Annotating If Authors of Tweets are Located in the Locations They Tweet About

Vivek Doudagiri, Alakananda Vempala and **Eduardo Blanco**
Human Intelligence and Language Technologies Lab
University of North Texas
VivekReddyDoudagiri@my.unt.edu, AlakanandaVempala@my.unt.edu, eduardo.blanco@unt.edu

Abstract

The locations in a tweet do not always indicate spatial information involving the author of the tweet. In this paper, we investigate whether authors are located or not located in the locations they tweet about, and temporally anchor this spatial information in the tweet timestamp. Specifically, we work with temporal tags centered around the tweet timestamp: longer than 24 hours before or after tweeting, within 24 hours before or after tweeting, and at the time of tweeting. We introduce a corpus of 1,200 location mentions from 1,062 tweets, discuss several annotation samples, and analyze annotator disagreements.

Keywords: Spatial knowledge, Social Networks, Twitter

1. Introduction

Twitter has quickly become one of the most popular social media sites. It has 313 million monthly active users, and 500 million tweets are published daily. People tweet about their current locations (e.g., *I'm at Walgreens in Anaheim, California*), as well as past (e.g., *I miss living in Kerman*) and (probable) future locations (e.g., *Can't wait to visit Italy!*). Tweets often contain hints regarding how long the author is in a particular location, either implicitly (e.g., people usually stay at pharmacies such as Walgreens for a few minutes to an hour, not days) or explicitly (e.g., *My 2 day vacation to San Diego beach starts tomorrow*).

In this work, we present a corpus annotating whether the author of a tweet is located in the locations mentioned in his tweets. Going beyond named entity recognition, we annotate whether the author is located or *not* located in the location he tweets about with respect to the time he tweeted (before, during and after). To the best of our knowledge, this problem has not been explored before. We found that no spatial relationship can be inferred between authors of tweets and the locations they tweet about in 21% of instances. In other words, 1 out of 5 locations in a tweet do not indicate any spatial information about the author.

The major contributions of this paper are:

1. We create a corpus of 1,062 tweets containing 1,200 location named entities, and annotate whether the authors are located or *not* located in those locations with respect to the time they tweeted (when the author tweeted, within 24 hours before and after he tweeted, and longer than 24 hours before and after he tweeted).[1]

2. We provide several annotation examples and the label distributions per temporal tag.

3. We present detailed inter-annotator agreement calculations, including Cohen's κ and confusion matrices per temporal tag.

2. Previous Work

The work presented here is inspired by Li and Sun (2014), who work with points of interest in tweets using the Foursquare API (their points of interest are similar to our locations). Li and Sun determine if the author of a tweet is present at a point of interest in the past, present or future with respect to the tweet timestamp (three binary decisions). In their corpus, 47.3% of points of interest are annotated invalid, meaning that their methodology to extract points of interest using Foursquare is not very effective. In contrast to their work, we (a) present a corpus with few invalid locations ($\approx 6\%$), and (b) work with finer-grained temporal information (when somebody tweets, within 24 hours before and after he tweeted, and longer than 24 hours before and after he tweeted).

Jurgens et al. (2015) present a thorough evaluation of nine state-of-the-art network-based approaches to perform geolocation inference. They propose several evaluation methods, discuss possible sources of ground truth and their soundness, and conclude, among others, that real-world performance is much lower than initially reported. Mahmud et al. (2014) extract home locations of Twitter users at different granularities (e.g., city, state, time zone, geographic region). Their approach is a combination of statistical classifiers and heuristics, and takes into account the content of tweets (the actual words). Unlike them, we annotate spatial information from any location a person tweets about, we do not target the place of residence. Jurgens (2013) shows that social relationships help determining locations. They present an algorithm grounded on propagating spatial information through a user's social network. The algorithm does not rely on the specifics of any social network (the only requirement is that there are social relationships), and pinpoints the location of 50% of the users in a Twitter-based social network within 10 km. In contrast to this previous work, we work with specific locations mentioned in a tweet, and annotate whether the author was located there with respect to the time he tweeted.

