Leveraging Spatial Community Information in Location Recognition in Tweets

Phuc Nguyen
University of North Texas
Denton, Texas
phucnguyen2@my.unt.edu

Yan Huang
University of North Texas
Denton, Texas
yan.huang@unt.edu

Joshua R. Trampier
Department of Defense
Springfield, Virginia

ABSTRACT

Location names are very helpful in event extraction. Informal social texts pose significant challenges for recognizing location names. However, social texts have an advantage that can be leveraged: spatial and social network contexts. We address the location recognizing task as a part of named entity recognition, and introduce a new approach which leverages community contexts and captures language variations among groups of users. Specifically, we incorporate a community component into a topic modeling method and harness unlabeled tweets. Experiments on a large Twitter dataset show that our proposed method can improve the location classification F1 score by 5%.

KEYWORDS

location recognition, twitter, topic modeling, named entity recognition

ACM Reference Format:
Phuc Nguyen, Yan Huang, and Joshua R. Trampier. 2017. Leveraging Spatial Community Information in Location Recognition in Tweets. In *LENS'17: LENS'17:1st ACM SIGSPATIAL Workshop on Analytics for Local Events and News , November 7–10, 2017, Los Angeles Area, CA, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3148044.3148048

1 INTRODUCTION

With 313 million monthly active users, Twitter[1] is one of the busiest microblogs worldwide. Its terse nature–each tweet is at most 140 character length–encourages users to share concise pieces of texts telling us what is happening, e.g., social activities (e.g., scientific conferences, jazz festivals, game releases), natural disaster occurrences (e.g., earthquakes, tornadoes, winter storms), or incidents (e.g., traffic jams, fires, shootings). Identifying location mentions in text plays an essential role for detecting events and locating them from Twitter [1, 21, 23]. This paper presents location recognition as a part of named entity recognition (NER), and proposes an approach to improve the NER performance on location names.

Traditional NER methods on formal texts have been adopted and improved on short texts. The improvements include gathering context features from multiple tweets [15, 20], exploiting external knowledge from gazetteers and Freebase [15, 20, 24], solving tweet normalization and (or) entity linking together with NER [6, 16, 24]. However, NER performance on short texts is still lagging behind that on formal texts. In spite of some well-known challenges, Twitter brings more metadata, social connections, and interactions, which potentially provide further contextual information. This paper reports a generative model harnessing community information and unlabeled tweets in order to improve NER on location. Our approach captures the intuition that mentioning conventions of entities especially locations are often well understood among a group of users but unclear to the outsiders. For example, in the following tweet, "GAB" refers to the General Academic Building on a campus. GAB is not found in any gazetteer and shallow parsing is unlikely to help identifying GAB as a location.

- NCA Future Pros Carrer Panel Tonight in the Black Box! *GAB* 321.

One way to deal with this is to find other tweets that contain GAB as contexts to provide distant supervision. However, the top three tweets using a simple search do not help at all:

- Tried to imitate the Jonah pose... Almost had it. Photobombed by *gab*
- at least *gab* just snorted pencil shavings
- Who's paying the bill? Report casts questions on *GAB* and John Doe payments

When we limit the search to the related spatial area, all of the top three tweets refer to the same location and some of the mentions can be recognized using part of speech and shallow parsing. This example shows that spatial connections can be found among the content of the tweets from a region.

- @UNTCSA General Body meeting tomorrow *GAB* 310 Taste of Di Island" and Gaza Vs Gully Discussion pome out for a great time
- After the @untpbso meeting make sure you stop by @UNTPJ at 9:30PM outside the *GAB*!
- Caribou Coffee, served at Cafe *GAB*, is the newest addition to campus coffee choices

The idea is to allow users from a community to help each other in location recognition. A *community* refers to a group of users who share some interests, and (or) have close social relations, and (or) are located in a particular geographic area. In this paper we focus on community information by proximity in space. We look for linguistic variations in tweets posted by users from spatial communities and use them for location entity classification, a later phase in NER. Specifically, our contributions are summarized as follows:

[1] https://about.twitter.com/company, 06/30/2016

LENS'17, November 7–10, 2017, Los Angeles Area, CA, USA

ACM ISBN 978-1-4503-5500-1/17/11…$15.00
https://doi.org/10.1145/3148044.3148048

(1) We propose C-LLDA, a named entity classification model that integrates community information into a Labeled LDA model [18] to leverage cross tweet language similarity inside a community. The model is distant supervised and hence can leverage unlabeled data.
(2) We tested the proposed model on about 2 millions tweets posted by 6,940 users. In our experiments, we use geographically defined communities where a city is considered as a community. The results show that our model can improve the location classification F1 score by 5% and overall F1 score for all entity categories by 1.6%.
(3) We generated and share a corpus of 2,000 labeled tweets through our intuitive annotation tool for Amazon Mechanical Turk. The tool is carefully designed to help the workers understand the context better and label the tweets with ease. The workers' majority agreement is above 85% for locations and persons; and is ~ 60%-70% for organizations and miscellaneous.

The paper is organized as follows. Section 2 formalizes the problem of location recognition as a part of NER, outlines essential components of the NER system, and describes the incorporation of community information into a Labeled LDA model. In Section 3, we detail the Twitter data collection and annotation, and report the performance of C-LLDA. Section 4 summarizes previous investigations on NER with a focus on location names, and discusses related work on geographical topic models. We conclude the paper and bring up future work in Section 5.

2 COMMUNITY-BASED APPROACH

A tweet is an 140-character message which often consists of one or two sentences. Sometimes a tweet just contains an exclamation or a phrase. Most of them use simple grammar structures and informal wordings, including Internet slangs (e.g. afaik, ymmv, lol) and emoticons. Although tweets are short and informal, groups of tweets often collectively describe happenings and events.

