Anaphora Resolution for Twitter Conversations: An Exploratory Study

Berfin Aktaş **Tatjana Scheffler** **Manfred Stede**
firstname.lastname@uni-potsdam.de
SFB 1287
Research Focus Cognitive Sciences
University of Potsdam
Germany

Abstract

We present a corpus study of pronominal anaphora on Twitter conversations. After outlining the specific features of this genre, with respect to reference resolution, we explain the construction of our corpus and the annotation steps. From this we derive a list of phenomena that need to be considered when performing anaphora resolution on this type of data. Finally, we test the performance of an off-the-shelf resolution system, and provide some qualitative error analysis.

1 Introduction

We are interested in the task of pronominal anaphora resolution for conversations in Twitter, which to our knowledge has not been addressed so far. By 'conversation', we mean tree structures originating from the `reply-to` relation; when using replies, people often (though not always) interact with each other across several turns.[1] Hence, anaphora resolution needs to attend both to the general and well-known problems of handling Twitter language, and potentially to aspects of conversation structure.

In order to study the properties of coreference relations in these conversations, we built a corpus that is designed to represent a number of different relevant phenomena, which we selected carefully. We annotated pronouns and their antecedents, so that the data can be used for systematically testing anaphora resolvers, and we conducted experiments with the Stanford system (Clark and Manning, 2015).

The paper is structured as follows: Section 2 introduces general phenomena found in Twitter conversations and describes earlier research. Section 3 discusses our approach to corpus construction and annotation. Section 4 shows in detail which "non-standard" phenomena we encountered in annotating the Twitter conversations in our corpus, and which need to be tackled by a coreference resolver. Section 5 outlines our experiments with the Stanford resolver and presents the results; finally we draw some conclusions in Section 6.

[1] For an overview of constructing corpora of this kind and some annotation tasks, see (Scheffler, 2017).

2 Overview of the Task and Related Work

In this section, we provide an overview of research that has addressed anaphora resolution specifically in the context of dialogue, multilogue, or social media. There we encounter the following phenomena that are potentially relevant for our scenario of Twitter conversations (and which are largely not present in monologue and hence in the "standard" work on anaphora resolution):

1. Pronouns referring to speakers
2. Other exophoric reference
3. Conversation structure as a factor for antecedent selection
4. Phenomena specific to spoken conversation
5. Phenomena specific to social media text

Obviously, not all of these phenomena are equally relevant in all interactive dialogue settings — in fact, certain settings basically do not require attending to such phenomena. For instance, the early work on TRAINS/TRIPS (Tetreault and Allen, 2004) emphasized the role of semantic features for pronoun resolution, while the factor of conversation structure was not so relevant, as the human-machine dialogues were relatively simple. Likewise, early work by Strube and Müller (2003) on the Switchboard corpus demonstrated that existing approaches to statistical pronoun resolution

could carry over to conversational data, but the authors focused on non-nominal antecedents and did not emphasize the need for using additional interaction features.

2.1 Reference to Speakers

In addition to using proper names, speakers can refer to one another using pronouns, and several early systems implemented simple rules for resolving *I* and *you* (e.g., (Jain et al., 2004)). In multilogue, it is also possible that third-person pronouns *he/she* refer to conversation participants; we are not aware of systems addressing this.

2.2 Other exophoric reference

This phenomenon was already prominent in TRAINS (see above), but largely handled by using semantic type constraints. It also occurs in Maptask dialogue and similar task-solving interactions like the Pentomino puzzle studied by Schlangen et al. (2009). Why is it potentially relevant for Twitter conversations? Because messages may contain embedded images, and speakers occasionally refer directly to entities therein. This is also possible with URLs and prominent objects present in the target page.

2.3 Conversation structure

The role of the turn structure in dialogue has received a lot of attention for anaphora resolution. Both (Poesio et al., 2006) and (Stent and Bangalore, 2010) were interested in the relative performance of specific dialogue structure models (the Grosz/Sidner stack model and Walker's cache model). Luo et al. (2009) worked with the mixed-genre ACE 2007 data and showed that features capturing the identity of the speaker and the same/different turn distinction can be very helpful for anaphora resolution, yielding an improvement of 4.6 points for telephone conversations. In contrast, Désoyer et al. (2016) used French spoken dialogues and could not find improvements when using information on speaker identity and the distance measured in number of intervening turns.

