Singapore Management University

Institutional Knowledge at Singapore Management University

Research Collection School Of Computing and Information Systems | School of Computing and Information Systems

4-2017

# Collective entity linking in tweets over space and time

Wen Haw CHONG
*Singapore Management University*, whchong.2013@phdis.smu.edu.sg

Ee-peng LIM
*Singapore Management University*, eplim@smu.edu.sg

William COHEN
*Carnegie Mellon University*

Follow this and additional works at: https://ink.library.smu.edu.sg/sis_research

 Part of the Databases and Information Systems Commons, Social Media Commons, and the Software Engineering Commons

## Citation

CHONG, Wen Haw; LIM, Ee-peng; and COHEN, William. Collective entity linking in tweets over space and time. (2017). *Advances in Information Retrieval: 39th European Conference on IR Research, ECIR 2017, Aberdeen, UK, April 8-13, Proceedings*. 10193, 82-94.
**Available at:** https://ink.library.smu.edu.sg/sis_research/3720

This Conference Proceeding Article is brought to you for free and open access by the School of Computing and Information Systems at Institutional Knowledge at Singapore Management University. It has been accepted for inclusion in Research Collection School Of Computing and Information Systems by an authorized administrator of Institutional Knowledge at Singapore Management University. For more information, please email cherylds@smu.edu.sg.

# Collective Entity Linking in Tweets Over Space and Time

Wen-Haw Chong[1(✉)], Ee-Peng Lim[1], and William Cohen[2]

[1] Singapore Management University, 80 Stamford Road, Singapore 178902, Singapore
whchong.2013@phdis.smu.edu.sg, eplim@smu.edu.sg
[2] Carnegie Mellon University, Pittsburgh, PA 15213, USA
wcohen@cs.cmu.edu

**Abstract.** We propose *collective entity linking* over tweets that are close in space and time. This exploits the fact that events or geographical points of interest often result in related entities being mentioned in spatio-temporal proximity. Our approach directly applies to geocoded tweets. Where geocoded tweets are overly sparse among all tweets, we use a relaxed version of spatial proximity which utilizes both geocoded and non-geocoded tweets linked by common mentions.

Entity linking is affected by noisy mentions extracted and incomplete knowledge bases. Moreover, to perform evaluation on the entity linking results, much manual annotation of mentions is often required. To mitigate these challenges, we propose *comparison-based evaluation*, which assesses the change in linking quality when one linking method modifies the output of another. With this evaluation we show that differences between collective linking and local linking, i.e. linking entities in each tweet individually, are statistically significant. In extensive experiments, collective linking consistently yields more positive changes to the linking quality, than negative changes. The ratio of positive to negative changes varies from 1.44 to 12, depending on the experiment settings.

**Keywords:** Concept linking · Entity linking · Entity disambiguation

## 1 Introduction

We explore entity linking for mentions in tweets. In entity linking, one links mentions in text, usually of named entities, to the referent entities in a given Knowledge Base (KB). Entity linking is also referred to as *entity disambiguation* or *concept linking* and is very similar to *Word Sense Disambiguation* (WSD). WSD aims to identify the correct sense of a word in a piece of text. Compared to WSD, entity linking focuses on named entity mentions. However it is unrealistic to expect detected mentions in tweets to match named entities only. This is due to social media content being written in an informal, case-insensitive manner. This increases mistakes by Named Entity Recognition (NER) tools, e.g. mistaking the term 'Merry' in the phrase 'Merry Christmas' as a named entity. Although prior work [5,11,14] do not consider such cases, in practical applications, they are

impossible to exclude entirely. Hence in our entity linking work, we also cover cases of such noisy mentions. We use Wikipedia as the KB for linking. Thus our work can also be considered as Wikification [4,11].

Entity linking for social media content is challenging, as social media documents are short e.g. tweets, Foursquare shouts etc. Thus mentions arise in short documents, which may lack enough content or context for deriving features. This motivates the use of collective linking, i.e. exploiting information from multiple tweets to link mentions in a single tweet. Prior work [4,14] had considered collective linking over multiple tweets from the same user, and tweets linked by common terms or hashtags. In this work, we focus on the orthogonal aspects of space and time for collective linking. This is motivated by observations of tweeting behaviour with respect to events and geographical effects.