2.1 Problem Definition

The problem that this paper addresses is: Given a set of tweets and the locations where the authors of tweets are located, segment and classify all location mentions in the tweets as accurately as possible. The location recognition is considered as as a part of named entity recognition (NER) which also recognizes organization names and personal names. The constraint is that the performance of recognizing other entities stays the same or better even with methods specified targeted at improving location recognition performance.

NER itself can be divided into two subtasks: segmentation and classification (as shown inside the rounded rectangle in Figure 1).

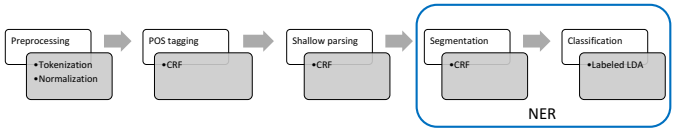

Figure 1: Components of a named entity recognition system.

ST1. Named entity segmentation (NES): Given a tweet, the named entity segmenter labels every token in the tweet as a part of an entity or not.

Example: "Now the $[Giants]_{ENTITY}$ are four games behind the $[New\ York\ Mets]_{ENTITY}$ for the $[National\ League]_{ENTITY}$ wild-card spot with seven to play."

Here, a *token* is an instance of a *word* or a *punctuation*. For example, in the sentence "AT&T *Stadium* is the home *stadium* of the Dallas Cowboys football team," there are two tokens of the word "stadium."

Name entity segmentation is not the focus of this work but an indispensable step in NER. We utilize an existing segmenter [20] which uses a conditional random fields model based on orthographic, contextual, dictionary, POS, and chunking features.

ST2. Named entity classification (NEC): Given a tweet with identified entities, the named entity classifier assigns an entity type label to every entity.

Example: "Now the $[Giants]_{ORG}$ are four games behind the $[New\ York\ Mets]_{ORG}$ for the $[National\ League]_{ORG}$ wild-card spot with seven to play."

This paper proposes an approach to improve NEC performance in four entity types: location (LOC), organization (ORG), person (PER), and other (MISC).

2.2 The Pipeline

We first describe the general framework that our proposed community-based location recognition method is based on. The tweets are firstly split into tokens. Then, a conditional random field (CRF) classifier labels the part-of-speech (POS) of each token using Brown clusters [4], POS dictionaries, spelling, and contextual features. Next is to identify noun phrases, verb phrases, and prepositional phrases (shallow parsing) using another CRF model with Brown clusters, POS tags, and features described in [22]. Labels from the POS tagger, shallow parser, and Brown clusters are combined with the orthographic, contextual, and dictionary features in a CRF classifier for named entity segmentation. These modules are connected together as a pipeline as shown in Figure 1. Next section discusses the details of the module for named entity classification which is the focus of our method.

2.3 Incorporating Community Information in Entity Classification

We firstly discuss Latent Dirichlet allocation (LDA), the model which our classifier is built upon. LDA [3] is a generative statistical model that discovers the underlying structure of a collection of observations. When observations are words collected into documents, it posits that each document is a mixture of topics, and topic distribution is assumed to have a Dirichlet prior. In LDA models, a document is generated in the following fashion: (1) decide on the number of words N the document will have, e.g. from a Poisson distribution; (2) choose a topic mixture for the document from a multinomial distribution over a fixed set of K topics; (3) generate each word w_n in the document by: first picking a topic from the multinomial distribution used before; then use the topic to sample a

Left context	Entity	Right context
Hundreds of graduate workers, students, & faculty rally to demand	UChicago	administration stop stalling vote on unionization
Interesting study from my colleague here at	Uchicago	...
@Aztec_Daves come back to	uchicago	we miss you!
Bring yo dog to	UChicago	this weekend
	UChicago	was a lot like a Led Zeppelin tour, in that people wouldn't stop talking about Tolkien

Table 1: Tokens surrounding the same name are to be collected as a document.

word from the topicâĂŹs multinomial distribution over the vocabulary. LDA model learns the parameters from existing documents and tries to backtrack a set of topics that are likely to have generated the collection.

uchicago, sonja, woods, francesco, woody, canes, ray charles, billy dec, forest hills, errol morris, francesca, kodak, white castle, billy crystal, china, michael harvey, golden, toronto sun, stern, dna, dns, bon jovi, yahoo, verizon fios, intake, matilda, wang, hyatt, anthony mackie, orchard beach

Table 2: Example of the 30 most frequent named entities identified by a segmenter.

We adopt a framework for named entity classification and incorporate community information into an LDA model. First, we use the segmenter mentioned in the previous section to identify entities in unlabeled tweets. The most frequent entities (as exemplified in Table 2) then are used to build a vocabulary of the most frequent tokens found in [-3, +3] windows: three tokens to the left and three tokens to the right of every entity. Tokens extracted from context of the instances of the same entity, as shown in Table 1, are combined together into an *entity document*. The *topics* of this entity document indicate the potential *types* of the corresponding entity. The goal of the model is to classify the instances of entities in each entity document into a topic which is the type of the entity.

In designing the LDA model for location recognition, we strive to capture the two intuition: 1) entity documents exhibit multiple topics/types, and these topics/types can be constrained by a set of possible entity types; and 2) entity documents also display linguistic characteristics of the author communities that need to be leveraged. We propose the following generative process of the entity documents:

(1) Randomly choose a distribution over communities,
(2) Randomly choose a distribution over topics/types,
(3) For each word in the entity document:
- Randomly choose a community from the distribution in step #1,
- Randomly choose a topic/type from the distribution over topics/types in step #2,
- Randomly choose a word from the corresponding distribution over the vocabulary of the chosen community.