Niraula and Rus (2014) conducted a thorough analysis on the influence of turn structure for anaphora resolution in tutoring system dialogues. Following their corpus analysis, they implemented a single "discourse" feature, viz. the location of the antecedent candidate on the dialogue stack; this turned out to be one of the most predictive features in their classifier.

2.4 Spoken conversation

Not much work has been done on speech-specific features for anaphora resolution; we mention here the influence of hesitations that Schlangen et al. (2009) studied for referring to Pentomino pieces. The potential connection to Twitter is the fact that Twitter users often borrow from speech, for example emphasis markers such as vowel lengthening (*honeyyyy*) and hesitation markers (*hmm*).

2.5 Social media text

The need for pre-processing Twitter text is widely known and not specific to anaphora resolution. As just one example, Ritter et al. (2011) worked on Named-Entity Recognition on Tweets. They show that performance can be significantly improved when a dedicated preprocessing pipeline is employed. But we are not aware of Twitter-specific work on coreference or anaphora.

Finally, we mention an early study on threaded data, as found for instance in email, blogs or forums. (Hendrickx and Hoste, 2009) studied the performance of coreference resolution (implemented following the mention-pair model) when moving from standard newspaper text to online news and their comments, and to blogs. They found performance drops of roughly 50% and 40%, respectively.

3 Corpus

3.1 Collecting Twitter Threads

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 hashtags or 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. We then used a script from (Scheffler, 2017), which constructs the conversational full tree structure for any tweet that generated replies. Now, a single *thread* (in our terminology) is a path from the root to a leaf node of that tree. For the purposes of this paper, we were not interested in alternative replies and other aspects of the tree structure; so we kept only one of the longest threads (path) from each tree and discarded everything else. Therefore, the data set does not contain any overlaps in tweet sequences.

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

thread length	3	4-10	11-50	51-78
number of threads	20	120	43	2
pronouns per thread (avg.)	4	5	19	55

Table 1: Distribution of thread length and 3rd person pronoun frequency in the annotated corpus

We decided to start our study on 3rd person sg. pronouns, as these are the most relevant for anaphora resolution. Hence we leave the handling of first and second person pronouns (which are usually deictic, i.e., depending on who is replying to whom in the conversation structure) as well as plural pronouns for future work. To ensure a minimum conversation complexity, we selected only threads containing at least three tweets; the additional selection criterion is that the thread has at least one instance of one of the pronouns *he, him, his, himself, she, her, herself, it, its, itself*.

For the manual annotation of pronouns and antecedents, we randomly selected 161 threads containing *he, she* or inflected forms, and 24 threads with *it* or inflection. In this set, the length of threads varies between three and 78, with the average being 10 and median being 7. Table 1 gives more information on the distribution of thread length and pronoun frequency.

Finally, we note that 77 root tweets contain visual data (pictures, videos etc.), and 20 contain a quoted tweet[3]. Both of these aspects may potentially affect pronominal reference, as mentioned in the previous section.

3.2 Data Preparation

It is well known that tokenization is a crucial preparatory step for doing any kind of NLP on tweets. 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 the variety of challenges, such as:

- PTBTokenizer decides whether to split at apostrophes (whereas Twokenizer does not). For example:

  O'neill → O'neill
  d'Orsay → d'Orsay
  aren't → are, n't
  London's → London, 's
  The final example demonstrates the relevance of this feature for anaphora (or general reference) resolution.

- Twokenizer recognizes punctuation symbols such as sentence final full stop, question mark, exclamation mark, and also social media signs such as emoticons even if they are not surrounded by white space. PTBTokenizer was not designed to do this.

  U.S. → U.S.
  e.g., i.e. → e.g., i.e.
  here.Because → here, ., Because
  here:)Because → here, :), Because

We thus decided to use both systems: the output of Twokenizer is sent as input to the PTBTokenizer. One drawback of this approach might be duplicating over-tokenization errors. For instance, some URL forms such as *ftp://xxx.yyy* are considered as URL in Twokenizer, hence recognized as one token. But PTBTokenizer is not recognizing them as URLs and, therefore, divides them into smaller tokens. However, for our purposes, over-tokenization (i.e., producing too many tokens) is preferred to insufficient generation of token boundaries, because annotation tools (see below) can handle markables containing more than one token, but they do not allow for selecting a substring of a token as a markable.