**Event Effects.** Tweets may be event related [1]. When tweet-worthy events occur, users may tweet about related entities, leading to an excess of related mentions in a space-time cube, i.e. a certain time period defined over a geographical area. Within a space-time cube, we can conduct collective linking and share linkage information across tweets. For example, the following are two actual tweets close in space and time: <*Stones*> and <Waiting for @*RollingStones* to come on stage so we can rock out *Singapore*>. Consider the mentions in italics. The first tweet has insufficient context for linking *Stones*. The second tweet's mentions can be linked with much less ambiguity, since *RollingStones* refers to the band entity 'The Rolling Stones' with high probability [15]. Given both tweets' space-time proximity, one can now use the second tweet's results to link the first tweet's *Stones* to the band with much more certainty.

**Geographical Effects.** Besides events, locations also affect tweeting behaviour. Certain entities may be more prevalent and mentioned more frequently at certain locations. Thus we can exploit geographical effects by collectively linking tweets that are close in space. For example, compare the following two tweets with mentions in italics: <*MBS* #throwback>, <Standby for SHOWTIME! @ *Marina Bay Sands*>. *MBS* in the first tweet is the surface form for many possible entities. The probability that it refers to 'Marina Bay Sands', a Singapore tourist attraction, is extremely low [15] at 0.000155. However if the second tweet with unambiguous mentions to 'Marina Bay Sands' occurs spatially near the first tweet, then it is much more plausible for the latter to be mentioning the same entity. Both event and geographical effects are often coupled due to events at Points of Interest (POI), e.g. concerts at a tourist attraction.

**Challenges and Contributions.** Our main contribution is a new collective entity linking method to exploit event and geographical effects. We connect tweets close in space and time to form a tweet graph, and define a novel objective function over the graph. This mitigates the challenge of entity linking for overly brief content. In addition, we introduce a comparison-based evaluation approach (see Sect. 4), which addresses the following challenges in evaluation:

– Noisy mention extraction: Automated mention extraction is noisy with mentions often being extracted in part, e.g. extracting *Garden* given *Madison Square Garden* or non-named entities being mistaken as named entities.

– Incomplete Knowledge Base: Many mentioned entities are often not in knowledge bases, even for a comprehensive KB such as Wikipedia.
– Annotation Effort: Expensive manual annotation is required to construct ground truth linking for mentions in order to evaluate linking accuracy.

Based on comparison-based evaluation, the differences between collective linking and local linking, i.e. linking entities in each tweet individually, are statistically significant. Over the results of local linking, collective linking made many more positive changes, i.e. that improves linking quality, than negative changes.

## 2 Related Work

The work in [9] introduces a semantic relatedness measure to quantify coherence. The measure uses only Wikipedia hyperlink structure and is inexpensive to compute. The main idea is that semantically related entities should share many common neighbors in Wikipedia. We use the same measure in our work.

For entity linking in tweets, Meij et al. [8] employed extensive feature engineering on content, page links and lexical word form. They then trained decision trees for ranking entities that are related to each tweet (rather than each mention). For linking individual mentions, Liu et al.'s work [6] maximized an objective derived from coherence, mention-concept features and mention-mention features. The objective requires training of feature weights. In [14], the idea is to exploit user interest for linking. A user's interest scores over entities are initialized and propagated over a graph of entities linked by relatedness [9]. Given a new mention with multiple candidate entities, entities with higher interest score are preferred. Huang et al. [4] proposed label propagation over a different form of graph. Graph nodes are mention-entity tuples, connected based on weighted combination of various relations, e.g. coreferencing mentions, semantic-relatedness [9] etc. After label propagation, high ranking tuples provide the linking results.

Different from the above works, we focus on orthogonal aspects such as spatial and temporal proximity between tweets. In these aspects, the work by Fang and Chang [2] is related. They learned entity distributions over time and large geographical areas (smallest area considered is $100\,\mathrm{km}^2$) in a weakly supervised setting. In contrast, we work in the unsupervised setting and consider much smaller geographical areas spanning hundreds of meters. For an unsupervised approach, TAGME [3] is applicable. Its key idea is: within the same document, candidate entities across mentions vote for each other. For a given mention, the entity with the highest prior is then selected from the top most voted entities. We shall also implement TAGME as a non-collective entity linking baseline.

