Entity Linking on Microblogs with Spatial and Temporal Signals

Yuan Fang *†
University of Illinois at Urbana-Champaign
201 N. Goodwin Avenue
Urbana, IL 61801, USA
fang2@illinois.edu

Ming-Wei Chang
Microsoft Research
1 Microsoft Way
Redmond, WA 98052, USA
minchang@microsoft.com

Abstract

Microblogs present an excellent opportunity for monitoring and analyzing world happenings. Given that words are often ambiguous, entity linking becomes a crucial step towards understanding microblogs. In this paper, we re-examine the problem of entity linking on microblogs. We first observe that spatiotemporal (*i.e.*, spatial and temporal) signals play a key role, but they are not utilized in existing approaches. Thus, we propose a novel entity linking framework that incorporates spatiotemporal signals through a weakly supervised process. Using entity annotations[1] on real-world data, our experiments show that the spatiotemporal model improves F1 by more than 10 points over existing systems. Finally, we present a qualitative study to visualize the effectiveness of our approach.

1 Introduction

Microblogging services provide an immense platform for intelligence gathering, such as market research (Asur and Huberman, 2010), disaster monitoring (Sadilek et al., 2012) and political analysis (Tumasjan et al., 2010). Extracting entities from microblogs is an essential step for many such applications. Suppose that a marketing firm is interested in the sentiment about some product on Twitter. However, any sentiment analysis is potentially misleading if we cannot correctly retrieve the tweets mentioning the target product.

To retrieve tweets for a given entity (*e.g.*, a product or an organization), a straightforward approach is to formulate a keyword query. However, simple keyword matching is largely ineffective, since keywords are often ambiguous. For instance, "spurs" can refer to two distinct sports teams (San Antonio Spurs, which is a basketball team in the US, and Tottenham Hotspur F.C., which is a soccer team in the UK), besides being a non-entity verb or noun. Thus, retrieving tweets via keyword queries inevitably mixes different entities.

Given the ambiguity of keywords, in this paper, we study the task of entity linking (Bunescu and Pasca, 2006) on microblogs. As *input*, we are given a short message (*e.g.*, a tweet) and an entity database (*e.g.*, Wikipedia where each article is an entity). As *output*, we map every surface form (*e.g.*, "spurs") in the message to an entity (*e.g.*, San Antonio Spurs) or to $\emptyset$ (*i.e.*, a non-entity). This task is particularly challenging for microblogs due to their short, noisy and colloquial nature. Fortunately, we also observe two new opportunities, which are often missing in traditional data.

First, microblogs usually embed rich meta-data, most notably *spatiotemporal* (*i.e.*, spatial and temporal) signals. Specifically, all tweets are associated with a timestamp, and many can be mapped to a location. We observe that entity priors often change across time and location. Consider the example tweets in Fig. 1. To understand the meaning of the word "spurs," it is challenging to only rely on the textual features. However, with the help of lo-

*Work done during an internship at Microsoft Research.

†Also affiliated with Agency for Science, Technology and Research, 1 Fusionopolis Way, Singapore 138632.

[1]Can be downloaded at http://research.microsoft.com/en-us/downloads/84ac9d88-c353-4059-97a4-87d129db0464/.

Transactions of the Association for Computational Linguistics, 2 (2014) 259–272. Action Editor: Noah Smith.
Submitted 11/2013; Revised 3/2014; Published 10/2014.

Tweets mentioning Tottenham Hotspur F.C., a soccer team in the UK

UK: Who scored the 2 goals for **spurs** I had to go out or 5 mins and missed it all

UK: Defoe again 3 nil to **spurs** how you doing @USER at Arsenal

Tweets mentioning San Antonio Spurs, a basketball team in the US

US: @USER who cares, nobody wanna see the **spurs** play. Remember they're boring. He's a great coach ...

US: The fine on the **spurs** for 250k is ridiculous! I'm a laker fan; still think this is too much

Figure 1: **Examples illustrating the importance of spatiotemporal signals to entity linking.** The "UK" or "US" tag indicates the location of the posting user. Based on deep semantic understanding of lexical items and additional resources (such as the entire conversation and attached URL), the annotators can label "spurs" in the first two tweets as the soccer team, and in the other two as the basketball team. While most entity linking systems handle merely textual features, the task obviously becomes easier with location information.

cation information, the two teams can be easily distinguished. In particular, based on our labeled data, San Antonio Spurs accounts for 91% of the "spurs" tweets in the US, whereas it only accounts for 8% in the UK. Similar trends also exist across different time periods. Therefore, exploiting spatiotemporal signals is crucial to entity linking.

Second, we can leverage the predictions from the tweets carrying an easier surface form to help link the tweets carrying an ambiguous surface form at a similar time or location. The intuition is that certain surface forms are easier to disambiguate than others. For instance, while only saying "spurs" in a tweet can be quite ambiguous, many other tweets around the same time or location might carry an easier surface form, such as "SA spurs" or "san antonio spurs," which are not ambiguous.

In this work, we focus on the *offline* mining setting, where a corpus has been collected for analysis offline. Developing algorithms for the streaming setting is an important direction of future work. Our contributions are summarized as follows:

- We propose a spatiotemporal framework for entity linking on microblogs. To our best knowledge, this is the first work to model spatiotemporal signals for entity linking.
- We demonstrate the effectiveness of our framework through extensive quantitative experiments. In particular, we improve F1 by more than 10 points over the existing state of the art.
- We point out that entity linking should be evaluated for both information extraction and retrieval needs. In the former, we evaluate the extraction of entities from tweets, while in the latter we evaluate the retrieval of tweets for a query entity. A qualitative study is also presented for the latter.

2 Related Work

Earlier research on entity linking (Bunescu and Pasca, 2006; Cucerzan, 2007; Milne and Witten, 2008) has been focused on well-written documents such as news and encyclopedia articles. The TAC KBP track (Ji et al., 2010; Ji and Grishman, 2011) also includes an entity linking task with a slightly different setting—to link a given mention based on a background document. These efforts exploit the statistical power aggregated by a semantic knowledge bases, most notably Wikipedia. Various features in the target document are also leveraged (Han et al., 2011; Kulkarni et al., 2009; Hoffart et al., 2011; Shen et al., 2012) to assess both local compatibility and global coherence.

These techniques have also been adapted and tailored to short texts including tweets, for the problem of entity linking (Ferragina and Scaiella, 2010; Meij et al., 2012; Guo et al., 2013) as well as the related problem of named entity recognition (NER) (Ritter et al., 2011; Li et al., 2012a). However, given the short and noisy nature of microblogs, these approaches, which largely depend on textual features, often result in unsatisfactory performance. Fortunately, additional non-textual meta-data in microblogs can often help. A recent study (Shen et al., 2013) improves entity linking by utilizing user account information, based on the intuition that all tweets posted by the same user share an underlying topic distribution.