3.3 Annotation

In our annotation scheme, we so far consider only the *identity* relation. With tweets being structurally relatively simple, we were interested in lean annotation guidelines, and followed the strategy defined in (Grishina and Stede, 2015), with some modifications in the treatment of predicative nouns and appositives. In our scheme, predicative nouns and appositions are considered as markables indicating reference identity. We defined additional attributes to differentiate these markables (i.e., copula constructions and appositives) from the other mentions. Also, we annotate the structural relation (anaphora, cataphora and exophora) of the pronouns, in order to cover the phenomena we will explain in Section 4. For exophora, additional more fine-grained categories are used:

[3]Sharing a tweet by adding new content "on top" of it: https://developer.twitter.com/en/docs/tweets/data-dictionary/overview/entities-object

Threads:	185
Coreference chains:	278
Annotated mentions:	1438
Annotated pronouns:	853
Annotated predicative NPs:	65
Length of longest coreference chain:	56
Average length of coreference chains:	5
Median length of coreference chains:	3
Intra-tweet coreference chains:	100
Inter-tweet coreference chains:	178
Threads with username or hashtag ref.:	43

Table 2: Descriptive statistics of annotations in the corpus

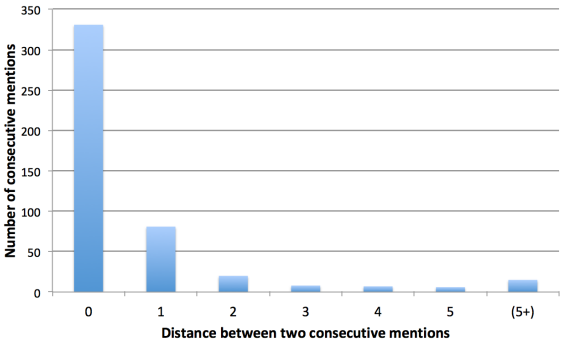

Figure 1: Distribution of distance between two consecutive items in reference chains

whether the antecedent is in the attached picture, quoted tweet, embedded link, or can be inferred by world knowledge.

Due to the data selection criteria, every thread contains at least one chain involving one or more 3rd person singular pronoun. For each pronoun, we annotated the complete reference chain (i.e., not just its antecedent). Hence, a chain can also include proper names and full NPs. The annotation tool is MMAX2 (Müller and Strube, 2006). Since it is important to know the authors of the tweets being annotated, both the user and the textual content of the tweet are shown together in the annotation window. Regarding the mention span, we do not allow discontinuous markables.

Since the annotation guidelines of (Grishina and Stede, 2015), on which ours are based, have already been evaluated with an inter-annotator agreement study (see that publication), we did not conduct one here. Our approach to quality control was that two annotators worked on separate files, but all chains marked by one annotator have been reviewed by the other, and were adjudicated when necessary. In a few cases (around 5), this did not lead to agreement; those threads were removed from the dataset. Altogether, our initial dataset of 225 threads shrank to the final size of 185 that we stated above. The majority of removed tweets were just incomprehensible or contained large portions of non-English content.

Table 2 gives an overview of the size of the annotations in the corpus. Also, to (partially) estimate the difficulty of the resolution problem, we calculated the distance for each consecutive pair of mentions in the coreference chains, in terms of the number of intervening turns. Figure 1 shows this information. Distance 0 means that the mentions are in the same tweet; 1 means they are in adjacent tweets, and so on. The longest distance values between a markable and its antecedent are 53, 37 and 19. In these chains the referring mention is either a definite NP or a named entity:

- Referring mention is a definite NP[4]:
  1:@10DowningStreet:*[The Prime Minister]$_i$ has started a refreh of [her]$_i$ ministerial team - updates will follow #CabinetReshuffle*
  ..
  54:@10DowningStreet:*[The PM]$_i$ with [her]$_i$ newest appointments to the Government Whips ' Office in Number 10 this afternoon #Reshuffle https://t.co/vgu9ioueu3*