[1] Available at https://alakanandav.bitbucket.io/

	Tweet	Location	Before >24	Before <24	During	After <24	After >24
1	I'm at Walgreens in Anaheim, Calif	Calif	PY	CY	CY	CY	PY
2	Just got home from Vegas and I'm cooking omg	Vegas	CY	PY	CN	PN	UNK
3	First time in Squaw Valley and it could not have been more perfect! #squawvalley	Squaw Valley	CN	CY	CY	CY	UNK
4	I adopted a child while in Mexico	Mexico	CY	PN	CN	PN	UNK
5	Found some really cute couches in Oakland and did not have a car big enough to carry them back RIP	Oakland	PY	CY	CN	PN	UNK
6	Tomorrow we are driving to Yosemite Valley	Yosemite Valley	UNK	PN	CN	PY	PY
7	I can't wait to be in Hawaii permanently.	Hawaii	UNK	UNK	UNK	UNK	UNK
8	Bruh im bouta fall asleep on Keyasia	Keyasia	INV	INV	INV	INV	INV

Table 1: Annotation examples. We show the original tweet, the location being annotated (we only detail one location per tweet), and the labels for the five temporal tags.

3. Corpus Creation

In this section, we detail the creation of the corpus. First, we present the steps to gather tweets and select locations to be annotated. Second, we describe the annotation process and the kind of spatial information annotators were asked about. Finally, we present annotation examples.

3.1. Selecting Tweets and Locations

We collected tweets containing at least one location named entity following 4 steps:

1. We downloaded over one million tweets published from California along with their metadata using the Twitter API.[2] Then, we discarded tweets (a) consisting of less than 3 tokens, or (b) not containing at least one pronoun.[3]

2. We extracted named entities using spaCy[4] and Stanford CoreNLP (Manning et al., 2014) after removing emoticons, URLs and newline characters from the original tweets. The corpus contains both the original tweet and the preprocessed version.

3. We identified locations in tweets if both Spacy and Stanford CoreNLP recognized a LOC or GPE named entity (location and geopolitical named entities respectively) spanning exactly the same tokens.

4. Finally, we randomly selected 1,200 locations from 1,062 tweets for annotation.

3.2. Annotation process

For each location in the selected tweets, we asked annotators the following question: "Is the author of the tweet present in the location . . . ":

1. at any point of time earlier than 24 hours before tweeting (Before $>$ 24)?

2. at any point of time within 24 hours before tweeting (Before $<$ 24)?

3. at the time he tweeted (During)?

4. at any point of time within 24 hours after tweeting (After $<$ 24)?

5. at any point of time later than 24 hours after tweeting (After $>$ 24)?

We decided to work with these temporal tags because people usually tweet about what is happening, about what has happened recently, or about what is about to happen (Sanagavarapu et al., 2017).

We allow annotators to choose from six labels inspired by previous work on factuality (Saurí and Pustejovsky, 2012):

- Certainly Yes (CY): I am certain that the author is located in the given location at the specified time.

- Probably Yes (PY): I am not certain that the author is located in the given location at the specified time, but it is probably the case.

- Certainly No (CN): I am certain that the author is *not* located in the given location at the specified time.

- Probably No (PN): I am not certain that the author is *not* located in the given location at the specified time, but it is probably the case.

- Unknown (UNK): There is not enough information to answer any of the four labels above.