Inspired by the use of non-textual features, we explore the spatiotemporal signals associated with tweets. Although the spatiotemporal aspect of social media has been studied in many papers, including temporal cycle tracking (Leskovec et al., 2009), spatial object matching (Dalvi et al., 2012), min-

ing emerging topics (Mei et al., 2006; Cataldi et al., 2010; Yin et al., 2011), event monitoring (Sakaki et al., 2010; Xu et al., 2012), and identifying geographical linguistic variations (Eisenstein et al., 2010; Wing and Baldridge, 2011), none of them addresses the problem of entity linking. In this paper, we propose a novel spatiotemporal framework for entity linking, building upon some of the previously observed patterns in social media.

While we believe that entity linking is the first step towards intelligence gathering, many existing studies filter or cluster tweets based on merely keywords. On the one hand, manual selection of keywords (Sakaki et al., 2010; Tumasjan et al., 2010) requires significant labor, and thus is not scalable to the vast number of entities. On the other hand, automatic approaches (Li et al., 2013) only identify coarse-grained topics (*e.g.*, crime or sports), falling short of recognizing specific entities.

Lastly, there is a line of research on record extraction from social media (Benson et al., 2011; Ritter et al., 2012). Although the problem is different from entity linking, they present an interesting insight into social media. They observe that the same record is often referenced by multiple messages, and exploit this redundancy to help with extraction. The redundant nature of social media can be potentially leveraged to improve entity linking as well.

3 Spatiotemporal Entity Linking

Our spatiotemporal framework for entity linking requires an input tweet m, as well as its associated timestamp t and location l. For each tweet m, the goal is to predict an output set $\{e_1, e_2, \ldots\}$ of entities that are mentioned in m.

3.1 Background

To build an entity linking system, we need both a *database* and a *lexicon*.

A database is a set of entities that a tweet can link to. Following earlier work (Sil and Yates, 2013), our database consists of the intersection of Freebase and Wikipedia. We then select the entities belonging to the *core* types (Guo et al., 2013)[2] based on the Freebase type information. In total, there are 2.7 million entities in our database.

A lexicon is a dictionary that maps a candidate anchor a (*i.e.*, a surface form) to its possible entity set $\mathcal{E}(a)$. Similar to existing studies (Cucerzan, 2007; Guo et al., 2013), our lexicon comprises information from disambiguation pages, redirect pages and anchor texts in Wikipedia[3]. To handle the case where a is not an entity, we add a special symbol $\emptyset$ to every $\mathcal{E}(a)$. We have 6 million anchors in total.

Given a tweet, the system generates a set of candidate anchors based on the lexicon. Specifically, each tweet is tokenized into individual words, keeping each punctuation as a separate token. To identify if any sequence of tokens matches an anchor in the lexicon, we use exact matching, but allow for case-insensitivity. Furthermore, we employ an NER system trained using structural perceptron (Collins, 2002) to filter the anchors of type *person*, *loc* or *org*, such that only the anchors recognized by the NER are retained.

3.2 Incorporating Spatiotemporal Signals

Consider a tweet m with timestamp t and location l. Given anchor a in tweet m, our system links a to the candidate entity e^* as follows:

$$\begin{aligned} e^* &= \arg\max_{e \in \mathcal{E}(a)} P(e|m, a, t, l) \\ &= \arg\max_{e \in \mathcal{E}(a)} P(e, m, a, t, l) \end{aligned} \quad (1)$$

We further adopt the following conditional independence assumption in our model.

ASSUMPTION: Given an entity e, *how* e is expressed (m, a), and *when* or *where* e is published (t, l), are conditionally independent. In other words, we have $P(m, a|t, l, e) = P(m, a|e)$. ■

Intuitively, the expression (m, a) of a given entity e is stable across most times and locations (t, l), which could be attributed to the imitation nature (Leskovec et al., 2009) of social media. That is, as stories of e propagate on the social media, users over different t and l often imitate each other (disregarding cross-lingual scenarios). While there could be a "burn-in" period for recent events of e, the distribution of m and a would eventually "stabilize" over most t and l.

[2] Their core types include *person*, *organization* (*org*), *location* (*loc*), *book*, *tvshow* and *movie*. We deal with two additional types, namely, *event* and *product*.

[3] We use a snapshot of Wikipedia taken on 04/03/2013.

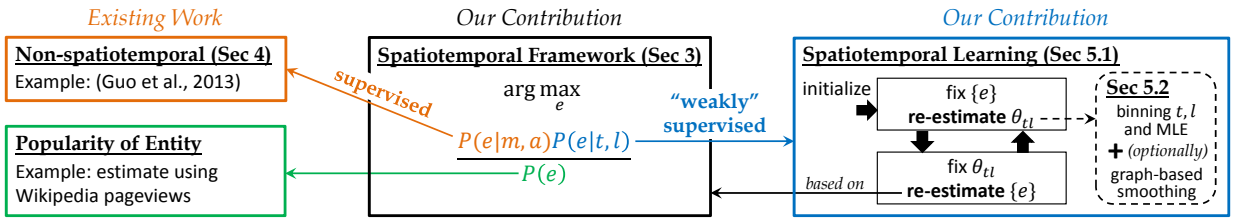


Figure 2: Overall framework of spatiotemporal entity linking.

Subsequently, we decompose the model below, to enable the reuse of existing non-spatiotemporal models for $P(e|m,a)$, as we shall see later.

$$\begin{aligned} & P(e,m,a,t,l) \\ &= P(m,a|t,l,e)P(t,l,e) \\ &= P(m,a|e)P(t,l,e) \\ &= P(e|m,a)P(m,a)P(e|t,l)P(t,l)/P(e) \end{aligned} \quad (2)$$

Note that in (2), we choose to rewrite $P(t,l,e)$ as $P(e|t,l)P(t,l)$ instead of $P(t,l|e)P(e)$. The choice is due to computational issues. In the latter, we need to estimate one distribution for every e; in the former, we need to estimate one distribution for every t,l. There are 2.7 million entities in our database, but with appropriate quantization the number of t,l "bins" are only on the order of thousands.

Thus, the prediction model (1) is equivalent to

$$e^* = \arg\max_{e \in \mathcal{E}(a)} P(e|m,a)P(e|t,l)/P(e). \quad (3)$$

Intuitively, if e at some t,l is more popular than usual, *i.e.*, $P(e|t,l) > P(e)$, we should promote e for a tweet at t,l; otherwise we should demote e.

The overall framework is summarized in Fig. 2, which boils down to the three factors in (3), namely, $P(e|m,a)$, $P(e|t,l)$ and $P(e)$. In Sect. 4, we will discuss briefly how $P(e|m,a)$ is modeled based on previous work. In Sect. 5, we will estimate spatiotemporal signals in the form of $P(e|t,l)$, which is jointly optimized with entity assignments. Lastly, we discuss $P(e)$ here.