- Referring mention is a named entity:
  17:@AustraliaToon1:@*cbokhove @oldandrewuk @Samfr @mikercameron I disagree with your analysis of [Andrew]$_i$'s form of arguing. ..*
  ..
  54:@littlewoodben:@*mikercameron @oldandrewuk .. What I find shocking , really shocking , is how [Andrew]$_i$ defends a man with a prolific history of odious misogynistic remarks. ..*

4 Pronominal anaphora in Twitter: Phenomena

Non-aligning replies A potential complication in any approach to analyzing Twitter conversations from a discourse perspective is possible mismatches between the `reply-to` ID and the

[4]If the conversational structure is important for demonstrating the phenomena, the examples are organized as follows: Tweet_order_in_thread:Username:Tweet_content

actual relation based on the contents of the tweets: In certain Twitter UIs, it may well happen that a user reads a sequence of related tweets, hits "reply" to tweet X, but then in fact responds to a different tweet Y in the neighborhood of X. We encountered a few clear cases in our threads. In general, they can obviously be hard to detect automatically, and it is not possible to estimate the frequency of the problem just on the basis of our relatively small sample. Hence we leave a deeper investigation for future work.

Hashtags In contrast to other social media conversations, Twitter offers the instrument of hashtags, which users employ gladly and frequently. Thus it is not surprising that hashtags can also work as referring expressions and hence as antecedents to pronouns. We distinguish two cases:

- Hashtag syntactically integrated:

  [# Oprah]$_i$ will be my favorite in 2020 selections. I will vote for [her]$_i$.

- Hashtag is not integrated:

  [She]$_i$ should be our president on 2020. [#Oprah]$_i$

The non-integrated case is challenging for annotation and automatic resolution, as this phenomenon is unknown from standard text. We decided to treat it on a par with cataphora (instead of looking for an antecedent in a previous tweet), assuming that hashtags at the beginning and end of tweets are textually-prominent entities.

Furthermore, we occasionally find references to substrings of hashtags, again with or without syntactic integration:

- *Let's #findClara, I hope she is safe.*
- *#findClara Our little girl is still missing. Please help us to find her.*

As we are doing a token based annotation and the hashtags are considered as single tokens in our scheme, we do not annotate these cases.

Usernames and display names These can act as referring expressions, too. Again, we find them both integrated in the syntax and disintegrated. The following example demonstrates how the username can become part of the syntax:
2:@Karen_LHL: *[@DannyZuker]$_i$ is funny*
3:@JanettheIntern: *@Karen_LHL Got [him]$_i$* !

Notice that in Twitter, the username of the replied tweet's writer is automatically added to the content of the reply message. Since this is not part of the text written by the user who is replying, we consider such usernames as part of the metadata of the tweet and outside the realm of reference annotation.

Multi-user conversations When more than two users are involved in a thread, 3rd person pronouns can refer to authors of previous messages. In those cases, we annotate the first occurrence of the username for the user being referred to as the referent for the pronouns. Then, the first (I), second (you) and third (he, she) person pronouns may refer to the same entity as indicated in section 2.1:
1: @realDonaldTrump: *[I]$_i$ 've had to put up with the Fake News from the first day [I]$_i$ announced that [I]$_i$ would be running for President. Now [I]$_i$ have to put up with a Fake Book, written by a totally discredited author. Ronald Reagan had the same problem and handled it well. So will [I]]$_i$*!
2: @shannao29522001: *[@realDonaldTrump]]$_i$ Stay strong. [You]$_i$ are our hero. I'm so proud to call [you]$_i$ MY president. As an educated female, I would be the first to stand up for [you]$_i$. I'm so tired of the fake news..* [..]
3:@Lisaword7: *@shannao29522001 @realDonaldTrump [He]$_i$ can quote things out [his]$_i$ mouth and you hear [him]$_i$. Come back two days later and say, fake news. [His]$_i$ base will agree with [him]$_i$.*[..]

As a complication, (part of) a Twitter username and (part of) a display name can be used interchangeably to refer to the same entity. For example:
@CBudurescu: *I have seen @[EdsonBarbozaJR]$_i$ fight and I have seen [@TheNotoriousMMA]$_j$ fight. I am pretty sure [Edson]$_i$ whoops [Conor]$_j$. Thats what @TeamKhabib meant when he said there are many fighters in lightweight division who would beat [Conor]$_j$. [Barboza]$_i$ is for sure one of them.*

In chain i, the display name of user @EdsonBarbozaJR is "Edson Barboza"; the parsing of either the display name or the username gives the relevant information that "Edson" refers to "Barboza".