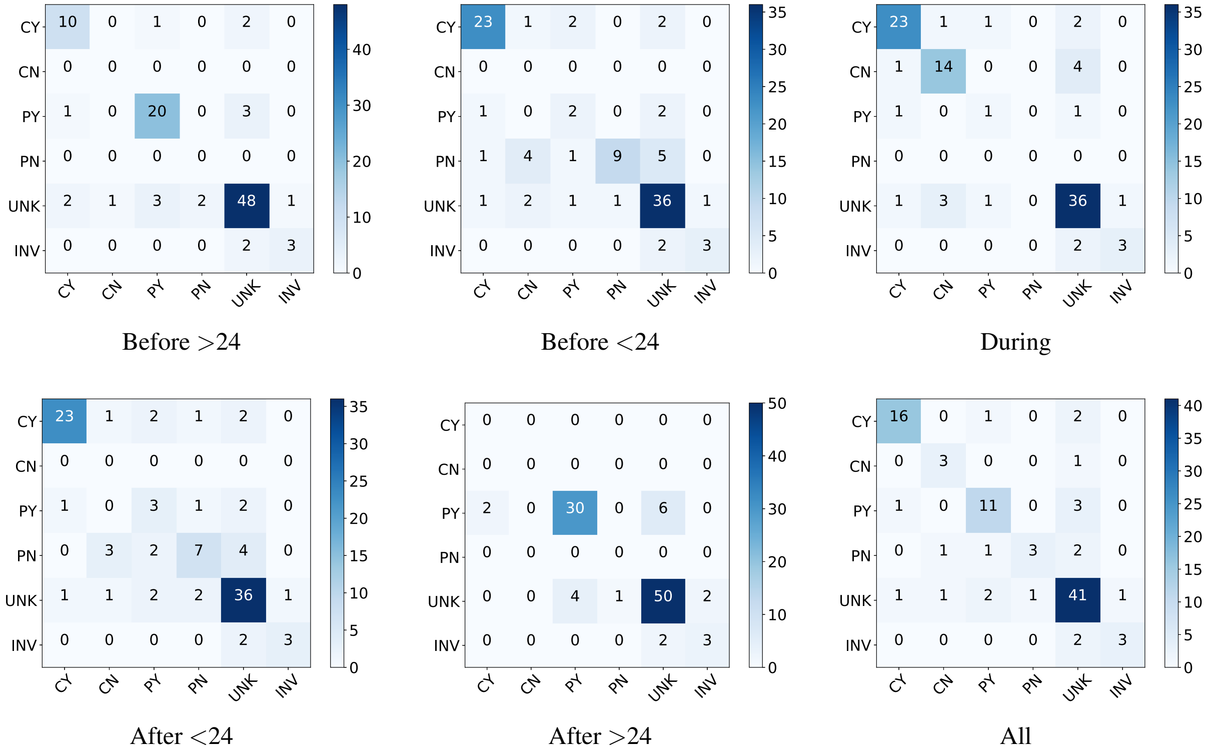
- Invalid (INV): The location is invalid, it is nonsensical to ask whether the author is (or is not) located there.

The entire corpus was annotated independently by two graduate students. Disagreements were adjudicated after in-person discussions between both annotators. Section 4.1. details the inter-annotator agreements.

[2] `https://developer.twitter.com/en/docs`

[3] Full list of pronouns: *I*, *we*, *you*, *he*, *she*, *it*, *they*, *him*, *her* and *them*. Tweets containing these pronouns are more likely to be authored by an individual (as opposed to companies or institutions).

[4] `https://spacy.io/`



	CY	CN	PY	PN	UNK	INV
CY	10	0	1	0	2	0
CN	0	0	0	0	0	0
PY	1	0	20	0	3	0
PN	0	0	0	0	0	0
UNK	2	1	3	2	48	1
INV	0	0	0	0	2	3

Before >24

	CY	CN	PY	PN	UNK	INV
CY	23	1	2	0	2	0
CN	0	0	0	0	0	0
PY	1	0	2	0	2	0
PN	1	4	1	9	5	0
UNK	1	2	1	1	36	1
INV	0	0	0	0	2	3

Before <24

	CY	CN	PY	PN	UNK	INV
CY	23	1	1	0	2	0
CN	1	14	0	0	4	0
PY	1	0	1	0	1	0
PN	0	0	0	0	0	0
UNK	1	3	1	0	36	1
INV	0	0	0	0	2	3

During

	CY	CN	PY	PN	UNK	INV
CY	23	1	2	1	2	0
CN	0	0	0	0	0	0
PY	1	0	3	1	2	0
PN	0	3	2	7	4	0
UNK	1	1	2	2	36	1
INV	0	0	0	0	2	3

After <24

	CY	CN	PY	PN	UNK	INV
CY	0	0	0	0	0	0
CN	0	0	0	0	0	0
PY	2	0	30	0	6	0
PN	0	0	0	0	0	0
UNK	0	0	4	1	50	2
INV	0	0	0	0	2	3

After >24

	CY	CN	PY	PN	UNK	INV
CY	16	0	1	0	2	0
CN	0	3	0	0	1	0
PY	1	0	11	0	3	0
PN	0	1	1	3	2	0
UNK	1	1	2	1	41	1
INV	0	0	0	0	2	3

All

Figure 1: Inter-annotator confusion matrices (percentages for all pairs of labels from both annotators) per temporal tag. Note that annotators almost never disagree between (CY, PY) and (CN, PN).

3.3. Annotation Examples

Table 1 presents sample tweets and the annotations for all temporal spans. We briefly interpret the annotations below:

- Regarding tweet 1, annotators understood that the author is certainly in *Calif* when he tweeted, and also at some point of time within 24 hours before and after (one cannot leave *Calif* in a split second). However, they annotated that the author is probably still in *Calif* at some point of time 24 hours before and after, as it is not guaranteed that they author is there for an extended period of time.