The aforementioned process is illustrated by Figure 2. The example shows how a *document* regarding named entity "UChicago" might be generated by the model. For every token in the document: (1) the community assignment is sampled from $P(community)$. (2) the topic/entity type assignment is sampled from $P(entity\ type)$, which is constrained by the dictionaries. In our example, the entity "UChicago" appears in the dictionaries as either a location or an organization; therefore, the model limits the choices for entity type assignments among $\{LOC, ORG\}$. Finally, (3) the token is sampled from $P(word|entity\ type, community)$, given the community and entity type assignments. Note that not all words of the vocabulary, as well as other named entity *documents* are shown in the illustration. Vice verse, Figure 4 shows the inference process, in which, only the documents and community assignments are known.

Symbol	Descriptions
D	The number of entity documents
N_d	The length of the d^{th} entity document
C	The number of communities
K	The number of topics (in case of NER, the number of entity types)
$\beta_{1:C,1:K}$	The topics/types in which each $\beta_{c,k}$ is a distribution over the vocabulary of the c^{th} community
θ_d	The topic/type proportions for the d^{th} entity document
$\theta_{d,k}$	The topic/type proportion for topic k in entity document d
z_d	The topic/type assignments for the d^{th} entity document
$z_{d,n}$	The topic/type assignment for the n^{th} word in entity document d
χ_d	The community proportions for the d^{th} entity document
$\chi_{d,c}$	The community proportion for community c in entity document d
c_d	The community assignments for the d^{th} entity document
$c_{d,n}$	The community assignment for the n^{th} word in entity document d
w_d	The observed words for entity document d
$w_{d,n}$	The n^{th} word in entity document d, which is an element from the fixed vocabulary
Λ_d	The list of binary topic/type presence/absence indicators $\Lambda_d = (l_1, \ldots, l_K)$ in entity document d, each $l_k \in \{0, 1\}$

Table 3: Notation used in our model, based on the basic LDA model [3].

According to the generative process and the notation in Table 3, the dependency between these factors can be shown by Figure 3. Specifically, for entity document d, we choose a multinomial community distribution χ_d under the Dirichlet prior with parameter ζ, and a multinomial entity type (topic) distribution θ_d under the Dirichlet prior with parameter α, based on the entity document's possible entity types Λ_d which are constrained by entity dictionaries acquired from Freebase[2]. Then, for each word $w_{d,n}$, we choose a

[2]Freebase (https://www.freebase.com) was an online collection of structured data harvested from sources such as Wikipedia, NNDB, Fashion Model Directory and MusicBrainz, as well as data contributed by its users. Freebase contains (but not limited to)

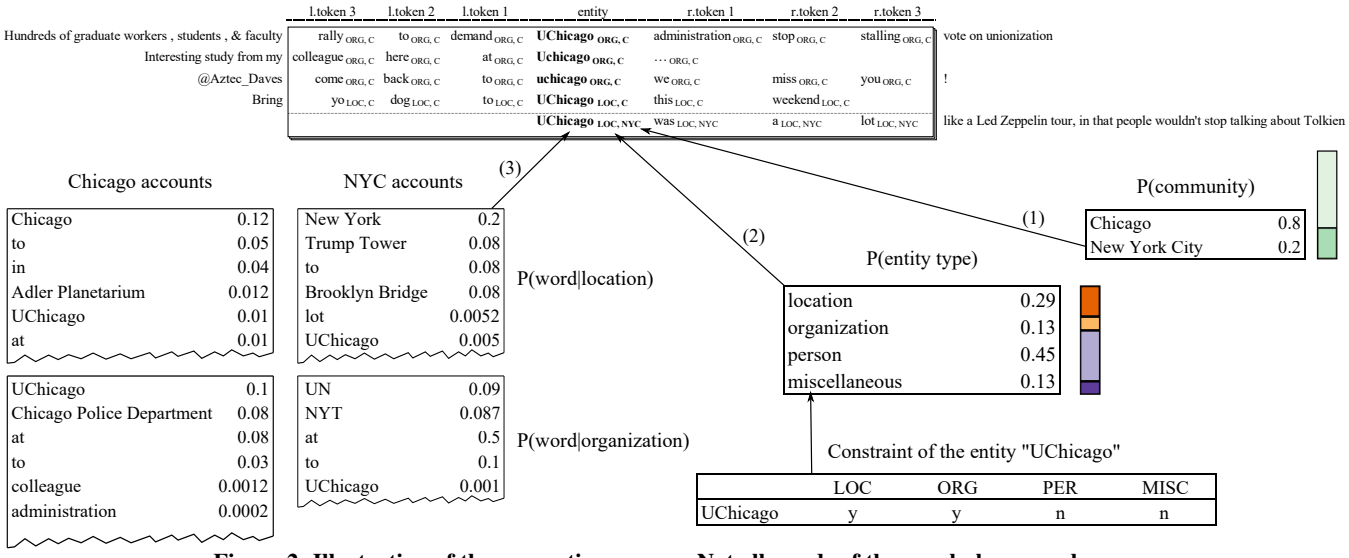

Figure 2: Illustration of the generative process. Not all words of the vocabulary are shown.