$P(e)$ measures the popularity of e, and is estimated by making it proportional to the Wikipedia pageviews of e. While this is a fair proxy for a general social media corpus, we acknowledge that more advanced method is required for a specialized corpus (*e.g.*, for users of a sub-community). Note that $P(e)$ is not estimated based on the joint optimization technique for $P(e|t,l)$ in Sect. 5. The reason is that our tweets used in the experiments only cover a one-month period, which does not necessarily reflect the general popularity of the entities.

4 End-to-End Entity Linking

The goal of this section is to describe our base system that does not consider spatiotemporal signals, *i.e.*, to model $P(e|m,a)$. Specifically, we adopt an end-to-end entity linking system (E2E), which is designed to jointly detect mentions and disambiguate entities. E2E is a supervised method largely based on a previous study (Guo et al., 2013).

For efficiency, we only adopt the first order model. Therefore, the prediction function can be decomposed for each anchor a independently,

$$e^* = \arg\max_{e \in \mathcal{E}(a)} \mathbf{w}^T \Phi(m,a,e). \quad (4)$$

where $\mathbf{w}$ is a linear model trained using structural SVM, and Φ is a feature function over message m, anchor a, and candidate output e.

We use all the basic features and the cohesiveness score feature (Guo et al., 2013)[4]. Additional features are also included. First, for each mention and candidate entity pair, we add a feature to capture the number of highly correlated candidates carried by other mentions in the same tweet with respect to the current candidate. Second, we include a binary feature that will be active if the type of the mention (when recognized by our NER system) and the type of the candidate entity (according to the Freebase type information) agree with each other.

Probability conversion. It is crucial to have a well calibrated probability distribution for the predictions. In order to convert the output of the structural SVM model, we adapt an existing approach

[4]Described in Table 4 and Sect. 4.3 of the reference.

(Platt, 2000) to our case. We define

$$P(e|m,a)=\frac{\exp(b_1+b_2\mathbf{w}^T\Phi(m,a,e))}{\sum_{e'\in\mathcal{E}(a)}\exp(b_1+b_2\mathbf{w}^T\Phi(m,a,e'))},$$

where b_1 and b_2 are the calibration parameters that will be tuned using labeled data.

Given a labeled development set, let $G(e|m,a) = 1$ if and only if anchor a in tweet m is labeled to link to entity e, and let $G(\emptyset|m,a) = 1$ if and only if a in m is not labeled to link to any entity. Note that $\sum_{e\in\mathcal{E}(a)} G(e|m,a) = 1$. Thus, G represents the ground-truth distribution, and we want $P(e|m,a)$ to be as close to $G(e|m,a)$ as possible. To this end, we optimize b_1 and b_2 by minimizing the cross entropy between G and P:

$$\min_{b_1,b_2} \quad -\sum_{m,a} G(e|m,a)\log P(e|m,a).$$

Alternative base system. We also consider *Link-Probability* (LP) as an additional base system. As pointed out earlier (Guo et al., 2013), mention detection is an important step for end-to-end entity linking, and its design is crucial to the ultimate performance. Hence, to detect candidate anchors, LP uses the same design of database and lexicon discussed in this paper and elsewhere (Cucerzan, 2007; Guo et al., 2013), which is believed to be effective. Given a potential anchor a in message m, $P(e|m,a)$ is simply modeled as $P(e|a)$, which can be estimated from Wikipedia anchor statistics. In fact, anchor statistics constitute one of the most useful features in more sophisticated systems (Shen et al., 2012; Guo et al., 2013). Given its robust mention detection mechanism and the utility of anchor statistics, the simple LP turns out surprisingly well.

5 Estimating Spatiotemporal Signals

There are two critical challenges for successfully estimating the spatiotemporal signals in the form of $P(e|t,l)$. First, it is impractical to collect sufficient labeled data to directly estimate it. Second, we need to properly handle the continuous space of the spatiotemporal signals.

In the following, we detail the overall model for learning spatiotemporal signals in a weakly supervised fashion (Sect. 5.1), and then discuss two ways of handling continuous signals (Sect. 5.2).

5.1 Spatiotemporal Learning Model

We model the spatiotemporal signals by a generative model $P(e|t,l) \sim \text{Multi}(\theta_{tl})$, where θ_{tl} is the parameter for the multinomial distribution over all entities at t,l. Since there is no ground truth of the entity assignment, our model will jointly optimize the entity assignment and θ_{tl}. Based on (3), we use the following objective function $\Omega(\{e\},\theta_{tl})$:

$$\sum_{m,a}(\log P(e|m,a) + \log P(e|\theta_{tl}) - \log P(e)),$$

where m is an unlabeled message at time t and location l, a is an anchor in m, and $\{e\}$ is a set of entity assignments for the set of m.

We use a block-coordinate ascent method to find the best θ_{tl} and $\{e\}$ iteratively. For each iteration, the following two steps will be executed.

- Fix θ_{tl}. Find the entity assignments $\{e\}$ that maximizes Ω. Note that if θ_{tl} is fixed, the most likely assignment can be found using

  $$\arg\max_e(\log P(e|m,a)+\log P(e|\theta_{tl})-\log P(e)).$$

  In fact, this equation is the same as (3), where $P(e|m,a)$ can be estimated by a supervised base system (*e.g.*, E2E or LP, see Sect. 4).

- Fix $\{e\}$. Re-estimate θ_{tl} by maximizing Ω. Once the assignment of entities has been generated by the previous step, we can re-estimate it by maximizing the objective function with

  $$\arg\max_{\theta_{tl}} \sum_{m,a} \log P(e|\theta_{tl}).$$

  In other words, we are looking for the maximum likelihood estimate (MLE) of θ_{tl}, given the (previously inferred) entity assignments $\{e\}$. Since θ_{tl} is multinomial, its MLE can be computed as the relative frequency of each e, that is,

  $$\theta_{tl}(e) = \frac{\#\text{ tweets containing } e \text{ at } t,l}{\sum_{e'} \#\text{ tweets containing } e' \text{ at } t,l}.$$

  Of course, t,l are continuous, so direct counting is infeasible. Instead, we resort to discrete bins, which will be treated in Sect. 5.2.

This process will be executed repeatedly, and can be considered as a variant of the Hard EM algorithm. In practice, we run it for up to five iterations, as Hard EM often converges fast.

We call this learning process a "weakly supervised" model. For some time t and location l, the objective function $\Omega(\{e\}, \theta_{tl})$ sums over all messages at t, l. These messages themselves are unlabeled, but some supervision is provided by the base system indirectly. In other words, even though we do not know the ground truth entity assignments in the messages, we can leverage the predictions from the base system to update the entity assignments.