- In tweet 2, the author describes the return from a trip to *Vegas*. Annotations reveal that the author is not in *Vegas* when he tweeted or shortly after (after <24, people usually don't travel to to the same destination within a day), but it is unknown whether he will go back in the long run (after >24). Annotations also reveal that the author was in *Vegas* prior to tweeting.

- In tweet 3, the author is also tweeting about a trip, this time to *Squaw Valley*. Annotators understood that the author tweeted shortly after arriving (before <24: CY) and he had not been there earlier (before >24: CN, keywords: *First time in*). They also annotated that the author is in *Squaw Valley* within 24 hours after tweeting, or in other words, that he was not leaving when he tweeted. Finally, there is not information to determine whether he was (or will be) there 24 hours after tweeting (after >24: UNK).

- Tweet 4 describes a past event. It is clear that the author was a participant in the event, and thus he was in *Mexico* at some point of time in the past. Annotators were certain that the author (a) was not in *Mexico* when he tweeted and (b) he was in *Mexico* at some point of time 24 hours before he tweeted (presumably adopting a child takes longer than a day).

- In tweet 5, the author describes a recent event (*Found some really cute couches*) that took place in *Oakland*. Annotators interpreted that *finding couches* occurred in the immediate past (before <24: CY), and maybe in the more distant past (before >24: PY). They also understood that the author left *Oakland* before tweeting (during: CN).

- Tweet 6 describes future plans for a trip to *Yosemite Valley*. Annotators understood that the author is likely to be there within 24 hours of tweeting (after <24: PY), and he will stay for longer than a day (after >24: PY). They also annotated that the author is certainly not there when he tweeted, and probably not there within a day before tweeting (before <24: PN). Finally, they couldn't determine if the author has or has not been at *Yosemite Valley* prior to 24 hours from tweeting (before >24: UNK).

- Tweet 7 describes a (possible) future state (the author wishes to *be in Hawaii*). Unlike in tweet 6, where the author appears to have made plans to drive to *Yosemite Valley* the day after tweeting, in tweet 7 it is not clear

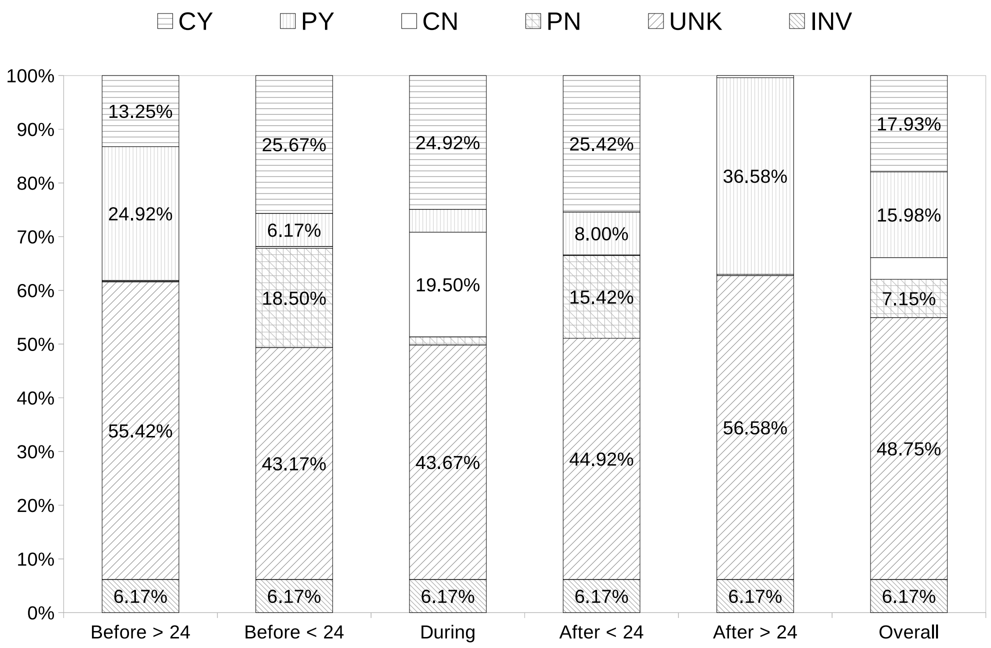


Figure 2: Label distribution per temporal tag and overall distribution after adjudicating disagreements. Percentage values are not shown if they are lesser than 5 %

	Before > 24	Before < 24	During	After < 24	After > 24
α	0.71	0.72	0.73	0.66	0.70

Table 2: Cohen's kappa (k) agreements per temporal tag.

if the author is simply stating a wish. Thus, annotators chose UNK for all temporal tags.

