calculation, we calculated the agreement between the annotators and got 0.86. Our labeled dataset contains 154 tweets with resolved pronominal references to characters. 59% (i.e., 91) of the labeled tweets contained both a third-person singular pronoun and the character that was being referenced by the pronoun, and 41% (i.e., 63) mentioned a third-person singular pronoun and did not contain the character that was being referenced by the pronoun.

3.2 US Presidential debate dataset

Similar to section 3.1, we collected 46,142 time-stamped tweets that contained the word *"debate"* while the Presidential debate was airing and identified tweets containing a third-person singular pronoun and their corresponding timestamps. There were 10 candidates who participated in the debate, 7 of which were male and 3 female. From our dataset, we observed that on the average, 431 posts were published per minute in our Presidential debates.

Labeling: While the debate was airing, NYTimes had a live-blog with some political reporters discussing and analyzing the debate in real-time[1]. We collected these live-blog discussions and analysis and showed them to 3 annotators who had watched the debate. Given a tweet with a mention of a third-person singular

[1] https://www.nytimes.com/interactive/2019/06/26/us/politics/democratic-debate-live-chat.html

pronoun and the timestamp in which the tweet was published, the annotators were asked to identify the Presidential candidate that was being referenced by the pronoun. A candidate was selected if at least 2 of the annotators identified the same candidate as the one being referenced in the tweet. Using kappa calculation, we calculated the agreement between the annotators and got 0.83. Our labeled dataset is made up of 141 tweets with resolved pronominal references to Presidential candidates. 84% (i.e., 118) of the labeled tweets contained both a third-person singular pronoun and the candidate that was being referenced by the pronoun and 16% (i.e., 23) of the labeled tweets mentioned a third-person singular pronoun and did not contain the candidate that was being referenced by the pronoun.

The annotators mostly agreed, however, some of the tweets in which the annotators did not agree on made reference to characters using third-person singular pronouns without mentioning any characters in the tweet, as shown by the following examples from GoTS7: (1) *"lol he can't stop himself #gots7"*, (2) *"noo shes my favorite #gots7"*, and (3) *"now it's time to get her rocks off #gots7"*.

4 Our Algorithm

Our algorithm has 3 steps:

4.1 Step 1: Candidate/Character Identification:

Prior to the Presidential debate, NYTimes published the names of the Presidential candidates who would be debating[2]. We selected the candidate names from the NYTimes article. For GoTS7, we identified characters by selecting all the character names listed in the Wikipedia page of GoTS7.

Some event-related tweets mention people by their names or aliases (Andy et al., 2017), hence for each candidate/character in each of these events, we construct an alias list which consists their first name (which is unique in both the Presidential debates and GoTS7), their last name if it is unique in the event, and the nickname listed in the first paragraph of the character (person entity) Wikipedia page. We also selected the gender of each candidate and character from the candidates and shows Wikipedia page, respectively.

4.2 Step 2: Identifying the context in tweets

To determine the context of pronoun mentions and candidate/character mentions, we use the pre-trained BERT (Devlin et al., 2018) language representation model. We input each tweet containing the pronoun or candidate/character mentions to the pre-trained BERT-Base model and extract the 768-dimensional contextual embeddings of the pronoun or candidate/character tokens, which we take as their contextual representations in the tweet, generated from the final hidden layer of the pre-trained model. For multi-token candidates/characters, we use the average of their token embeddings.

4.3 Step 3: Pronoun resolution

A tweet with a pronominal mention either contains possible persons being referred to by the pronoun (section 4.3.2) or it does not (i.e., either because it does not contain any person mention or it does not contain person mentions with the same gender as the pronoun) (section 4.3.1). Our algorithm handles both cases.