5.2 Handling Continuous Time and Location

Given the continuous space of time t and location l, we propose two methods to estimate θ_{tl}. While it could be cast as an instance of the well-studied density estimation problem (Vapnik and Mukherjee, 1999; Scott, 2009), we resort to a simple binning method, which has also been applied to other spatiotemporal contexts (Wing and Baldridge, 2011; Xu et al., 2012). The binning approach can already demonstrate the significance of spatiotemporal signals in entity linking.

Binning. Our first approach segments the continuous space into discrete bins. Time is divided into a set of equal intervals Δ_{T} (*e.g.*, every one hour), and location is divided into a set of equal squares Δ_{L} (*e.g.*, each 100×100 sqkm area[5]). Let $\Delta = \Delta_{\mathrm{T}} \times \Delta_{\mathrm{L}}$ denote the set of bins over time and location. We further index the bins by $\mathcal{I} = \{1, \ldots, n\}$ where $n = |\Delta|$, and refer to each bin δ_i through its index $i \in \mathcal{I}$. Correspondingly, we denote the multinomial parameter at bin δ_i by θ_i, which can be estimated using maximum likelihood (*i.e.*, relative frequency of entities in δ_i).

It can be shown that with sufficiently small bins, θ_{tl} can be approximated accurately by θ_i when $t, l \in \delta_i$. In practice, if the bins are too small, there are often inadequate tweets in a bin, and thus θ_i cannot be well estimated. On the contrary, if the bins are too large, θ_i cannot reliably model θ_{tl}. In this paper, we tune the bin size using development data.

Graph-based smoothing. Our algorithm contains two steps: first, we estimate $\hat{\theta}_i$ for each bin δ_i as in the above binning approach; next, we smooth the initial estimate to obtain the final estimate θ_i using graph-based regularization.

[5]The location grid is actually defined in longitude and latitude. Here we show the equivalent distance on Earth's surface.

As a common insight in graph-based learning (Zhu et al., 2003; Fang et al., 2012; Fang et al., 2014), if δ_i and δ_j are close to each other, θ_i and θ_j should be similar. Moreover, θ_i should not deviate too much from the initial estimate $\hat{\theta}_i$. The intuition can be captured by the following optimization:

$$\min_{(\theta_1,\ldots,\theta_n)} \quad \tfrac{1}{2}(1-\epsilon)\textstyle\sum_{i,j\in\mathcal{I}} W_{ij}\|\theta_i - \theta_j\|^2 + \epsilon \textstyle\sum_{i\in\mathcal{I}} D_{ii}\|\theta_i - \hat{\theta}_i\|^2, \tag{5}$$

where $\epsilon \in (0, 1)$ is a regularization parameter, W is an affinity matrix such that W_{ij} measures the "closeness" of δ_i and δ_j, and D is a diagonal matrix such that $D_{ii} = \sum_{j\in\mathcal{I}} W_{ij}$. In particular, we design the affinity matrix as follows:

$$W_{ij} = \begin{cases} (d_0 + d(\delta_i, \delta_j))^{\gamma} & i \neq j \\ 0 & i = j, \end{cases} \tag{6}$$

where $d_0 > 0$ and $\gamma < 0$ are parameters, and $d(\cdot)$ is a symmetric distance function for bins. Given that $\gamma < 0$, W_{ij} follows a polynomial decay when δ_i and δ_j become farther apart, as previously suggested (Dalvi et al., 2012).

It can be shown that the optimization problem (5) is equivalent to finding θ_i such that $\forall i \in \mathcal{I}$,

$$\theta_i = (1-\epsilon)\textstyle\sum_{j\in\mathcal{I}} \frac{W_{ij}}{D_{ii}}\theta_j + \epsilon\hat{\theta}_i. \tag{7}$$

Interestingly, we can consider (7) as a generalization of previous observations on social media. It is proposed that the temporal model on social media needs to account for two factors: *imitation* and *recency* (Leskovec et al., 2009). For imitation, users often imitate one another, so that a past story can be picked up and propagated by other users by writing new articles about the same story. For *recency*, more recent stories are more likely to be imitated. We generalize their idea to both temporal and spatial signals, and extend it to entity linking. Specifically, (7) can be interpreted as a result of imitation: tweets mentioning entity e in δ_i are imitated from those in δ_j. In addition, a closer δ_j has a higher chance to be imitated, which captures recency.

To solve (7), let $Q = [\theta_1, \ldots, \theta_n]^T$ and $\hat{Q} = [\hat{\theta}_1, \ldots, \hat{\theta}_n]^T$, treating θ_i and $\hat{\theta}_i$ as column vectors. Through the following iterative updates (Fang et al., 2013), $Q^{(t)}$ converges to Q as $t \to \infty$, starting from

an arbitrary $Q^{(0)}$:

$$Q^{(t+1)} = (1-\epsilon)(D^{-1}W)Q^{(t)} + \epsilon\hat{Q}. \quad (8)$$

The time complexity is $O(tnk)$, where k is the number of neighboring bins on the graph. Such cost is reasonable, given that the update generally converges for $t < 50$, and k is a constant if we only consider k nearest neighbors on the graph.

Joint vs. separate modeling. Finally, while modeling time and location *jointly* is more expressive than modeling each signal separately, the joint approach also creates much more bins (and hence more multinomial parameters), making data scarcity a severe issue. In particular, jointly we have $|\Delta_T| \times |\Delta_L|$ bins, but separately we only have $|\Delta_T| + |\Delta_L|$ bins. By assuming the conditional independence of t and l given e, we can rewrite (3) to model time and location separately:

$$e^* = \arg\max_{e \in \mathcal{E}(a)} P(e|m,a)P(e|t)P(e|l)/P(e)^2 \quad (9)$$

We refer to the joint model (3) as "T×L", and the separate model as "T+L" (9). For "T+L", graph-based smoothing can be carried out separately for the time bins and the location bins. The two models will be compared in our experiments.

6 Experiments

In this section, we assess the effectiveness of our approach through quantitative experiments.

6.1 Settings

Dataset. Our experiments are conducted on Twitter data. We collect all English tweets from verified accounts in December 2012, excluding retweets.[6] These tweets amount to 7 million. As we assume an offline analysis scenario instead of an online streaming scenario, these tweets are fetched and stored locally in advance.

We further select tweets with both spatial and temporal information. For time, we take the posting time of each tweet. Locations are harder to obtain, given that fewer than 5% of our tweets are geotagged. For tweets without geo-coordinates, we use the location in the user profile and map it to coordinates based on a lookup table containing major cities in the US[7]. Such mapping exists for about 25% of the tweets. For the remaining tweets, their user profiles are either uninformative (*e.g.*, "home") or report a non-US location. In the end, 1.8 million tweets remain in our dataset. Note that some studies (Dalvi et al., 2012; Li et al., 2012b) enable the inference of missing locations based on user generated content or user network. We do not apply these methods, which are beyond our focus.