## 3 Approach

### 3.1 System Architecture

Our system architecture comprises of **Pre-processing**, **Local linking** and **Collective linking**. Given a set of tweets for entity linking, the first pre-processing step is mention extraction with an NER tool. The process is often

noisy with mentions being omitted or extracted partially. To mitigate this, we apply TweetNLP [10], which was specially developed for tweets. Next, for each extracted mention, we use the Google lexicon [15] to identify candidate Wikipedia entities. The lexicon lists possible mentions $\{m\}$ for each entity $e$ along with the occurrence probability $p(e|m)$ derived from web hyperlinks.

In local linking, mentions to entities are linked individually for each tweet, without considering information from other tweets. We implemented two local linking methods: TAGME [3] and Loclink, introduced in Sect. 3.2. Local linking can be used to initialize the entity assignments for collective linking.

In collective linking, each mention in a tweet is linked using information within that tweet and from other tweets. Collective linking comprises three steps:

- **Tweet Graph Construction:** We first construct a graph that connects tweets by spatio-temporal proximity. The tweet graph is used to propagate information. Section 3.3 describes the construction process.
- **Initialization:** This means assigning an initial entity to each mention for subsequent refinement. This can be done using the results from local linking or with some other heuristics. We have opted for the former.
- **Optimization:** We define an objective function over the tweet graph and search for entity assignments to optimize it. Refer to Sect. 3.3.

### 3.2 LocLink: A Local Linking Method

Local linking processes each tweet individually, assigning entities that are semantically related to each other to make each tweet coherent. To quantify coherence, we adopt the semantic relatedness measure proposed in [9]. Consider entity $e_a$. Denote other entities with outgoing links to $e_a$ as the set $I(e_a)$. Equivalently, regard $e_a$ as having $|I(e_a)|$ incoming neighbors. For a pair of entities $e_a$, $e_b$ with overlapping incoming neighbors, semantic-relatedness is then computed as:

$$SR(e_a, e_b) = 1 - \frac{log(max\{|I(e_a)|, |I(e_b)|\}) - log|I(e_a) \cap I(e_b)|}{log(|W|) - log(min\{|I(e_a)|, |I(e_b)|\})} \tag{1}$$

where $I(e_a) \cap I(e_b)$ are entities which link to both $e_a$, $e_b$ in Wikipedia and $W$ is the total number of Wikipedia entities. If $I(e_a) \cap I(e_b) = \emptyset$, we set $SR(e_a, e_b) = 0$.

**Intra-tweet Coherence.** Let $d_i$ represent the $i$-th tweet containing $|m_i|$ mentions with set of linked entities $\mathbf{e}_i$. Also let $m_{ia}$ be the $a$-th mention of $d_i$, with corresponding linked entity $e_{ia}$. We define the intra-tweet coherence as average semantic relatedness between its assigned entities:

$$C(d_i, \mathbf{e}_i) = \frac{1}{0.5|m_i|(|m_i| - 1)} \sum_{a=1}^{|m_i|} \sum_{b>a}^{|m_i|} SR(e_{ia}, e_{ib}) \tag{2}$$

Maximizing intra-tweet coherence makes each tweet as coherent as possible. However assigned entities can be rather obscure or rare. Hence a prior $p(e|m)$ is

usually included [6,12,14] to favor more popular entities. In fact using only the prior for entity linking is a surprisingly strong baseline [5,11], while including the notion of coherence improves performance further. We use the prior from [15] and define the objective function for tweet $d_i$ as:

$$Q_i(d_i, \mathbf{e}_i) = \xi.C(d_i, \mathbf{e}_i) + \frac{\tau}{|m_i|} \sum_{a=1}^{|m_i|} p(e_{ia}|m_{ia}) \tag{3}$$

where $\xi$ and $\tau$ are combination weights. In the unsupervised setting, we simply let $\xi = \tau$ and assign entities to maximize $Q_i$. For single-mention tweets, coherence is undefined and we simply assign the entity with the highest prior to the mention. We call the above local linking method as **LocLink**.

### 3.3 Collective Linking in Space and Time

**Inter-tweet Coherence.** For collective linking, we exploit the fact that different tweets close in space and time may be related to the same event or have a common geographical effect, e.g. mentioning a common location. Therefore we expect some of the tweets to be *inter-coherent*. For computational efficiency, we shall only consider tweet pairs. Given tweets $d_i$ and $d_j$ with respective linked entity sets $\mathbf{e}_i$ and $\mathbf{e}_j$, we define the inter-tweet coherence as:

$$C(d_i, d_j, \mathbf{e}_i, \mathbf{e}_j) = \frac{1}{|m_i|.|m_j|} \sum_{a=1}^{|m_i|} \sum_{b=1}^{|m_j|} SR(e_{ia}, e_{jb}) \tag{4}$$