In chain j, the display name of user @TheNo-

Antecedent in the attached media (threads):	12
Antecedent in the quoted tweet (threads):	3
Antecedent in the attached link (threads):	2

Table 3: Exophoric reference statistics

toriousMMA is "Conor McGregor". Here, unless we know what the display name is, it is not possible to relate @TheNotoriousMMA with "Conor", as the username itself gives no hint about this.

Exophoric reference On the one hand, this concerns the use of 1st and 2nd person pronouns as also mentioned as a natural result of multi-user conversations above:
1:@user1: *[[my]$_a$ aunt]$_i$ won't eat anything.*
2:@user2: *@user1 [[my]$_b$ aunt]$_j$ eats everything.*
3:@user3: *@user1 @user2 hope [[your$_{a/b?}$ Auntie]$_{i/j?}$ picks up soon.*

Resolving such coreference chains requires knowledge of tweet authors and of the *reply-to* structure.

On the other hand, as mentioned earlier, Twitter allows users to insert images, videos and URLs into their tweets. It is also possible to quote (embed) a previous tweet and comment on it.

For anaphora, this means that antecedents can be entities found in embedded images, videos, and even material somewhere in a referred URL or an embedded tweet, or its author. We annotate these anaphors where the antecedent is out of the current linguistic domain (i.e., the text of the tweet or its preceding tweets) as exophora, using the categories given in Table 3. As the numbers in the table show, in most cases of exophora the antecedents can be found in the attached pictures, as in the following example:
1:@LondonCouple2:*Few more of me on the way to work had to get the Train into day as Toms car in the Garage so he had to take mine did I sit opposite you today on the train if I did did u notice my stocking Xxx PICTURE_URL*
..
4:@cheknitallout:*@LondonCouple2 i know i would have enjoyed the view ! make eye contact , gesture her to show me more*

A final category of exophoric reference results from Twitter's listing the top keywords or hashtags being currently discussed ("trending topics") in the UI. For example, this a tweet that appeared after the 2017 Golden Globe awards:

Come onn! How can she be a president?!

Most probably, *she* refers to Oprah Winfrey, as her possible presidential candidacy was a trending topic emerging from the ceremony. In such cases, We annotate *she* as an exophoric type of pronoun and assign the attribute "antecedent can be inferred by world knowledge" (cf. Section 3.3).

There are cases where the antecedent of the pronoun is to be found in the text but it is ambiguous. In the example below, the ambiguity can be resolved only by inference:
1:@jessphilips:*Watching [@lilyallen]$_i$ and @stellacreasy stand their ground for last few days is inspiring for those who need resilience. Oh for the days of reasonable discourse where issues could be explored.*
2:@CorrectMorally: *@jessphillips @lilyallen @stellacreasy It started when [she]$_i$ insinuated Maggie Oliver was part of a right wing agenda to make Labour look bad. I couldn't let that go unchallenged, I'm surprised that you find [her]$_i$ stance so admirable, some of the things [she]$_i$ has said about the victims have been vile,*

The pronouns *she* and *her* in the second tweet are ambiguous as they can both refer to @lilyallen and to @stellacreasy. Knowing that @lilyallen's comments on some victims of a well known incident are criticized on the date of conversation and the second tweet has a reference to *victims*, the feminine 3rd person pronouns are inferred as referring to @lilyallen instead of @stellacreasy. This example is illustrating that all the participants are aware of the relevant discussions, so there is – presumably – no ambiguity in resolving the pronouns for them.

General Twitter challenges Finally, we mention some of the phenomena that are well-known problems in Twitter language, focusing here on those that can have ramifications for reference resolution.

- Typos affecting referring expressions: *@kennisgoodman: @Karnythia @TheRealRodneyF She not qualified to **he** president why?*
- Name abbreviations are frequent. E.g., *Barack Obama* can be referred to as *BO, O.*, etc.