Lastly, we collect the Wikipedia pageviews in the year 2012 in order to estimate $P(e)$ in (3).

Development set. We randomly sample 250 tweets as the development set and label their *core* entities (see Sect. 3.1). There are two human annotators. Each annotator labels half of the tweets, which are then counter-checked by the other to reach an agreement. To label a tweet, the annotators are given all available information to understand its content, *e.g.*, the attached URL and the conversation.

On the other hand, the base system E2E is trained on an independent set of 1000 tweets randomly sampled from the year 2010.

Algorithms. For the binning approach, we tune the bin size on the development set, ranging from 10 minutes to 1 day for time, and from 10×10 sqkm to 1500×1500 sqkm for location. Although the optimal bin size may be entity specific, we use the same size optimized over all entities for scalability considerations. That said, if we are only interested in a small set of target entities, it is possible to optimize the bin size for each entity, in order to attain better performance. Unless otherwise stated, in the rest of the paper, binning is the default method to estimate spatiotemporal signals.

For graph-based smoothing, we use the finest bin for both time and location, and tune its parameters $\epsilon \in \{.01, .05, .1, .5\}$, $d_0 = \{1, 10\}$ and $\gamma \in \{\text{-}.5, \text{-}1, \text{-}2, \text{-}4, \text{-}8\}$ on the development set. To construct the graph, we also need to define the distance function $d(\cdot)$ in (6). For time bins, we use the

[6] The identities of *verified accounts* are validated by Twitter, and thus their tweets contain little spam. Moreover, we ignore retweets here as their spatiotemporal behavior might significantly differ from that of the original tweets.

[7] The lookup table is compiled from geonames.org. We only consider city and state names in the US, discarding all others. As city names may be ambiguous, we apply conservative matching using city-state pair. We also account for popular abbreviations (*e.g.*, IL for Illinois and NYC for New York City).

elapsed time between the middle timestamps of two bins. For location bins, we use the distance between the geographical centers of two bins on Earth's surface. For joint time-location bins, we take the average of time (min) and location (km) distances.

Finally, in both of the above, we use the inferred entity assignments $\{e\}$ to estimate θ_{tl} (see Sect 5.1). Since the inferred assignments can be quite noisy, we only count confident ones such that their inferred probability is at least 0.5.

6.2 Evaluation Methodology

We adopt two different evaluation policies, which are driven by information *extraction* (IE) and *retrieval* (IR) needs, respectively.

IE-driven evaluation. The IE-driven evaluation is similar to the standard evaluation for an end-to-end entity linking system, which evaluates the entities "extracted" by linking. We randomly sample 250 tweets as the test set, which are labeled in the same manner as the development set. For evaluation, we compute the F1 score over the test set, as defined earlier (Guo et al., 2013).

According to the labels, 40% of the tweets in the test set contain at least one entity. A total of 179 entity instances are identified, or an average of 0.72 per tweet. These entities belong to different types, including *person* (24%), *org* (36%), *loc* (16%), *event* (9%), and others (15%).

IR-driven evaluation. One key application of entity linking is to enable intelligence gathering for a *query entity*, where the first step is to "retrieve" the tweets mentioning the query entity. As listed in Table 1, we sample ten query entities of different types, where each entity is known to be influenced by spatiotemporal signals, and has one or more ambiguous anchors. For instance, Hillary Rodham Clinton may be mentioned by "clinton" only, which could be mistakenly linked to Bill Clinton.

Type	Query entity	
person	Hillary Rodham Clinton*	Catherine, Duchess of Cambridge*
org	San Antonia Spurs*	Big Bang (South Korean band)*
loc	Washington (state)*	Newtown, Connecticut
event	Hanukkah*	Winter solstice
movie	Les Misérables (2012 film)	Django unchained (2012 film)

Table 1: Query entities for IR-driven evaluation.

For each query entity, we randomly sample 100 tweets as the test set (totaling 1000 tweets), such that each tweet contains an ambiguous anchor of the entity (through a lookup from our lexicon). Each tweet is labeled to indicate whether it mentions the query entity or not. About 37% of the sampled tweets did not actually mention the query entity since the anchors are ambiguous. For evaluation, we test if the system can correctly identify the presence or absence of the query entity in every tweet. In particular, we compute the F1 score over the test set.

Note that this task is harder than it seems. Besides ambiguous anchors, many entities are not dominant for their anchor. For instance, for the anchor "big bang", Big Bang (South Korean band) is less popular than The Big Bang Theory (a TV show). In fact, six of the entities are not dominant, marked by $\star$ in Table 1. Thus, merely linking to the most popular one (*i.e.*, the base system LP) is not a good strategy.

Significance test. To establish the statistical significance of our results, we randomly divide the test set into 10 splits of equal number of tweets, and compute the F1 score on each split for each algorithm. Two-tail paired t-test is then applied to determine if the F1 scores of two algorithms over the 10 splits are significantly different.

6.3 Results and Discussion

We present the empirical findings for the following research questions.

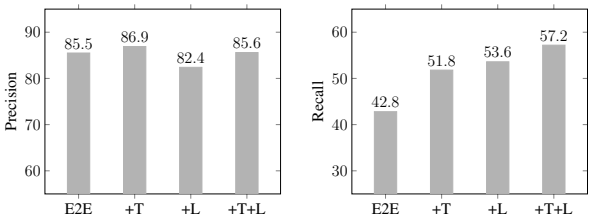
Q1: How do our base systems perform?
Q2: Are spatiotemporal signals indeed useful?
Q3: Does the graph-based smoothing help?
Q4: What causes the errors? How to recover them?

Base system comparison (Q1). To show that our base systems, in particular E2E, already outperform other systems, we compare with Wikiminer (Milne and Witten, 2008) and Illinois (Ratinov et al., 2011) systems.[8] As existing systems are more geared for the IE scenario, we report in Table 2 the IE-drive F1 on the test set.

[8]We use the authors' implementations. AIDA (Hoffart et al., 2011) is not compared, as it mostly links to *person*, *org* and *loc* only. TAGME (Ferragina and Scaiella, 2010) and Cucerzan (Cucerzan, 2007) were already compared in an earlier paper (Guo et al., 2013) which E2E is largely based on. To be fair, we discard non-core entities linked by Wikiminer or Illinois.

	E2E		LP	
	IE	IR	IE	IR
base	57.0 –	58.4 –	48.3 –	48.5 –
+T	64.9 ***	71.4 ***	**52.4** *	59.7 ***
+L	65.0 **	76.1 ***	50.3 *	**61.8** ***
+T+L	**68.6** ***	**79.0** ***	49.0	53.3 ***
+T×L	66.2 ***	74.1 ***	50.6	61.2 ***

(a) F1 scores



(b) Precision (c) Recall

Figure 3: **Effect of using spatiotemporal signals.** (a) F1 scores. ***, **, *: Significantly different from the base system at .01, .05, .1 levels. (b, c) IE-driven evaluation of precision and recall, using E2E as the base system.