**Tweet Graph Construction.** Denote tweet $d_i$'s timestamp as $t_i$ and its location as $l_i$. In the simplest graph building scenario, we first retrieve geocoded tweets from a desired time interval and geographical area. For convenience, we call this a space-time cube although the geographical area need not be rectangular. For every pair of tweets $d_i$ and $d_j$, we connect them if $|t_i - t_j| \leq \delta_t$ and $dist(l_i, l_j) \leq \delta_d$, where $\delta_t$ and $\delta_d$ are the respective thresholds for temporal and spatial proximities, and $dist()$ measures geographical distance.

We can relax the spatial requirement to include non-geocoded tweets. This assumes that non-geocoded tweets related to an event/POI may mention similar entities as the geocoded tweets. Thus from geocoded tweets in the initial space-time cube, we first extract mentions. We then query for more tweets with similar mentions and from same-city users (based on their profiles). We now have a mixture of tweets with and without location information. To consistently form the graph, we connect tweets based only on temporal proximity, i.e. $|t_i - t_j| \leq \delta_t$. Note that although individual edges are based on temporal proximity, the overall graph incorporates spatial-proximity since tweets are constrained to be from the initial space-time cube or users in the same city.

**Objective Function.** Let $D$ and $E$ be the set of nodes and edges respectively in the tweet graph. We define our objective function for collective linking:

$$Q(D, E, \mathbf{e}) = \frac{\alpha}{|D|} \sum_{i=1}^{|D|} C(d_i, \mathbf{e}_i) + \frac{\beta}{|E|} \sum_{(d_i,d_j)\in E}^{|E|} C(d_i, d_j, \mathbf{e}_i, \mathbf{e}_j) + \frac{\gamma}{|M|} \sum_{i=1}^{|T|} \sum_{a=1}^{|m_i|} p(e_{ia}|m_{ia}) \quad (5)$$

where $|M|$ is the total number of mentions, with set of linked entities $\mathbf{e}$; and $\alpha$, $\beta$ and $\gamma$ are global combination weights. Essentially $Q$ is a linear combination of intra-tweet coherence, inter-tweet coherence and the entity prior term. Thus $Q$ encapsulates our earlier discussed intuitions about coherence and entity popularity. For a fixed set of weights, the optimization problem is to assign entities to mentions to maximize $Q$. For optimization, we use the decoding algorithm [6].

**Parameter Settings.** We consider unsupervised collective linking where labeled data is unavailable. Given that tuning/training is not possible, we consider two intuitive cases of averaging. In the first case, we use uniform weights in $Q$, i.e. $\alpha = \beta = \gamma$. We referred to this setting as *Uniform*. Alternatively, one can regard coherence and entity prior as very different notions and assign them equal importance. Hence in the second case, one averages over coherence and the entity prior, i.e. $\alpha = \beta, \gamma = \alpha + \beta$. We denote this setting *Avg(Coh,prior)*.

## 4 Comparison-Based Evaluation

Instead of heuristics/random initialization, we use local linking to initialize collective linking. This leads to a comparison-based evaluation approach. Essentially we compare initial and final linkings and determine if a change is an improvement (positive change), a degradation (negative change) or neither. As will be explained, there are several advantages in such an evaluation.

**Annotation Effort.** Firstly, we only need to compare linkings which are different between two linking results. This reduces the data annotation effort, compared to traditional evaluation using accuracy [13], i.e. proportion of correctly linked mentions. For example, to compute accuracy for a dataset of 100 mentions, each mention first has to be linked to the correct KB entity, typically via manual annotation [8]. In our evaluation framework, the annotation effort depends on the linkage differences between techniques and is usually less. For example, if all 100 mentions are linked by local linking and collective linking suggested 5 changes, then we only need to examine 5 changes. Clearly, more positive than negative changes is desired and implies improved performance.

**Incomplete KB and Imperfect Linking.** No KB can cover all mentioned entities. One can ignore unlinkable mentions or link them to the catch-all NIL entity [7,13,14]. However this discards data that may be useful for evaluation. Related to this, there is also the notion of how fine-grained a linkage needs to be, in order to be considered correct. Mentions can be linked to entities at different type or instance granularities. If one considers all coarse-grained linkages as wrong, many linkages useful for comparing techniques will be discarded.