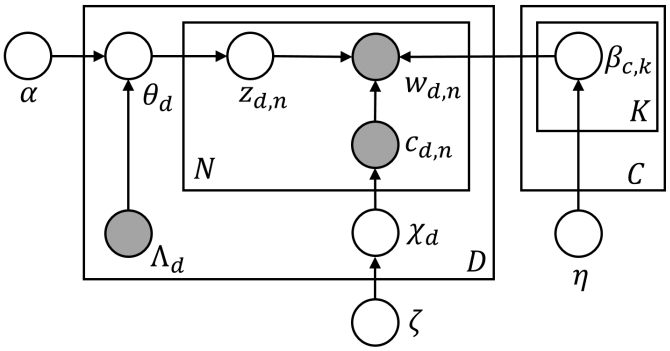

Figure 3: Graphical illustration of Labeled LDA model incorporating community information (C-LLDA), both the labels set Λ and the topic prior α influence the topic mixture θ.

Algorithm 1

for all $c = 1...C$ **do**
  for all $k = 1...K$ **do**
    Generate $\beta_{c,k} \sim \text{DirSym}(\eta)$
for all entity document $d = 1...D$ **do**
  Generate θ_d over $\Lambda_d \sim \text{DirSym}(\alpha_{\Lambda_d})$
  Generate $\chi_d \sim \text{DirSym}(\zeta)$
  for all word position $n = 1...N_d$ **do**
    Generate $z_{d,n} \sim \text{Mult}(\theta_d)$
    Generate $c_{d,n} \sim \text{Mult}(\chi_d)$
    Generate the word $w_{d,n} \sim \text{Mult}(\beta_{c_{d,n}, z_{d,n}})$

community $c_{d,n}$ from multinomial community distribution χ_d, and an entity type (topic) $z_{d,n}$ from multinomial entity type (topic) distribution θ_d. Finally, the word $w_{d,n}$ is generated from a multinomial word distribution $\beta_{c_{d,n}, z_{d,n}}$ under a Dirichlet prior with parameter η. The generative process of C-LLDA is formally described in Algorithm 1.

Assuming all of the words of an entity mention and its context, W (this differs from the bag of words w_d of all mentions), are sampled from the multinomial distribution $\text{Mult}(\beta_{c,z})$ of a single entity type z and a single community c, the inference can be carried out by optimizing the posterior distribution over entity types:

$$P(z|w_i) \propto \prod_{w_i \in W} P(w_i|z : \beta_{c,z}) P(z : \theta)$$

The inferred entity type z is one of the possible types constrained by the dictionaries via Λ_d. In this paper we use 19 types from Freebase which are detailed in Section 3.

We report the performance of a model in which the community assignment of every word is known as user's home city[3]. In comparison to the original LDA model [3], the proposed model maintains a distribution $\beta_{c,z}$ over the vocabulary for each community c, providing a localized distribution for the community and can be used to make better predictions on location names.

We extend the Labeled LDA implementation provided at https://github.com/aritter/twitter_nlp, which applies Gibbs sampling [10] in learning, a resampling strategy [25] in inference, and uses HBC [5] to generate parts of the program.

3 EVALUATION

3.1 Experimental Setup

Dataset We use the dataset provided by Li et al. [14] for our experiments. The dataset contains about 50 millions tweets posted

types and entities belong to these types, e.g., Arnold Schwarzenegger was described as an actor, bodybuilder, and politician. The Freebase API was shut-down on Aug 31 2016, and Google is maintaining its latest dump. Currently the similar service can be found on https://www.wikidata.org.

[3] In another case where the community labels χ are unobserved and unconstrained, a community can be considered as a group of users interested in the same topic, and is inferred by the model. An experiment with the unobserved model yields no positive results on 670,000 tweets from 3,000 London users, suggesting the hardship in detecting fine-grained communities in a city.

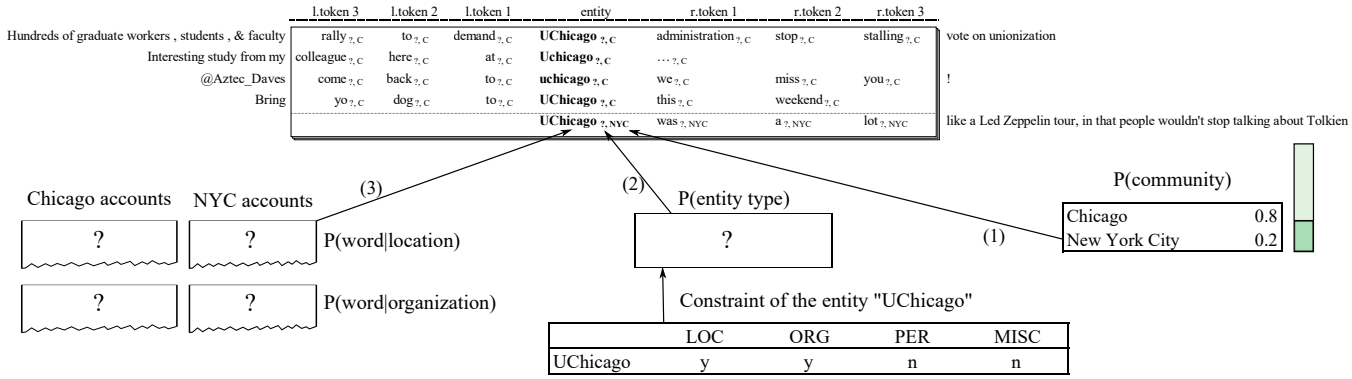

Figure 4: Illustration of the inference process. Only the documents and community information are known.