- Missing apostrophe in contracted copula: *Hes my best.*
- Intentional misspellings: *Its **himmm** who does it.*
- Frequent elision, e.g., of subjects

5 Experiments

5.1 Setting

As a starting point for performing automatic anaphora resolution on the data set described above, we decided to test the performance of an off-the-shelf system. Thus we compared the output of the Stanford statistical coreference resolution system (Clark and Manning, 2015) with our manually annotated data. The input to the system was in an XML format that includes information on speakers and turns for each tweet.[5]

The Stanford resolver does not produce singletons in the output, and therefore, we also removed all singletons from our annotated data for this evaluation process. Further, we noted above that in our data we only annotated the coreference chains including 3rd person singular pronouns; other chains are left out of the scope of the annotation. In contrast, automatic resolution systems extract all the coreference chains in the input text. In order to make the Stanford resolver's output comparable to our annotations, we therefore needed to filter out some coreference chains in the resolver's output (viz. the chains with no 3rd person pronouns and the chains belonging to different entities than we annotated). Thus we extracted the coreference chains with the 3rd person singular pronouns and also the chains with at least one overlapping item with our mentions from the Stanford resolver's output and used only those chains for the evaluation.

5.2 Evaluation

In our experiments, the resolver's algorithm option is set to the value of "clustering". There is also an option for activating the "conll" settings in the Stanford resolver. When this setting is on, the resolver does not mark the predicative nominals and appositives, because in the CoNLL 11/12 shared tasks, these were not treated as markables[6]. We preferred to keep that setting off, as our dataset is annotated for those mentions.

Metric	Recall	Precision	F1
MUC	58.24	48.97	53.21
BCUBED	45.8	40.75	43.13
CEAFM	57.55	49.06	52.97
CEAFE	52.57	47.69	50.01
BLANC	43.27	21.25	28.49

Table 4: Evaluation results with speaker and turn info included in the input data (CONLL 2012 scorer)

Category	Count	%
Wrong/missing items in chain:	279	60
Missing chains:	55	12
Incorrect mention spans:	130	28
Total error count :	464	100

Table 5: Statistical information on error categories

The metric scores are calculated using the reference implementation of the CoNLL scorer (Pradhan et al., 2014). The results are given in Table 4[7]. It is difficult to compare them to results published on standard monologue text datasets, due to our selecting only a subset of the output chains. The scorer considers a mention to be correct only if it matches the exact same span in the manual gold annotations. The partial match scoring (e.g. checking the matching of heads for each phrase) might be more insightful for our data as the impact of differences in annotation schemes will be reduced by this way. The comparison of the metric results with each other may create more understanding on the strong and the weak aspects of the resolver on Twitter conversations. We leave the partial matching scoring and analysis of the differences in metric results as future work. However, for the present study, we made a qualitative analysis of the errors existing in the automated results and present them in the next section.

5.3 Error Analysis

We classified the errors in the automatically created coreference chains into 3 categories for which general statistical information can be found in Table 5.

[5] We also conducted experiments with the raw input text (i.e., with no speaker or turn info provided), but it is ongoing work to interpret the difference in the results we found.

[6] http://conll.cemantix.org/2012/task-description.html

[7] The numbers for MUC measure are different than the ones published in CRAC Proceedings at NAACL, 2018. There was a mistake in copying the relevant values to the table and we corrected those values in this version.

5.3.1 Wrong items or missed references in the chain

1. Wrong or missing antecedent in the chain:

   This error classification indicates that the pronouns are captured correctly in the chain but a wrong antecedent is assigned to them or no antecedent at all exists in a chain. This is a generic classification, there could be different reasons for these mismatches but as we didn't observe any clear pattern for the reasons of these wrong/missing assignments, we decided to present them in a generic classification.

   We observed that 39% of the errors in this category are of this type.

2. Missing matches due to lack of world knowledge:

   In the following thread, "Hillary Clinton" and "The Secretary of States" are referring to the same person, but this chain could not be captured correctly by the automated system due to the lack of knowledge that "Hillary Clinton" was "The Secretary of States".
   1:@TheRealJulian: *The only Russia collusion occurred when* ***Hillary Clinton*** *conspired to sell US Uranium to a Russian oligarch* [..] 12:@jolyeaker: ***The Secretary of States*** *should* [..]

   We observed that 23% of errors in this category are of this type.