For example, consider Table 1. The tweet was sent from the game venue during a college football match between Duke and Indiana University. Linking the mention *Duke* to Wikipedia, the most fine grained entity is $e_1$, i.e. Duke University's football team. However a linking technique may miss this perfect linking and choose other entities. Table 1 also lists Wikipedia entities in decreasing order of relatedness to the actual football team. Consider two techniques, one linking *Duke* to $e_2$, the other to $e_4$. Clearly the former provides useful information, even though both techniques miss out on $e_1$. In such cases, we still want to differentiate both techniques instead of regarding both linkings as equally wrong. If $e_1$ is not in the KB, but parent organizations such as $e_2$ and $e_3$ are present, it is still possible and reasonable to compare linking performance on *Duke*, instead of just discarding the mention as unlinkable. This calls for a comparison-based kind of evaluation.

**Table 1.** A sample tweet with mentions (in Italics). Row 2 lists candidate Wikipedia entities for the mention *Duke*, in decreasing relatedness.

| Go *Duke*! *#PinstripeBowl* @ *Yankee Stadium* |
|---|
| • $e_1$: Duke_Blue_Devils_football: Duke University's football team |
| • $e_2$: Duke_Blue_Devils: Duke University's varsity sports team |
| • $e_3$: Duke_University: Duke University |
| • $e_4$: Duke: Monarch ruling over a duchy |

**Noisy Mention Extraction.** Automated mention extraction is noisy. Often, incomplete sub-mentions are extracted. Even in cases where a mention should link to a unique entity, the notion of correct/wrong linking is less clear when sub-mentions are involved. Fortunately in comparison-based evaluation, we can compare entity assignments and pick the better one. For example, consider the tweet <Watching *Jeff Dunham* @*star* performing arts centre with the family>, where mentions (in italics) were extracted with TweetNLP [10]. The complete venue mention is *star performing arts centre*. However the sub-mention *star* was extracted, constraining entity linking to link *star*. Instead of discarding such cases, one can still compare linking results, e.g. linking to 'Movie_star' is intuitively preferred over 'Star': a luminous sphere of plasma in space. On a related note, if an extracted mention is in fact that of a non named-entity, such comparisons can also be used for evaluation.

### 4.1 Evaluating Changes

To evaluate changes, we define what constitutes each outcome. Firstly, we observe changes to often reduce or increase the specificity/granularity of linked entities. This leads to the consideration of parent-child relationships between entities in a type hierarchy. For brevity of discussion, we overload the term of entity types such that types can refer to semantic categories, organizations or

locations. A super-type is decomposable into sub-types of finer granularities and this is applicable to semantic categories, instances, organizations and locations. For example entity $e_1$:'Duke_Blue_Devils_Football' is a sports team instance under the semantic category of 'American_football', and also a child organization of 'Duke_University'. For a location example 'New_York_City' (NYC) contains (and is the parent of) 'Madison_ Square_Garden', a multi-purpose indoor arena.

Clearly, we are considering more parent-child relationships beyond the semantic categories in ontologies. Hence any automated evaluation using only ontologies, e.g. the Dbpedia ontology[1] will be highly incomplete. Instead we compare type information using Wikipedia content when assessing linkage changes, e.g. $e_1$'s Wikipedia page starts with '*The Duke Blue Devils Football team represents Duke University in the sport of American football*'.

We now discuss *positive changes* using Table 1:

- **Incorrect linking to parent entity/correct linking:** In this case, initial linking is unrelated and wrong, e.g. linking *Duke* to 'Duke', ruler of a Duchy. Changing the linking to either 'Duke_University' (a parent entity) or 'Duke_Blue_Devils_football' (the correct linking) is a positive change.
- **Parent entity to correct linking:** Eg. changing the linking for *Duke* from 'Duke_University' to 'Duke_Blue_Devils_football'. Intuitively, this provides more specific information to the system user.
- **Ancestor entity to parent entity:** In this case, the final linking is still not perfect, however the information specificity is increased, eg. changing the linking for *Duke* from 'Duke_University' to 'Duke_Blue_Devils'.
- **Incorrect sibling entity to parent entity:** We regard coarse-grained, related information as more useful than specific, but wrong information, e.g. if *Duke* is initially linked wrongly to 'Duke_Blue_Devils_men's_basketball' and changed to 'Duke_Blue_Devils', it counts as a positive change.