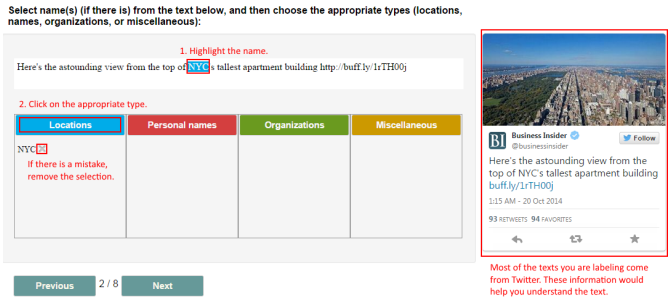

Figure 5: The annotation tool provides additional information for workers.

by 140,000 users in May 2011. These users have locations in their profiles and are selected from 3 million users who have at least 10 friends. We consider a city as a geographically defined community and choose two cities: Chicago and New York City. These two cities have comparable number of users (3,425 and 3,515 respectively in the dataset) and tweets (about 1 millions tweets of each) in this dataset. Retweets beginning with "RT" are excluded to avoid repetition. We also remove the manual labels of entities started with @, which are easy to recognize and likely leads to performance inflation. In the following experiments, we investigate whether spatial community affects named entity classification.

Type	3 votes	2 votes	1 vote
LOC	41.65%	45.63%	12.72%
PER	32.57%	55.46%	11.97%
ORG	15.33%	52.92%	31.75%
MISC	6.76%	51.25%	41.99%

Table 4: Agreement among Mechanical Turk workers on the type of the entities.

Data Annotation[4] We randomly choose 1,000 tweets from each city. For every tweet, we ask 3 different workers on Amazon Mechanical Turk to highlight the entities and select a type from 4 types *PERSON*, *ORGANIZATION*, *LOCATION*, and *MISCELLANEOUS*. An intuitive tool is built to aid the workers in this specific task (Figure 5), providing the context of the tweets to the workers. The final assignment of type for every token is carried out by majority voting.

[4]The annotated data is published on https://github.com/phucng/lens17.

Table 4 shows how the annotators agree with each other on the labels of the entities as a whole. Figure 6 shows the distribution of the entity types in these tweets. Most of the tweets contain entities of a kind, suggesting the popularity of entities on the microblog.

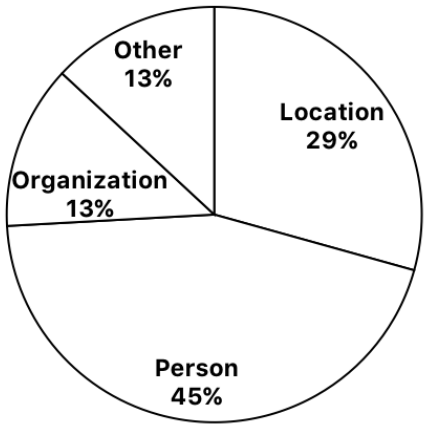

Figure 6: Distribution over 4 entity types of 2,145 annotated entities from 2,000 tweets.

4-fold cross-validation is carried out and the average performance is reported. After classification, a mapping from 19 Freebase types to 4 standard types (Table 5) to obtain the final assignments.

Baseline We compare the performance of the proposed community-based classifier (**C-LLDA**) with Ritter et al.'s **T-NER** [20], using the same training and test data.

3.2 Effects of Community Information on Location Recognition

Table 6 shows the comparison of the community-based classifier and the baseline. Table 7 and Table 8 demonstrate the consistent improvements of the proposed model in Chicago and New York, respectively. Community information boosts the performance in term of F1 measure by 5%. Location names tend to have locality than other types. For example, "GAB" is used by a community to refer to "General Academic Building" and has totally different meanings outside that community. As a result, location entities may benefit more from community information. In the same community, some

Freebase type	Entity type
person_combined	PER
film.film	MISC
company_combined	ORG
tv.tv_program	MISC
government.politician	PER
sports.sports_team	ORG
sports.sports_league	MISC
cvg.computer_videogame	MISC
book.newspaper	ORG
time.holiday	MISC
award.award	MISC
tv.tv_network	ORG
government.government_agency	ORG
sports.pro_athlete	PER
automotive.make	ORG
music.musical_group	ORG
location	LOC
facility	LOC
product	MISC

Table 5: The mapping from Freebase types to standard entity types.

Type	P	R	F_1
LOC_{T-NER}	68.40%	25.62%	37.17%
LOC_{C-LLDA}	**71.21%**	**30.24%**	**42.20%**

Table 6: Comparison of C-LLDA and the baseline (T-NER) in *named entity classification* on location. The performance is shown in term of precision, recall, and F_1 score.

tweets may have better context to infer the location types and other tweets from the same community difficult to infer can benefit from them.

Type	P	R	F_1
LOC_{T-NER}	72.88%	26.84%	39.10%
LOC_{C-LLDA}	**73.19%**	**32.10%**	**44.34%**

Table 7: Performance comparison on tweets from Chicago-based users in *named entity classification* on location.

Type	P	R	F_1
LOC_{T-NER}	63.72%	24.35%	35.15%
LOC_{C-LLDA}	**68.91%**	**28.07%**	**39.71%**

Table 8: Performance comparison on tweets from New York-based users in *named entity classification* on location.

3.3 Error Analysis

Table 9 shows the entities identified as locations by C-LLDA from 400 randomly sampled tweets, 200 from each city. Local places such as "Pheasant Run St Charles", "ClickZ", and "The High Line" (displayed in bold texts) are successfully classified while the baseline failed. The names are found from these texts:

(1) Zanies Comedy Night Club - St. Charles @ Pheasant Run St Charles
(2) I'm at ClickZ (120 Broadway, New York, NY, btw pine and cedar, New York)
(3) One of my favorite places in the City (@ The High Line w/ 6 others

On the other hand, the model struggles with names which can be either location or organization/show such as "MoMA", "Writers' Theatre", and "International Auto Show" (in italic). In these false positive examples, there is likely insufficient context information for the model to distinguish location from other types. The underlined names are correctly classified by the baseline, while the other italic ones are misclassified by both models.