The results clearly show that E2E performs better than other base systems. Also note that LP obtains similar or better F1 than Illinois or Wikiminer, although its precision is lower.

	Precision	Recall	F1	Significance
Wikiminer	78.9	24.7	37.6	***
Illinois	77.3	34.9	48.1	**
LP	49.7	**47.0**	48.3	**
E2E	**85.5**	42.8	**57.0**	–

Table 2: IE-driven comparison of base systems. ***, **, *: Significantly different from E2E at .01, .05, .1 levels.

Spatiotemporal signals (Q2). To showcase our key insight that spatiotemporal signals are crucial for entity linking, we compare the base systems with their time or location-aware counterparts.

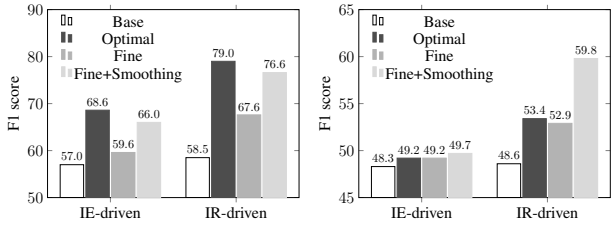
As reported in Fig. 3a, using either temporal (+T) or spatial (+L) signal can indeed improve entity linking over the base systems. In particular, the improvements in IR-driven evaluation are more significant since the chosen entities are known to be influenced by time or location.

Next, we combine both temporal and spatial signals. As Sect. 5 discussed, while the joint model (+T×L) is ideal in theory, it is not necessarily better than the separate model (+T+L) due to data scarcity, as shown in the last two rows of Fig. 3a. In the following, we will use "+T+L" as the default setting. Also note that using both time and location on LP may be worse than using either alone, given that LP has low precision and thus employing both signals may aggregate more errors.

Let us also examine the precision and recall in Fig. 3b and 3c. Interestingly, while precision is not compromised, recall is greatly increased with spatiotemporal signals. The reason is that E2E has high precision, but misses a lot of mentions (*i.e.*, linking them to $\emptyset$). However, if an entity e is "trending" at some time t or location l, that is, e is mentioned by more tweets than usual at t or l, we shall expect $P(e|t,l) > P(e)$. Thus, the system is more likely to link to e instead of $\emptyset$ based on our spatiotemporal model (3), resulting in higher recall. A more in-depth error analysis will be presented in Q4.

Graph-based smoothing (Q3). Next, we evaluate the utility of smoothing in Fig. 4. The model called "Fine" discretizes time and location using the finest bins. The one called "Fine+Smoothing" is our graph-based smoothing algorithm applied to the "Fine" model. We also compare with "Optimal", which discretizes time and location using the optimal bins tuned on the development set.

As we can see, smoothing is crucial when a small bin size is used. Compared to "Optimal", our smoothing is better on LP but not on E2E. However, smoothing is still useful, as "Optimal" depends a lot on choosing the right bin size, while smoothing is less sensitive to its parameters.



(a) Base system: E2E (b) Base system: LP

Figure 4: F1 scores by different ways of estimating spatiotemporal signals in the +T+L setting.

Error analysis (Q4). Let us investigate what causes the errors. In Table 3, we first break down the errors into false positives and false negatives. Each type is further categorized as "mention" (the mention itself

is wrong) or "linking" (the mention is correct but linking is wrong). Note that if the mention is wrong, the linking would also be wrong.

	False Positives			False Negatives		
	mention	linking	total	mention	linking	total
E2E	7	5	12	90	5	95
+T+L	9	7	16	64	7	71

Table 3: Sources of error in IE-driven evaluation.

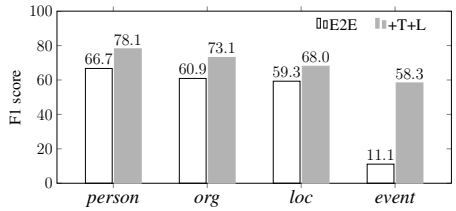
On the one hand, both systems make few false positives, suggesting high precision (see Fig. 3b). On the other hand, while the base system (E2E) makes many false negatives, the spatiotemporal model (+T+L) substantially reduces these errors, resulting in a significant increase in recall (see Fig. 3c). +T+L is particularly effective in recovering false negatives due to incorrect mention, *i.e.*, linking to $\emptyset$ when it should not, which account for the vast majority (84%) of all the errors in E2E.

Next, we examine how these errors can be recovered in +T+L. We zoom into entity types of *person*, *org*, *loc* and *event*, which collectively cover 85% of the entities in the test set. Fig. 5a reveals that +T+L is consistently helpful to different types of entities. In particular, +T+L improves the *event* entities the most, since events are generally more correlated to time and location, which are useful signals for recovering the errors.

We illustrate in Fig. 5b some entities that are missed by E2E, but recovered by useful spatial or temporal signals. Taking tweet #2 for illustration, E2E is not confident linking to the entity California State University, Fullerton based on textual content alone, and thus incorrectly links to $\emptyset$. This is a false negative due to incorrect mention, the most frequent kind in E2E (see Table 3). Fortunately, +T+L is confident linking to the entity correctly, given that 1) the tweet was posted during a campus emergency, *when* many other tweets were also discussing this entity; and 2) the posting user was in California, *where* the users of many other tweets discussing this entity were also located.

7 Qualitative Study: Entity vs. Keyword

The goal of this section is twofold. First, we provide a qualitative analysis to visualize the performance of our entity linking system. The results show that



(a) F1 scores

Tweet	Example entity	Posting time	User location
#1	*person*: Colin Kaepernick	during his game	*(not useful)*
#2	*org*: Cali. State U. Fullerton	in campus emergency	California
#3	*loc*: Los Angeles	*(not useful)*	California
#4	*event*: New Year's Eve	on 31 December	*(not useful)*

(b) Example entities recovered by spatiotemporal signals

Figure 5: IE-driven evaluation by entity type.