3. First, second and third person pronouns corefer:

   Occasionally, first, second and third person pronouns are erroneously put in the same chains. Although coversation structure may in principle allow all these pronouns to refer to the same entity as indicated in Section 4, the chains we inspected do not seem to follow a logical selection mechanism on the input structure. A representative example is the one below, where the first person pronouns are put in the same chain with *@EricTrump* who is obviously not one of the conversation participants.
   1:@ALT_uscis:*[@EricTrump]$_i$, [his]$_i$ wife/guests wore sombreros during [his]$_i$ . . wait for it ... Mexican themed birthday party , while [his]$_i$ dad is DEPORTING THEM and wants to build a WALL on the border . ..*
   2:@ActualEPAFacts:@*ALT_uscis @EricTrump So , the irony [I]$_i$* get. [I]$_i$* am a 45 year old man whose family frequents a TexMex restaurant in DC . On my birthday , I have worn a sombrero a few times . It isn't unusual.*

   We observed that 15% of errors in this category are of this type.

4. Missing matches due to hashtags and at-sign:

   "#Borisjohnson", "@Borisjohnson" and "Boris" were not recognized as the same entity below:
   3:@angelneptustar:*To B sure* ***#Borisjohnson*** *held 4 huge consultations*
   ..
   7:@angelneptustar:.. *But sadly a raving anti semite , totally divisive.* ***@Borisjohnson*** *'s biggest achievement , he united London.*
   8:@WMDivision:.. *given* ***Boris*** *has published articles brimming ..*

   We observed that 7% of errors in this category are of this type.

5. Missing matches due to case sensitivity:

   The usage of upper and lower case in Twitter posts deviates from conventional usage in many forms. The resolver makes case-sensitive decisions, but the problems can lead to missing matches, such as in the next case where "LINDA SARSOUR" and "Linda Sarsour" were not recognized as the same entity:
   1:@yongaryisback:*#IranianProtests THE DEMOCRATS AND* ***LINDA SARSOUR*** *HATE THESE PROTESTS*
   2:@mattfwood:@*yongaryisback .. you do n't even look at her feed , you 'd see* ***Linda Sarsour*** *tweeting against ..*

   We observed that 2% of errors in this category are of this type.

6. Missing or wrong mention matches with unclear reason:

   This is a generic category to capture unclear cases of mention mismatches. We observe that 14% of errors in this category are of this type.

   We also observed errors due to the Twitter phenomena we presented in Section 4. Since

we don't have clear statistical information for these cases, we put these errors under this generic type. For instance, in the following example, both "he" and "his" refers to the same entity present in the attached media, but they were not put in the same chain by the resolver:
1:@MockingTheDraft: *Agree or disagree?* VISUAL_MEDIA_URL
3:@cmilner2: *@ChrisJBSnow @MockingTheDraft* ***He*** *'s 6' 3*
5:@bdbsport: *@ChrisJBSnow @cmilner2 @MockingTheDraft I 'm not saying anything until I hear* ***his*** *hand size.*

5.3.2 Missing chains

We are aware that the automated system that we tested against our data does not show singleton chains in the resulting files. But there are also non-singleton chains which do not appear in the automated results.

As indicated in Table 5, 12% of total errors are of this category.

5.3.3 Incorrect mention spans

1. Twitter names included in the span:
   Lists of usernames and hashtags in tweets can cause difficulties for the resolver. This holds in particular for the automatically-added usernames (mentioned in Section 3), which can erroneously be identified as antecedents. Removing these elements from the text could thus be an effective preprocessing step. But in general, usernames, display names and hashtags can also be used as linguistic constituents in the way that we mentioned in Section 4. Therefore, the preprocessing should be done with this consideration.
   7:@ToddXena:*@TippyStyle @nedryun there is a lot of " noise .. I would suggest is go back research the [Reagan]$_i$ years ..*
   8:@TippyStyle:*[@ToddXena @nedryun Todd Reagan]$_i$* actually had early onset Alzheimer 's during his presidency . Not giving me the warm an fuzzies here.*
   We observed that 36% of errors in this category are of this type.

2. Miscellaneous mention span errors:
   There are variety of errors with selecting the mention span, such as including emoticons[8] or unnecessary punctuations in the span.
   We observed that 64% of errors in this category are of this generic type.