- Finally, tweet 8 contains an invalid location. *Keyasia* is a girl's name, but it was selected for annotation because both spaCy and Stanford CoreNLP tagged it as a location.

4. Corpus Analysis

In this section we analyze the corpus. Specifically, we discuss the quality in terms of inter-annotator agreement and the label distribution. Our corpus contains 1,200 location mentions extracted from 1,062 tweets (i.e., 1.14 locations per tweet on average).

4.1. Quality

Figure 1 presents confusion matrices per temporal tag and for all temporal tags (each matrix shows percentages for all pairs of labels from both annotators). Note that disagreements are relatively minor: most disagreements are between CY and PY, CY and UNK, CN and PN, CN and UNK, or UNK and INV. In other words, while annotators do not agree all the time, the sources of disagreements are not a major source of concern.

Table 2 presents inter-annotator agreements prior to adjudicating disagreements. The overall Cohen's kappa coefficient is 0.71, which is considered substantial (Landis and Koch, 1977). Coefficients range between 0.66 and 0.73 depending on the temporal tag. The highest agreement was obtained when annotating whether the author is (or is not) in a location when he tweeted (*during*: 0.73), and the lowest agreement when annotating whether the author is (or is not) in a location within 24 hours after tweeting (*after <24*: 0.66). The remaining agreements range from 0.70 to 0.72.

4.2. Label Distribution

Figure 2 presents the label distribution per temporal tag and overall after adjudicating disagreements. Overall, labels that allow us to extract spatial knowledge (CY, PY, CN and PN labels) account for 45.08% of labels. This percentage is similar across temporal tags, although we observe larger percentages of UNK with *before >24* and *after >24* (55.42% and 56.58% vs. 43.17%–44.92%). Authors are more likely to be located at the locations they tweet about than *not* located (CY+PY: 33.91% vs. CN+PN: 11.15%).

Overall, we can infer the location of authors of tweets with around 20% certainty (CY+CN). For within a day before, during and within a day after posting the tweet, we can infer the location of the author with greater certainty (CY+CN: 26%, 44.42%, 26% vs. CY+CN: 13.25%, <5%). Also, we can infer the location of the author with greater certainty for before rather than after posting the tweet.

5. Conclusions

In this paper, we present a corpus of tweets annotated with temporally-anchored spatial information involving the author. Specifically, we annotated whether authors of tweets are located in the locations they tweet about when they tweet, within 24 hours before or after tweeting, and longer than 24 hours before or after tweeting.

The corpus has substantial inter-annotator agreement. Label distributions indicate that many locations present in tweets do not indicate any spatial information about the author. Additionally, annotators were more certain (CY and CN labels) when annotating spatial information for temporal tags close to the tweet timestamp (during, and within 24 hours before and after) than for more distant temporal tags (longer than 24 hours before and after).

6. Bibliographical References

Jurgens, D., Finethy, T., McCorriston, J., Xu, Y. T., and Ruths, D. (2015). Geolocation prediction in twitter using social networks: A critical analysis and review of current practice. *ICWSM*, 15:188–197.

Jurgens, D. (2013). That's what friends are for: Inferring location in online social media platforms based on social relationships. *ICWSM*, 13(13):273–282.

Landis, J. R. and Koch, G. G. (1977). The measurement of observer agreement for categorical data. *biometrics*, pages 159–174.

Li, C. and Sun, A. (2014). Fine-grained location extraction from tweets with temporal awareness. In *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*, pages 43–52. ACM.

Mahmud, J., Nichols, J., and Drews, C. (2014). Home location identification of twitter users. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 5(3):47.

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. J., and McClosky, D. (2014). The stanford corenlp natural language processing toolkit. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL)*.

Sanagavarapu, K. C., Vempala, A., and Blanco, E. (2017). Determining whether and when people participate in the events they tweet about. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 641–646, Vancouver, Canada, July. Association for Computational Linguistics.

Saurí, R. and Pustejovsky, J. (2012). Are you sure that this happened? assessing the factuality degree of events in text. *Comput. Linguist.*, 38(2):261–299, June.