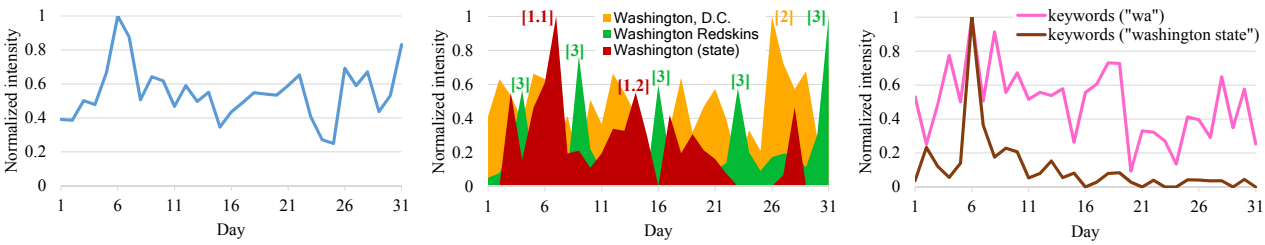
our system can consistently recover the real-world happenings or physical locations of the query entities. Second, we show that entity linking is more effective than keyword matching in the retrieving scenario (see Sect. 6.2).

Approach. We can retrieve a tweet by entity linking or keywords. For entity linking, we classify all the tweets using E2E+T+L, and a tweet is positive (*i.e.*, it will be retrieved) if its predictions contain the query entity. For keyword matching, a tweet is positive if it contains the given keywords which describe the query entity.

To visualize the retrieved tweets, we discretize time into 24-hour bins, and location into 100×100 sqkm bins. In a bin δ_i, define the *intensity* of an entity e as $\#(e, \delta_i)/\#(\delta_i)$, where $\#(e, \delta_i)$ is the number of positive tweets for e in δ_i, and $\#(\delta_i)$ is the total number of tweets in δ_i. To visualize the intensity of multiple entities simultaneously, we further compute the *normalized intensity* of e, which normalizes e's intensity by its maximum over all the bins.

Case study 1: "washington." We retrieve tweets with keyword "washington", a common anchor for three entities: Washington (state), Wasghinton D.C. and Washington Redskins. We also retrieve tweets for each of them with entity linking.

We first examine the intensity of the entities over December 2012. In Fig. 6a, the intensity based on keyword "washington" inevitably represents a mixture of different entities—it is unclear which "washington" entity corresponds to each intensity peak. Fortunately, entity linking separates the three enti-

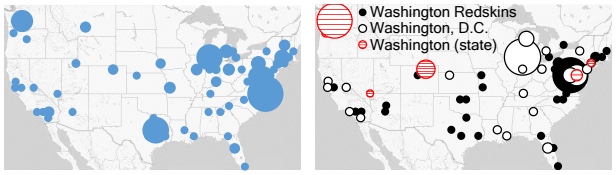


(a) with keywords ("washington") (b) with entity linking (c) with other keywords for Washington (state)

Figure 6: **Intensity of "washington" over December 2012.** In particular, entity linking (b) reveals the correspondence to several real-world happenings: *legalization of marijuana* in Washington (state) [1.1] and *Obama's response* [1.2], *fiscal cliff and weather alert* in Washington D.C. [2], *games* of Washington Redskins [3].

ties in Fig. 6b, where major intensity peaks correspond to real-world happenings for each entity. We also use alternative keywords specifically targeted for Washington (state), namely, "washington state" and "wa". While they are more precise than "washington" for the given entity, they have lower recall, resulting in different intensity profiles in Fig. 6c.

Lastly, we examine their intensity over the US. Using keywords, Fig. 6a does not isolate different entities. In contrast, using entity linking, Fig. 6b reveals more information about the three entities. For instance, tweets for Washington (state) are mostly concentrated in that state, and those for Washington Redskins are mostly in their home location but also have occurrences all over the US.



(a) with keywords ("washington") (b) with entity linking

Figure 7: Intensity of "washington" over the US. The size of dots indicates the normalized intensity.

Case study 2: "spurs." We retrieve tweets with keyword "spurs", a popular anchor for both San Antonio Spurs and Tottenham Hotspur F.C. We also retrieve tweets for each of them with entity linking.

As illustrated in Fig. 8, the intensity based on keyword "spurs" roughly matches that of San Antonio Spurs based on entity linking, but significantly differs from Tottenham Hotspur F.C. In other words, "spurs" only accounts for the former entity, even though it is also a standard anchor for the latter. The reason is that an overwhelming majority of the mentions of "spurs" in the US refer to San Antonio Spurs. That means, while keywords may work for the dominating entity, they are particularly weak for the other entity. In contrast, with entity linking, both entities can be identified, where the major peaks correspond to the game days of each team.

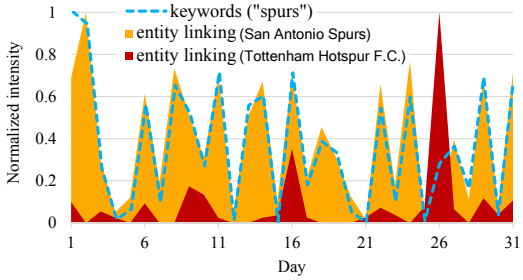


Figure 8: Intensity of "spurs" over December 2012.

8 Conclusion

As microblogging and other social media services become increasingly popular, the number of short and noisy messages are growing at an unprecedented rate. We demonstrate, for the first time, that spatiotemporal signals are critical in advancing entity linking. After all, it might not be a mere text-based language processing problem.

There are some important future directions for this work. First, updating the spatiotemporal model on the fly is a useful extension to cater to the online streaming setting. Second, we foresee a more general framework to integrate various meta-data such as authorship, surrounding conversation, attached URL and hashtags, in addition to the spatiotemporal signals. Finally, it would also pay off to analyze the interactions between the meta-data and the use of language. We believe that exploring the meta-data for entity linking on microblogs will be an interesting and active line of research.

References

S. Asur and B.A. Huberman. 2010. Predicting the future with social media. In *Web Intelligence*, pages 492–499.

E. Benson, A. Haghighi, and R. Barzilay. 2011. Event discovery in social media feeds. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 389–398.

R. C Bunescu and M. Pasca. 2006. Using encyclopedic knowledge for named entity disambiguation. In *Proceedings of the European Chapter of the ACL (EACL)*, pages 9–16.

M. Cataldi, L. Di Caro, and C. Schifanella. 2010. Emerging topic detection on twitter based on temporal and social terms evaluation. In *Proceedings of the International Workshop on Multimedia Data Mining (MDMKDD)*, pages 4:1–10.

M. Collins. 2002. Discriminative training methods for hidden Markov models: Theory and experiments with perceptron algorithms. In *Proceedings of the Conference on Empirical Methods for Natural Language Processing (EMNLP)*.

S. Cucerzan. 2007. Large-scale named entity disambiguation based on Wikipedia data. In *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 708–716.

N. Dalvi, R. Kumar, and B. Pang. 2012. Object matching in tweets with spatial models. In *Proceedings of International Conference on Web Search and Web Data Mining (WSDM)*, pages 43–52.

J. Eisenstein, B. O'Connor, N.A. Smith, and E.P. Xing. 2010. A latent variable model for geographic lexical variation. In *Proceedings of the Conference on Empirical Methods for Natural Language Processing (EMNLP)*, pages 1277–1287.