[8] https://en.wikipedia.org/wiki/Smiley

6 Conclusions

Twitter conversations have so far not received much attention from the perspective of coreference or anaphora resolution. We argued that this genre shares certain features with other social media, multi-party chat, but also with spoken language. We explained how we constructed a corpus of 185 conversation threads, and what decisions we made in annotating pronominal anaphora on this somewhat unusual genre. A number of specific challenges surfaced in our annotation work, and we explained how we responded to them. Finally, we reported on our first experiments with an off-the-shelf resolution system (Stanford), showing the results as well as an error analysis. Our next steps are to experiment with different variants of preprocessing for measuring the effect on the resolver performance, and then conclude what fundamental problems remain for a resolver trained on "standard" text, when being confronted with this genre, and how they may be tackled.

Acknowledgements

We are grateful to Constanze Schmitt for her help in the annotation and qualitative error analysis. Our work was funded by the Deutsche Forschungsgemeinschaft (DFG), Collaborative Research Centre SFB 1287, Project A03.

References

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)*, pages 1405–1415, Beijing, China. Association for Computational Linguistics.

Adèle Désoyer, Frédéric Landragin, Isabelle Tellier, Anas Lefeuvre, and Jean-Yves Antoine. 2016. Coreference resolution for french oral data: Machine learning experiments with ancor. In *7th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing)*, Konya, Turkey.

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.

Yulia Grishina and Manfred Stede. 2015. Knowledge-lean projection of coreference chains across languages. In *In Proceedings of the 8th Workshop on Building and Using Comparable Corpora*, Beijing, China. Association for Computational Linguistics.

Iris Hendrickx and Vronique Hoste. 2009. Coreference resolution on blogs and commented news. In *Anaphora Processing and Applications*, Lecture Notes in Artificial Intelligence 5847, pages 43–53. Springer, Berlin/Heidlberg.

Prateek Jain, Manav Ratan Mital, Sumit Kumar, Amitabha Mukerjee, and Achla M. Raina. 2004. Anaphora resolution in multi-person dialogues. In *Proceedings of the 5th SIGdial Workshop on Discourse and Dialogue*, pages 47–50, Cambridge, Massachusetts, USA. Association for Computational Linguistics.

Xiaoqiang Luo, Radu Florian, and Todd Ward. 2009. Improving coreference resolution by using conversational metadata. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers*, pages 201–204, Boulder, Colorado. Association for Computational Linguistics.

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 Sabine Braun, Kurt Kohn, and Joybrato Mukherjee, editors, *Corpus Technology and Language Pedagogy: New Resources, New Tools, New Methods*, pages 197–214. Peter Lang, Frankfurt a.M., Germany.

Nobal B. Niraula and Vasile Rus. 2014. A machine learning approach to pronominal anaphora resolution in dialogue based intelligent tutoring systems. In *Computational Linguistics and Intelligent Text Processing*, pages 307–318, Berlin, Heidelberg. Springer.

M. Poesio, A. Patel, and B. Di Eugenio. 2006. Discourse structure and anaphora in tutorial dialogues: An empirical analysis of two theories of the global focus. *Research on Language and Computation*, 4(2-3):229–257.

Sameer Pradhan, Xiaoqiang Luo, Marta Recasens, Eduard Hovy, Vincent Ng, and Michael Strube. 2014. Scoring coreference partitions of predicted mentions: A reference implementation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 30–35, Baltimore, Maryland. Association for Computational Linguistics.

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

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

David Schlangen, Timo Baumann, and Michaela Atterer. 2009. Incremental reference resolution: The task, metrics for evaluation, and a Bayesian filtering model that is sensitive to disfluencies. In *Proceedings of the SIGDIAL 2009 Conference*, pages 30–37, London, UK. Association for Computational Linguistics.

Amanda Stent and Srinivas Bangalore. 2010. Interaction between dialog structure and coreference resolution. In *Proceedings of the IEEE Spoken Language Technology Workshop*, pages 342–347. IEEE.

Michael Strube and Christoph Müller. 2003. A machine learning approach to pronoun resolution in spoken dialogue. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 168–175, Sapporo, Japan. Association for Computational Linguistics.

Joel Tetreault and James Allen. 2004. Semantics, dialogue, and pronoun resolution. In *Proceedings of the SemDial Conference (CATALOG)*, Barcelona, Spain.