Community	Location names
Chicago	Broadway, the Gulf, Illinois, Regal Webster Place 11, Wacker, Hyde Park, Alumni Memorial Union, AMU, Oak Park, 3rd Coast Comics, the Grotto, Woburn, **Pheasant Run St Charles**, Harold Washington Library, Chick - Fil - A, Hawaii, **Syria**, *Google Transparency Engineering*, *Cystic Fibrosis Foundation*, *Kindles*, *PittPatt*, *Goose Island Beer*, *MoMA*, *House Republicans*, *French-Mexican Fusion Fare*, *TARP*
New York	Central Park, **ClickZ**, **"120 Broadway, New York, NY"**, The square, the East Village, Norway, Shelter Island, Spokane, the Mudtruck, **The High Line**, NY, Asheville, West End, Haiti, Glencoe, LA, Shake Shack, "154 E 86th St, Btw Lexington & 3rd Ave, New York", *Red Dog*, *NEXT TO LOVE*, *Ha*, *Bradley*, *International Auto Show*, *Writers' Theatre*, *Immobilier Boulogne Billancourt*, *URI*, *Yale*, *Rapids*, *Spencer Reed Matan Shalev*

Table 9: Entities which C-LLDA classified as locations from 400 randomly sampled tweets.

Type	P	R	F_1
PER_{T-NER}	76.60%	**70.25%**	**72.99%**
PER_{C-LLDA}	**77.19%**	68.88%	72.42%
ORG_{T-NER}	21.96%	**41.57%**	28.45%
ORG_{C-LLDA}	**22.68%**	40.46%	**28.73%**
$MISC_{T-NER}$	23.26%	**40.41%**	29.22%
$MISC_{C-LLDA}$	**25.75%**	43.07%	**32.03%**
ALL_{T-NER}	42.02%	45.17%	40.10%
ALL_{C-LLDA}	**44.24%**	**45.70%**	**41.75%**

Table 10: Comparison of C-LLDA and the baseline (T-NER) in *named entity classification* on the other types.

3.4 Effects of Community Information on Other Entity Types

Table 10 shows the performance of C-LLDA and the baseline on classifying organization names, personal names, and miscellanea, in which community information raises the performance in most cases. The general improvement is subtle but noticeable in both cities (Table 11 and Table 12). Dissimilar from location names, a person is more likely to be called the same across places, making his name harder to benefit from locality.

Type	P	R	F_1
PER_{T-NER}	66.88%	**68.52%**	67.65%
PER_{C-LLDA}	**67.55%**	67.96%	**67.67%**
ORG_{T-NER}	**24.66%**	**40.91%**	**30.71%**
ORG_{C-LLDA}	23.56%	39.83%	29.51%
$MISC_{T-NER}$	26.27%	**48.04%**	33.49%
$MISC_{C-LLDA}$	**28.11%**	45.32%	**34.43%**
ALL_{T-NER}	41.86%	**46.76%**	40.74%
ALL_{C-LLDA}	**43.37%**	45.81%	**41.65%**

Table 11: Performance comparison on tweets from Chicago-based users in *named entity classification* on the other types.

Type	P	R	F_1
PER_{T-NER}	85.14%	**71.77%**	**77.70%**
PER_{C-LLDA}	**85.80%**	69.71%	76.66%
ORG_{T-NER}	19.51%	**42.17%**	26.41%
ORG_{C-LLDA}	**21.76%**	41.12%	**27.91%**
$MISC_{T-NER}$	19.59%	31.12%	24.02%
$MISC_{C-LLDA}$	**23.36%**	**40.79%**	**29.58%**
ALL_{T-NER}	42.17%	43.57%	39.46%
ALL_{C-LLDA}	**45.12%**	**45.59%**	**41.84%**

Table 12: Performance comparison on tweets from New York-based users in *named entity classification* on the other types.

3.5 Segmentation and Discussion

This paper does not incorporate community information into segmentation and the performance of segmentation stays the same for both. For completeness, however, we report the overall results including segmentation. Table 13 shows the comparison including both segmentation and classification. Because each mis-segmented entity is counted as a mis-classified one, we observe reduced accuracy in comparison to that of classification for both C-LLDA and T-NER. Overall the proposed recognizer still performs better than the baseline.

As in [6], we observe the changes in performance of NER systems from dataset to dataset. T-NER's F1 is less than the performance reported in [20], suggesting its susceptibility to the dataset. This phenomenon would be caused by the veriety in content of the microblog and its ever-changing underlying distribution.

Type	P	R	F_1
LOC_{T-NER}	50.00%	16.53%	24.46%
LOC_{C-LLDA}	**52.28%**	**16.92%**	**25.31%**
PER_{T-NER}	**66.92%**	**49.83%**	**56.84%**
PER_{C-LLDA}	64.75%	47.96%	54.78%
ORG_{T-NER}	**7.62%**	10.63%	8.45%
ORG_{C-LLDA}	7.60%	**13.29%**	**9.29%**
$MISC_{T-NER}$	3.77%	4.12%	3.88%
$MISC_{C-LLDA}$	**5.46%**	**6.27%**	**5.81%**
ALL_{T-NER}	28.77%	19.49%	21.92%
ALL_{C-LLDA}	**29.42%**	**19.98%**	**22.43%**

Table 13: Comparison of C-LLDA and the baseline (T-NER) in *named entity recognition*.