Y. Fang, B.-J. Hsu, and K. C.-C. Chang. 2012. Confidence-aware graph regularization with heterogeneous pairwise features. In *Proceedings of International Conference on Research and Development in Information Retrieval (SIGIR)*, pages 951–960.

Y. Fang, K. C.-C. Chang, and H.W. Lauw. 2013. Roundtriprank: Graph-based proximity with importance and specificity. In *Proceedings of International Conference on Data Engineering (ICDE)*, pages 613–624.

Y. Fang, K. C.-C. Chang, and H.W. Lauw. 2014. Graph-based semi-supervised learning: Realizing pointwise smoothness probabilistically. In *Proceedings of the International Conference on Machine Learning (ICML)*.

P. Ferragina and U. Scaiella. 2010. TAGME: on-the-fly annotation of short text fragments (by Wikipedia entities). In *Proceedings of ACM Conference on Information and Knowledge Management (CIKM)*, pages 1625–1628.

S. Guo, M.-W. Chang, and E. Kıcıman. 2013. To link or not to link? A study on end-to-end tweet entity linking. In *Proceedings of the Annual Meeting of the North American Association of Computational Linguistics (NAACL)*, pages 1020–1030.

X. Han, L. Sun, and J. Zhao. 2011. Collective entity linking in web text: a graph-based method. In *Proceedings of International Conference on Research and Development in Information Retrieval (SIGIR)*, pages 765–774.

J. Hoffart, M. Amir Yosef, I. Bordino, H. Fürstenau, M. Pinkal, M. Spaniol, B. Taneva, S. Thater, and G. Weikum. 2011. Robust disambiguation of named entities in text. In *Proceedings of the Conference on Empirical Methods for Natural Language Processing (EMNLP)*, pages 782–792.

H. Ji and R. Grishman. 2011. Knowledge base population: Successful approaches and challenges. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1148–1158.

H. Ji, R. Grishman, H.T. Dang, K. Griffitt, and J. Ellis. 2010. Overview of the TAC 2010 knowledge base population track. In *Text Analysis Conference (TAC)*.

S. Kulkarni, A. Singh, G. Ramakrishnan, and S. Chakrabarti. 2009. Collective annotation of Wikipedia entities in web text. In *Proceedings of International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 457–466.

J. Leskovec, L. Backstrom, and J. Kleinberg. 2009. Meme-tracking and the dynamics of the news cycle. In *Proceedings of International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 497–506.

C. Li, J. Weng, Q. He, Y Yao, A. Datta, A. Sun, and B.-S. Lee. 2012a. TwiNER: named entity recognition in targeted twitter stream. In *Proceedings of International Conference on Research and Development in Information Retrieval (SIGIR)*, pages 721–730.

R. Li, S. Wang, H. Deng, R. Wang, and K. Chen-Chuan Chang. 2012b. Towards social user profiling: unified and discriminative influence model for inferring home locations. In *Proceedings of International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 1023–1031.

R. Li, S. Wang, and K. Chen-Chuan Chang. 2013. Towards social data platform: Automatic topic-focused monitor for twitter stream. *Proceedings of the VLDB Endowment (PVLDB)*, 6(14).

Q. Mei, C. Liu, H. Su, and C.-X. Zhai. 2006. A probabilistic approach to spatiotemporal theme pattern mining on weblogs. In *Proceedings of the International World Wide Web Conference (WWW)*, pages 533–542.

E. Meij, W. Weerkamp, and M. de Rijke. 2012. Adding semantics to microblog posts. In *Proceedings of International Conference on Web Search and Web Data Mining (WSDM)*, pages 563–572.

D. Milne and I. H. Witten. 2008. Learning to link with Wikipedia. In *Proceedings of ACM Conference on Information and Knowledge Management (CIKM)*, pages 509–518.

J. Platt. 2000. Probabilistic outputs for support vector machines and comparison to regularize likelihood methods. In *Advances in Large Margin Classifiers*, pages 61–74.

L. Ratinov, D. Roth, D. Downey, and M. Anderson. 2011. Local and global algorithms for disambiguation to Wikipedia. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1375–1384.

A. Ritter, S. Clark, Mausam, and O. Etzioni. 2011. Named entity recognition in tweets: an experimental study. In *Proceedings of the Conference on Empirical Methods for Natural Language Processing (EMNLP)*, pages 1524–1534.

A. Ritter, Mausam, O. Etzioni, and S. Clark. 2012. Open domain event extraction from Twitter. In *Proceedings of International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 1104–1112.

A. Sadilek, H. Kautz, and V. Silenzio. 2012. Modeling spread of disease from social interactions. In *Proceedings of the International Conference on Weblogs and Social Media (ICWSM)*.

T. Sakaki, M. Okazaki, and Y. Matsuo. 2010. Earthquake shakes Twitter users: real-time event detection by social sensors. In *Proceedings of the International World Wide Web Conference (WWW)*, pages 851–860.

D.W. Scott. 2009. *Multivariate density estimation: theory, practice, and visualization*. John Wiley & Sons.

W. Shen, J. Wang, P. Luo, and M. Wang. 2012. LINDEN: linking named entities with knowledge base via semantic knowledge. In *Proceedings of the International World Wide Web Conference (WWW)*, pages 449–458.

W. Shen, J. Wang, P. Luo, and M. Wang. 2013. Linking named entities in tweets with knowledge base via user interest modeling. In *Proceedings of International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 68–76.

A. Sil and A. Yates. 2013. Re-ranking for joint named-entity recognition and linking. In *Proceedings of ACM Conference on Information and Knowledge Management (CIKM)*.

A. Tumasjan, T. O. Sprenger, P. G Sandner, and I. M Welpe. 2010. Predicting elections with Twitter: What 140 characters reveal about political sentiment. In *Proceedings of the International Conference on Weblogs and Social Media (ICWSM)*, pages 178–185.

V. Vapnik and S. Mukherjee. 1999. Support vector method for multivariate density estimation. In *Proceedings of the Conference on Advances in Neural Information Processing Systems (NIPS)*, pages 659–665.

B.P. Wing and J. Baldridge. 2011. Simple supervised document geolocation with geodesic grids. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 955–964.

J.-M. Xu, A. Bhargava, R. Nowak, and X. Zhu. 2012. Socioscope: Spatio-temporal signal recovery from social media. In *Proceedings of the European Conference on Machine Learning (ECML)*, pages 644–659.

Z. Yin, L. Cao, J. Han, C. Zhai, and T. Huang. 2011. Geographical topic discovery and comparison. In *Proceedings of the International World Wide Web Conference (WWW)*, pages 247–256.

X. Zhu, Z. Ghahramani, J. Lafferty, et al. 2003. Semi-supervised learning using Gaussian fields and harmonic functions. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 912–919.