For the above, reversing the change direction count as *negative changes*. In addition, changes can be neither positive nor negative, e.g. replacing an incorrect entity with another. Such "neither" changes also include changing an initial unrelated entity assignment to a sibling or child entity, although this arguably improves our understanding of the tweets involved. For example, if *Duke* in Table 1 is initially linked to 'Duke' and changed to 'Duke_Blue_Devils_men's_basketball', we count it as a neither. Section 5.2 provides examples from experiments.

## 5 Experiments

**Data.** We conduct experiments on New York City (NYC) and Singapore (SG) tweets. To obtain meaningful tweets for linking (instead of trivial blabber [8]), we collect tweets near POIs or in space-time cubes covering performance events. For NYC, we obtained geocoded tweets from the CHIMPS Lab[2] that are within

[1] http://mappings.dbpedia.org/server/ontology/classes/.
[2] http://cmuchimps.org/.

100 meters of five popular event venues. For each venue, we consider two evenings (18:00–22:00) in Dec 2015 with the most tweets, obtaining 10 space-time cubes with an average of 24.8 tweets. For each cube, we form a spatio-temporal tweet graph for collective linking where tweets within 1 h and 100 m of each other are connected. For Singapore, we relax the spatial proximity requirement as discussed in Sect. 3.3 and obtain an average of 46.47 tweets over space-time cubes covering 17 performance events. The tweets are a mixture of geocoded and non-geocoded tweets. We connect tweets within 1 h of each other. Note that although individual edges in the tweet graph are based on temporal proximity, there is still a coarse notion of spatial proximity as most tweets are from Singapore, a small geographical area.

Following tweet graph construction, we apply both manual and automated mention extraction. For the latter, we use TweetNLP. For manual mention extraction, we process all 10 space-time cubes for NYC and 8 space-time cubes (out of 17) for SG, selected based on largest number of tweets. We link all mentions regardless of whether the parent tweets are related to the POI or event.

**Local Linking Baselines.** We use collective linking to modify the results of local linking. Thus the latter are equivalent to baselines. We implement *LocLink* (Sect. 3.2) with uniform weights for the objective in Eq. (3). We also implement *TAGME* [3], which is based on weighted voting among candidate entities.

### 5.1 Results

Results are summarized in Table 2 for New York City (NYC) tweets and Table 3 for Singapore (SG) tweets. Comparing collective linking to local linking, we see linkage improvements across all experiment settings. Consistently, collective linking makes more positive changes than negative changes, when applied on the results of local linking. In most cases, the ratio of positive to negative changes is larger than 2. The highest ratio is 12, for the experiment using NYC tweets with manually extracted mentions, TAGME for local linking and averaging over coherence and entity for $Q$, i.e., $Avg(coh, prior)$. The lowest ratio is 1.44, again on NYC tweets and with TweetNLP, LocLink and $Avg(coh, prior)$.

**Table 2.** Results on NYC tweets. Bracketed numbers are counts of unique mentions over which changes occur. ($\Delta$: total changes, +ve: total positive, −ve: total negative, Ratio: +ve/−ve. **: significant at $p$-value = 0.01, *: sig. at $p$-value = 0.05)

| Local linking method | | LocLink | | | | TAGME | | | |
|---|---|---|---|---|---|---|---|---|---|
| Mentions | Setting | $\Delta$ | +ve | −ve | Ratio | $\Delta$ | +ve | −ve | Ratio |
| Manual | *Uniform* | 43 | 22 (14) | 9 (6) | 2.44** | 73 | 37 (18) | 6 (5) | 6.17** |
| Manual | *Avg(coh, prior)* | 20 | 13 (9) | 3 (3) | 4.33* | 62 | 36 (18) | 3 (3) | 12.00** |
| TweetNLP | *Uniform* | 61 | 23 (14) | 11 (10) | 2.09* | 103 | 38 (19) | 13 (12) | 2.92** |
| TweetNLP | *Avg(coh, prior)* | 50 | 13 (8) | 9 (7) | 1.44 | 95 | 35 (18) | 9 (7) | 3.89** |