4 RELATED WORK

Named-entity recognition tries to locate and classify named entities in text into pre-defined categories such as the persons, organizations, locations, times, quantities, and monetary values. Location recognition can be part of a general NER model. This paper focus on improving location recognition performance by incorporating collective cues from multiple tweets posted by users in close-by locations while maintaining/improving general NER performance. The most related work is in the area of name entity recognition that includes location recognition and geographical topic model that our new method is based on.

4.1 Location Recognition in Tweets

Being one of the most fundamental problems of information extraction in microblogs, named entity recognition has gained much attention in recent years. Liu et al. [15] propose a named entity recognizer for tweets that follows a two-stage prediction aggregation method [12]. They adopt a k-nearest-neighbor (KNN) classifier to conduct word level classification to leverage global evidence across tweets. These pre-labeling results, together with gazetteer-related features, are fed to a conditional random field (CRF) model [13] for the fine-grained labeling task. The KNN classifier and the CRF model work under a semi-supervised learning framework: 10,000 most recent labeled tweets with high confidence are maintained for retraining. They utilize BILOU encoding [19] for label representation and report a 78.9% F1 in location recognition on their data set. In a later work [16], they describe a graphical model simultaneously conducting named entity recognition and normalization–the task to transform named entities mentioned in tweets to their unambiguous canonical forms. This method boosts their F1 from 78.9% to 82.1% for NER in location.

Ritter et al. [20] build a specialized NLP pipeline for tweets with POS tagging, chunking, and NER. They train a CRF model for segmentation, and a Labeled LDA model for classification using information from 6M unlabeled tweets and Freebase dictionaries. The system (F1 = 0.59) outperforms the Stanford NER [9] (F1 = 0.29) and a co-training implementation (F1 = 0.49) with 3 entity types *PERSON*, *LOCATION*, and *ORGANIZATION*. Their named entity classifier yields a 77% F1 in location and the performance of

NER (including segmentation and classification) in location is not reported.

In the ACL 2015 Workshop on Noisy User-generated Text [2], there are 8 teams participated in a NER shared task. The most common features are part-of-speech (POS), orthographic, gazetteers, Brown clusters, and word embeddings. Many of them use conditional random fields as the model. Team ousia [24] shows the best overall performance in segmenting and categorizing with 56.41% F1, and achieves 66.42% F1 in location. They use entity linking (also known as entity disambiguation) to enhance the NER results. In the shared task in 2016, an additional data set for domain-specific task is proposed. Most participants perform better on this data than on general tweets and the LSTM (Long Short-Term Memory) models are found to be popular this year.

Derczynski et al. [6] describe a Twitter dataset for entity disambiguation, and conduct an extensive analysis of named entity recognition and disambiguation. They show that state-of-the-art NER approaches did not perform robustly on ill-formed, terse, and linguistically "compressed" microblog texts; in which, some Twitter-specific methods reached F1 measures of over 80%, but were still far from the state-of-the-art results achieved on newswire. They also discuss some causes such as poor capitalization, typographic errors, out-of-vocabulary words, lack of training data, and the diversity of entity types which are typical in microblogs. To address these problems, they find language identification, microblog-trained POS tagging, and normalization led to some improvements.

Regarding tweets labeling, Finin et al. [8] investigate the use of Amazon Mechanical Turk and CrowdFlower for collecting named entity annotations for Twitter status updates.

Previous efforts are limited to considering tweets as documents and disregarding community information which would affect the composition of tweets. In contrast, we take the community information into account and observe the effects of community on location mentions.

4.2 Geographical Topic Models and Community Effects on Topic

Eisenstein et al. [7] assume that regions and topics interact to shape lexical frequencies. They then present a generative model which jointly identifies words with regional affinity, geographically-coherent linguistic regions, and the relationship between regional and topic variation.

Yin et al. [26] propose and compare three strategies of modeling geographical topics including location-driven model, text-driven model, and a joint model called LGTA (Latent Geographical Topic Analysis) that combines both location and text information. Evaluation results show that their LGTA model works well for not only finding regions of interests but also providing effective comparisons of different topics across locations.

Hong et al. [11] address the problem of modeling geographical topical patterns on Twitter by introducing a sparse generative model, which utilizes both statistical topic models and sparse coding techniques to uncover different language patterns and common interests. They demonstrate the model's effectiveness on the task of predicting locations of new messages.

Paul et al. [17] present preliminary results on the detection of cultural differences from experiences of tourists and locals perspectives in some countries. They propose a model, which extends over LDA [3] and cross-collection mixture models, and provide analysis of the model on blogs and forums.

None of the aforementioned works deals with the problem of NER, which is the focus of our paper. In this paper, we look for regional variations in the ways people mention named entities, and methods to adopt these features to enhance named entity classification in location.

5 CONCLUSION AND FUTURE WORK

The paper tackles the problem of identifying location names in tweets, as a part of named entity recognition. We propose C-LLDA, an approach for named entity recognition harnessing community information. A dataset has been labeled for evaluation, and the experimental results show improvements over the baseline, with remarkable improvement on location names.

In future, we will experiment with a larger number of cities and extend this community idea into segmentation. We shall also investigate the effects of social relation-based communities instead of geographical ones.

6 ACKNOWLEDGMENTS

We would like to thank Rodney Nielsen and Eduardo Blanco for useful discussions. We thank Longbo Kong for his help with the annotation tool, and thank the workers on Amazon Mechanical Turk. This research was made possible by funding from the NGA University Research Initiatives (NURI). Computational resources were provided by UNT's High Performance Computing Services.