4.3.1 Case 1: Tweets with pronouns but no possible person referent

Given a tweet t with a third-person singular pronoun mention and the timestamp it was published, our algorithm identifies all the candidates/characters that were mentioned more than k times in tweets published in the same minute as tweet t and groups the tweets that mention the same candidate/character together. To identify the optimal value for k, we collected tweets published around 2 other episodes of GoTS7 and tweets published around another US Presidential debate and randomly selected tweets published in a 20 minute time period in each of these events; we varied the number of candidate/character mentions (between 1 to 5) per minute and observed that on the average, candidates/characters mentioned more than 3 times in tweets published in a minute, were referred to by a pronoun in the same minute,

[2] https://www.nytimes.com/2019/06/26/us/politics/democratic-debate-lineup.html

hence we choose *k=3*. For each group of tweets, where tweets in each group make reference to the same candidate/character, our algorithm calculates the cosine similarity between the BERT embedding of the third-person singular pronoun mention and the BERT embedding of the candidate or character mentions in each tweet in each group. The average cosine similarity between the pronoun mention and the candidate/character mentions in each group is calculated. Since we know the gender of each of the candidates/characters (section 4.1), our algorithm identifies the candidate/character being referred to by the pronoun in the tweet by selecting the candidate/character with the largest average cosine similarity to the pronoun embedding, above or equal to a threshold (0.3), and its gender matches the pronoun's. We chose 0.3 as the threshold by using selected tweets published in two other GoTS7 episodes and another US Presidential debate to compare candidate/character embeddings to pronoun embeddings published in the same minute and observed that a threshold of 0.3 gave optimal results for matching pronouns to their referents.

Algorithms	Precision	Recall	F1
Our Model	**0.79**	**0.65**	**0.71**
Last person mention	0.47	0.40	0.43
Spike per minute	0.68	0.48	0.56

Table 1: Results of applying our algorithm and baselines to tweets with pronoun mentions but no possible referent mention in our Presidential debate dataset

Algorithms	Precision	Recall	F1
Our Model	**0.73**	**0.65**	**0.68**
Last person mention	0.61	0.23	0.33
Spike per minute	0.69	0.62	0.65

Table 2: Results of applying our algorithm and baselines to tweets with pronoun mentions but no possible referent mention in our GoTS7 dataset

4.3.2 Case 2: Tweets with pronoun mentions and possible person referents

Given a tweet t with a third-person singular pronoun mention and person mentions, and the timestamp this tweet was published, our algorithm identifies all the candidates/characters that were mentioned more than *k=3* times in tweets published in the same minute as tweet t and groups the tweets that mention the same candidate/character together.

In some cases, the pronoun in the tweet t could be making reference to the candidate/character mentioned in t. Therefore, we also select the candidate/character mentioned in the tweet as a candidate/character mention; here, we do not give any additional weight or preference to the mentioned candidate/character because it is possible that the pronoun might be making reference to a different character as shown in the following examples : (1) *"Cersei is about to kill her"* and (2) *"Cersei is taunting her"*, where in both cases, *"her"* refers to a different character, *"Ellaria"*.

Similar to Section 4.3.1, for each group of tweets, where tweets in each group make reference to the same candidate/character, our algorithm calculates the cosine similarity between the BERT embedding of the third-person singular pronoun and the BERT embeddings of each of the candidate/character mentions. The candidate/character with the highest average cosine similarity to the pronoun embedding, above or equal to a threshold (0.3), and its gender matches the pronoun in the given tweet, is selected as the candidate/character that resolves the pronoun.

5 Experiments

5.1 Case 1: Tweets with pronouns but no possible person referent:

Here we compared our algorithm to the following baselines:

Algorithms	Precision	Recall	F1
Our Model	**0.95**	**0.82**	**0.88**
Stanford Coref	0.90	0.63	0.74
Spike per minute	0.91	0.49	0.64
Neural model	0.75	0.55	0.63

Table 3: Results of applying our algorithm and baselines to tweets with pronoun mentions and mention of possible person referents in our Presidential debate dataset

Algorithms	Precision	Recall	F1
Our Model	**0.91**	**0.82**	**0.86**
Stanford Coref	0.73	0.30	0.43
Spike per minute	0.78	0.60	0.68
Neural model	0.83	0.71	0.77

Table 4: Results of applying our algorithm and baselines to tweets with pronoun mentions and mention of possible person referents in our GoTS7 dataset

Most frequent candidate/character mention per minute (Spike per minute): Given a tweet t with a pronoun mention, the tweets published in the same minute as t are identified. The candidate/character with the most mentions in these tweets, and is the same gender as the referenced third-person singular pronoun is identified as the candidate/character the pronoun is referring to.