**Table 3.** Results on SG tweets. Notations as in Table 2.

| Local linking method | | LocLink | | | | TAGME | | | |
|---|---|---|---|---|---|---|---|---|---|
| Mentions | Setting | $\Delta$ | +ve | −ve | Ratio | $\Delta$ | +ve | −ve | Ratio |
| Manual | *Uniform* | 59 | 22 (10) | 7 (4) | 3.14** | 93 | 38 (14) | 8 (6) | 4.75** |
| Manual | *Avg(coh,prior)* | 28 | 16 (7) | 2 (2) | 8.00** | 78 | 37 (16) | 8 (6) | 4.63** |
| TweetNLP | *Uniform* | 83 | 29 (10) | 9 (7) | 3.22** | 168 | 61 (21) | 30 (8) | 2.03** |
| TweetNLP | *Avg(coh,prior)* | 44 | 23 (8) | 2 (2) | 11.5** | 128 | 54 (23) | 23 (6) | 2.35** |

Our results are statistically significant. Considering positive and negative changes, we conducted significance testing with the binomial test. The null hypothesis is that the proportion of positive and negative changes are equal. Except for one setting (TweetNLP, LocLink and $Avg(coh, prior)$), we are able to reject the null hypothesis at $p$-value of 0.05.

In both Tables 2 and 3, we also tabulate the number of unique mentions (in brackets) over which changes are made. This provides another view of the results accounting for mention diversity. In the trivial case, if all mentions are identical and initially wrongly linked, then it is easy to achieve many positive changes just from correcting one unique mention. However this overstates the performance advantage of collective linking due to a lack of mention diversity. From both tables, we see that the number of unique mentions for positive changes is consistently larger than that for negative changes, which is reassuring.

Collective linking exerts much of its influence through inter-tweet coherence. Recall that for $Uniform$, we use uniform weights for $Q$, while for $Avg(coh, prior)$, weight for the entity prior is set equal to total weights from intra and inter-tweet coherence. Thus in $Avg(coh, prior)$, inter-tweet coherence has smaller relative weight and plays a smaller role in affecting the linking results. This means that collective linking should suggest fewer changes. Indeed, we see that for a fixed mention extraction and local linking method, there are always fewer changes in $Avg(coh, prior)$ than $Uniform$.

### 5.2 Qualitative Analysis

Many, but not all changes are shared across experiments. Due to space constraints, we only illustrate changes for one experiment on NYC: TweetNLP for mention extraction, TAGME for local linking and uniform weighting for $Q$. Sample tweets are displayed in Tables 4 to 6, along with changes in the format: Initial entity → final entity. Readers can inspect Wikipedia entities by appending the entity name to the URL 'https://en.wikipedia.org/wiki/'.[3]

**Positive Changes.** Table 4 shows positive changes. Tweets N1 and N2 are from a college football match between Duke and Indiana University. The mention

[3] e.g. entity 'Duke_University' for tweet N1 (Table 4) is described in https://en.wikipedia.org/wiki/Duke_University.

**Table 4.** Examples of positive changes (in bold), with affected mentions in italics.

| | |
|---|---|
| N1 | LETS GO *DUKE*!! #PinstripeBowl @Yankee Stadium<br>**Duke → Duke_University** |
| N2 | May be the post-season but finally getting to see the *#Hoosiers* play<br>**Hoosiers → Indiana_Hoosiers_football** |
| N3 | *Syracuse* game with my dad at The Garden-we're both alumni #cuse #cusenation #nyc<br>**Syracuse,_Sicily → Syracuse,_New_York** |

*Duke* in N1 is initially linked by TAGME to 'Duke': ruler of a Duchy. Collective linking then changed it to 'Duke_University'. Although this is not perfect, it is an improvement since Duke University is the parent organization of the football team involved. For N2, the final entity for *Hoosier* is correct in the strictest sense. Tweet N3 illustrates geographical effects, where surrounding tweets linked to NYC-related entities drive changes in the initial linking. For example, N3 is about a basketball game involving Syracuse University. Its final linking is a positive change, since an unrelated entity (a location in Italy) has been changed to a parent entity (university's location in NYC).

**Negative Changes.** Table 5 illustrates negative changes. N5's mention *World* is not from a named entity, but has been extracted by TweetNLP. It is impossible to automatically filter out all such mentions, hence linking is still conducted. The final linking in N5 is overly specific and wrong. N5 originates from NYC and surrounding tweets mentioned entities that drive the negative change. For example, mentions of NYC will drive the linking towards 'World_Wrestling_Entertainment' (WWE) since WWE's event had been held in NYC before. For N6, initial linking is to 'Yankee', which discuss usage of the word, including its usage in referring to Americans. The final linking is wrong and refers to an American baseball team.