REFERENCES

[1] Einat Amitay, Nadav Har'El, Ron Sivan, and Aya Soffer. 2004. Web-a-where: geotagging web content. In *Proceedings of the 27th ACM SIGIR*. 273–âĂŞ280.
[2] Timothy Baldwin, Young-Bum Kim, Marie Catherine de Marneffe, Alan Ritter, Bo Han, and Wei Xu. 2015. Shared tasks of the 2015 workshop on noisy user-generated text: Twitter lexical normalization and named entity recognition. (2015), 126.
[3] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *The Journal of machine Learning research* 3 (2003), 993–1022.
[4] Peter F Brown, Peter V Desouza, Robert L Mercer, Vincent J Della Pietra, and Jenifer C Lai. 1992. Class-based n-gram models of natural language. *Computational linguistics* 18, 4 (1992), 467–479.
[5] Hal DaumÃľ III. 2008. hbc: Hierarchical bayes compiler. *Pre-release version 0.7, URL http://www.cs.utah.edu/˜hal/HBC* (2008).
[6] Leon Derczynski, Diana Maynard, Giuseppe Rizzo, Marieke van Erp, Genevieve Gorrell, RaphaÃńl Troncy, Johann Petrak, and Kalina Bontcheva. 2015. Analysis of named entity recognition and linking for tweets. 51, 2 (2015), 32–49. https://doi.org/10.1016/j.ipm.2014.10.006
[7] Jacob Eisenstein, Brendan O'Connor, Noah A. Smith, and Eric P. Xing. 2010. A latent variable model for geographic lexical variation. In *Proceedings of the 2010 EMNLP*. ACL, 1277–1287.
[8] Tim Finin, Will Murnane, Anand Karandikar, Nicholas Keller, Justin Martineau, and Mark Dredze. 2010. Annotating named entities in Twitter data with crowdsourcing. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*. ACL, 80–88.
[9] Jenny Rose Finkel, Trond Grenager, and Christopher Manning. 2005. Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43rd Annual Meeting on ACL*. 363–âĂŞ370.
[10] Thomas L. Griffiths and Mark Steyvers. 2004. Finding scientific topics. *Proceedings of the National Academy of Sciences* 101 (2004), 5228–5235. Issue suppl 1.
[11] Liangjie Hong, Amr Ahmed, Siva Gurumurthy, Alexander J. Smola, and Kostas Tsioutsiouliklis. 2012. Discovering Geographical Topics in the Twitter Stream. In

Proceedings of the 21st International Conference on World Wide Web (WWW '12). ACM, New York, NY, USA, 769–778.

[12] Vijay Krishnan and Christopher D. Manning. 2006. An Effective Two-stage Model for Exploiting Non-local Dependencies in Named Entity Recognition. In *Proceedings of the 21st COLING and the 44th Annual Meeting of the ACL*. ACL, 1121–1128.

[13] John Lafferty, Andrew McCallum, and Fernando CN Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. (2001).

[14] Rui Li, Shengjie Wang, Hongbo Deng, Rui Wang, and Kevin Chen-Chuan Chang. 2012. Towards social user profiling: unified and discriminative influence model for inferring home locations. In *KDD*. 1023–1031.

[15] Xiaohua Liu, Shaodian Zhang, Furu Wei, and Ming Zhou. 2011. Recognizing Named Entities in Tweets. 359–âĂŞ367.

[16] Xiaohua Liu, Ming Zhou, Furu Wei, Zhongyang Fu, and Xiangyang Zhou. 2012. Joint inference of named entity recognition and normalization for tweets. In *Proceedings of the 50th Annual Meeting of the ACL: Long Papers-Volume 1*. ACL, 526–535.

[17] Michael Paul and Roxana Girju. 2009. Cross-cultural Analysis of Blogs and Forums with Mixed-collection Topic Models. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 3 - Volume 3 (EMNLP '09)*. Association for Computational Linguistics, Stroudsburg, PA, USA, 1408–1417.

[18] Daniel Ramage, David Hall, Ramesh Nallapati, and Christopher D. Manning. 2009. Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of the 2009 EMNLP: Volume 1-Volume 1*. ACL, 248–256.

[19] Lev Ratinov and Dan Roth. 2009. Design challenges and misconceptions in named entity recognition. In *Proceedings of the Thirteenth CoNLL*. 147–âĂŞ155.

[20] Alan Ritter, Sam Clark, and Oren Etzioni. 2011. Named entity recognition in tweets: an experimental study. In *EMNLP*. ACL, 1524–1534.

[21] Jagan Sankaranarayanan, Hanan Samet, Benjamin E. Teitler, Michael D. Lieberman, and Jon Sperling. 2009. TwitterStand: News in Tweets. In *Proceedings of the 17th ACM SIGSPATIAL (GIS '09)*. ACM, New York, NY, USA, 42–51.

[22] Fei Sha and Fernando Pereira. 2003. Shallow parsing with conditional random fields. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*. Association for Computational Linguistics, 134–141.

[23] David A. Smith and Gregory Crane. 2001. Disambiguating Geographic Names in a Historical Digital Library. In *Proceedings of the 5th ECDL*. Springer-Verlag, 127–136.

[24] Ikuya Yamada, Hideaki Takeda, and Yoshiyasu Takefuji. 2015. Enhancing Named Entity Recognition in Twitter Messages Using Entity Linking. (2015), 136.

[25] Limin Yao, David Mimno, and Andrew McCallum. 2009. Efficient methods for topic model inference on streaming document collections. In *Proceedings of the 15th ACM SIGKDD*. ACM, 937–946.

[26] Zhijun Yin, Liangliang Cao, Jiawei Han, Chengxiang Zhai, and Thomas Huang. 2011. Geographical Topic Discovery and Comparison. In *Proceedings of the 20th International Conference on World Wide Web (WWW '11)*. ACM, New York, NY, USA, 247–256.