Last person mentioned in tweet by author: For each given tweet t with a third-person singular pronoun, we select the last candidate/character that was mentioned in a tweet by the author of t.

Tables 1 and 2 show the results from our algorithm compared to these baselines on the Presidential debate and GoTS7 datasets, respectively.

5.2 Case 2: Tweets with pronoun mentions and possible person referents:

In this section, we compare our algorithm to the baseline, *Spike per minute*, described in Section 5.1. We also compare our algorithm to the Stanford coreference toolkit (Manning et al., 2014) and a neural coreference resolution model (neural model) (Lee et al., 2017); this model, when given a span of text, determines if any of the previous spans of text is an antecedent. Tables 3 and 4 show the results.

6 Error Analysis and Future Work

In this work, we focused on gathering context in the form of tweets published in the same minute as the tweet with the pronomial mention. One of the challenges we observed is that in some cases a pronomial mention might make reference to a character or person not mentioned in the same minute, hence in the future, we plan to explore how far back in time we should expand this context prior. With regards to tweets with pronouns but no possible person referent, one future avenue to explore is to prepend these tweets with previous tweets published by the same authors and apply state-of-the-art coreference model on these expanded tweet "paragraphs". However, as we observe in our experiments, the last tweet prior may not contain the referent (as evident from the low recall of the "Last person mention" baseline). Therefore, we plan to explore how far back in tweet (or time) we should prepend these tweets.

7 Conclusion

In this work, we develop an algorithm that resolves third-person singular pronouns in Twitter data related to two events: a Presidential debate and GoTS7. We show that our algorithm can help resolve third-person singular pronouns in these event-related tweets, even in cases where there are no possible person referent mentioned in a tweet. We also show that our method outperforms baselines.

References

Artem Abzaliev. 2019. On gap coreference resolution shared task: insights from the 3rd place solution. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 107–112.

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.

Anietie Andy, Mark Dredze, Mugizi Rwebangira, and Chris Callison-Burch. 2017. Constructing an alias list for named entities during an event. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pages 40–44.

Anietie Andy, Derry Tanti Wijaya, and Chris Callison-Burch. 2019. Winter is here: Summarizing twitter streams related to pre-scheduled events. In *Proceedings of the Second Workshop on Storytelling*, pages 112–116.

Joe Cheri, Abhijit Mishra, and Pushpak Bhattacharyya. 2016. Leveraging annotators' gaze behaviour for coreference resolution. In *Proceedings of the 7th Workshop on Cognitive Aspects of Computational Language Learning*, pages 22–26.

Kevin Clark and Christopher D Manning. 2015. Entity-centric coreference resolution with model stacking. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, volume 1, pages 1405–1415.

Pascal Denis and Jason Baldridge. 2007. A ranking approach to pronoun resolution. In *IJCAI*, volume 158821593.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Vid Kocijan, Oana-Maria Camburu, Ana-Maria Cretu, Yordan Yordanov, Phil Blunsom, and Thomas Lukasiewicz. 2019. Wikicrem: A large unsupervised corpus for coreference resolution. *arXiv preprint arXiv:1908.08025*.

Kenton Lee, Luheng He, Mike Lewis, and Luke Zettlemoyer. 2017. End-to-end neural coreference resolution. *arXiv preprint arXiv:1707.07045*.

Christopher D Manning, Mihai Surdeanu, John Bauer, Jenny Rose Finkel, Steven Bethard, and David McClosky. 2014. The stanford corenlp natural language processing toolkit. In *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations*, pages 55–60.

Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 2018. Gender bias in coreference resolution. *arXiv preprint arXiv:1804.09301*.

Qingyu Yin, Yu Zhang, Weinan Zhang, Ting Liu, and William Yang Wang. 2018. Zero pronoun resolution with attention-based neural network. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 13–23.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender bias in coreference resolution: Evaluation and debiasing methods. *arXiv preprint arXiv:1804.06876*.