**Table 5.** Examples of negative changes (in bold), with affected mentions in italics.

| | |
|---|---|
| N5 | *World*'s Most Famous Arena for my sixth sporting event in two weeks...<br>**World → World_Wrestling_Entertainment** |
| N6 | Incredible spread by the @*yankees*. Choice of pork,<br>chicken, hot dogs and burgers. Salad bar<br>**Yankee → New_York_Yankees** |

**Neither.** Table 6 shows two examples where the final linking arguably improves our understanding of the tweet content. N9 is generated during a college football game. After collective linking, its mention *Bowl* is linked to a different series of football game, much better than the initial linking to 'Bowl', a container. N10's mention *WWF* is finally linked to a WWF wrestler, a more related entity than the initial linking to a nature conservation organization. Nonetheless such cases

**Table 6.** Sample changes (bold) for affected mentions (italics) that arguably improve tweet understanding, but are not counted as positive changes.

| | |
|---|---|
| N9 | *Bowl* Games with Famiky #CandyStripes NotPinstripes #PinstripeBowl<br>**Bowl → Super_Bowl** |
| N10 | I Met Former UFC Fighter & amp; *WWF* Wrestler Dan The Beast Severn At The MMA World Expo. Dan Is A...<br>**World_Wide_Fund_for_Nature → Hulk_Hogan** |

do not fall into our discussed scenarios in Sect. 4.1 and can be subjective to assess. Hence we do not count them as positive change.

## 6 Conclusion

Motivated by event and geographical effects, we have proposed a collective entity linking approach for tweets over space and time. In addition, we proposed a comparison-based evaluation strategy, that focuses on the linkage differences between competing entity linking techniques. This reduces manual annotation effort and mitigate challenges such as noisy mention extraction and incomplete KB. Our results show that collective linking over space and time performs much better than local linking techniques that process individual tweets. In extensive experiments, collective linking improves the linking quality of local linking.

**Acknowledgements.** This research is supported by DSO National Laboratories, and the National Research Foundation, Prime Ministers Office, Singapore under its International Research Centres in Singapore Funding Initiative.

We also wish to thank Dan Tasse and Jason Hong of CHIMPS Lab for providing access to the NYC tweets.

## References

1. Atefeh, F., Khreich, W.: A survey of techniques for event detection in Twitter. Comput. Intell. **31**(1), 132–164 (2015)
2. Fang, Y., Chang, M.-W.: Entity linking on microblogs with spatial and temporal signals. TACL **2**, 259–272 (2014)
3. Ferragina, P., Scaiella, U.: Tagme: on-the-fly annotation of short text fragments (by wikipedia entities). In: CIKM (2010)
4. Huang, H., Cao, Y., Huang, X., Ji, H., Lin, C.-Y.: Collective Tweet Wikification based on semi-supervised graph regularization. In: ACL (2014)
5. Ling, X., Singh, S., Weld, D.S.: Design challenges for entity linking. TACL **3**, 315–328 (2015)
6. Liu, X., Li, Y., Wu, H., Zhou, M., Wei, F., Lu, Y.: Entity linking for Tweets. In: ACL (2013)
7. Mazaitis, K., Wang, R.C., Dalvi, B., Cohen, W.W.: A tale of two entity linking and discovery systems. In: TAC (2014)
8. Meij, E., Weerkamp, W., de Rijke, M.: Adding semantics to microblog posts. In: WSDM (2012)

9. Milne, D., Witten, I.: An effective, low-cost measure of semantic relatedness obtained from Wikipedia links. In: AAAI (2008)
10. Owoputi, O., O'Connor, B., Dyer, C., Gimpely, K., Schneider, N., Smith, N.A.: Improved part-of-speech tagging for online conversational text withword clusters. In: NAACL (2013)
11. Ratinov, L., Roth, D., Downey, D., Anderson, M.: Local and global algorithms for disambiguation to Wikipedia. In: ACL (2011)
12. Shen, W., Wang, J., Luo, P., Wang, M.: LIEGE: link entities in web lists with knowledge base. In: KDD (2012)
13. Shen, W., Wang, J., Luo, P., Wang, M.: Linden: linking named entities with knowledge base via semantic knowledge. In: WWW (2012)
14. Shen, W., Wang, J., Luo, P., Wang, M.: Linking named entities in Tweets with knowledge base via user interest modeling. In: KDD (2013)
15. Spitkovsky, V.I., Chang, A.X.: A cross-lingual dictionary for English Wikipedia concepts. In: LREC (2012)