Collective Tweet Wikification based on Semi-supervised Graph Regularization

Hongzhao Huang[1], Yunbo Cao[2], Xiaojiang Huang[2], Heng Ji[1], Chin-Yew Lin[2]
[1]Computer Science Department, Rensselaer Polytechnic Institute, Troy, NY 12180, USA
[2]Microsoft Research Asia, Beijing 100080, P.R.China
{huangh9,jih}@rpi.edu[1],
{yunbo.cao,xiaojih,cyl}@microsoft.com[2]

Abstract

Wikification for tweets aims to automatically identify each concept mention in a tweet and link it to a concept referent in a knowledge base (e.g., Wikipedia). Due to the shortness of a tweet, a collective inference model incorporating global evidence from multiple mentions and concepts is more appropriate than a non-collecitve approach which links each mention at a time. In addition, it is challenging to generate sufficient high quality labeled data for supervised models with low cost. To tackle these challenges, we propose a novel semi-supervised graph regularization model to incorporate both local and global evidence from multiple tweets through three fine-grained relations. In order to identify semantically-related mentions for collective inference, we detect meta path-based semantic relations through social networks. Compared to the state-of-the-art supervised model trained from 100% labeled data, our proposed approach achieves comparable performance with 31% labeled data and obtains 5% absolute F1 gain with 50% labeled data.

1 Introduction

With millions of tweets posted daily, Twitter enables both individuals and organizations to disseminate information, from current affairs to breaking news in a timely fashion. In this work, we study the wikification (Disambiguation to Wikipedia) task (Mihalcea and Csomai, 2007) for tweets, which aims to automatically identify each *concept mention* in a tweet, and link it to a *concept* referent in a knowledge base (KB) (e.g., Wikipedia). For example, as shown in Figure 1, *Hawks* is an identified mention, and its correct referent concept in Wikipedia is *Atlanta Hawks*. An end-to-end wikification system needs to solve two sub-problems: (i) concept mention detection, (ii) concept mention disambiguation.

Wikification is a particularly useful task for short messages such as tweets because it allows a reader to easily grasp the related topics and enriched information from the KB. From a system-to-system perspective, wikification has demonstrated its usefulness in a variety of applications, including coreference resolution (Ratinov and Roth, 2012) and classification (Vitale et al., 2012).

Sufficient labeled data is crucial for supervised models. However, manual wikification annotation for short documents is challenging and time-consuming (Cassidy et al., 2012). The challenges are: (i) *unlinkability*, a valid concept may not exist in the KB. (ii) *ambiguity*, it is impossible to determine the correct concept due to the dearth of information within a single tweet or multiple correct answer. For instance, it would be difficult to determine the correct referent concept for "*Gators*" in t_1 in Figure 1. Linking "*UCONN*" in t_3 to *University of Connecticut* may also be acceptable since *Connecticut Huskies* is the athletic team of the university. (iii) *prominence*, it is challenging to select a set of linkable mentions that are important and relevant. It is not tricky to select "*Fans*", "*slump*", and "*Hawks*" as linkable mentions, but other mentions such as "*stay up*" and "*stay positive*" are not prominent. Therefore, it is challenging to create sufficient high quality labeled tweets for supervised models and worth considering semi-supervised learning with the exploration of unlabeled data.

Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, pages 380–390,
Baltimore, Maryland, USA, June 23-25 2014.

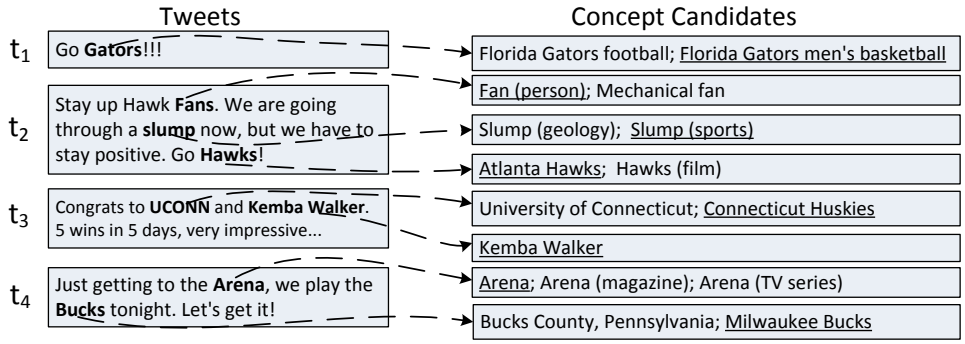

Figure 1: An illustration of Wikification Task for Tweets. Concept mentions detected in tweets are marked as bold, and correctly linked concepts are underlined. The concept candidates are ranked by their prior popularity which will be explained in section 4.1, and only top 2 ranked concepts are listed.

However, when selecting semi-supervised learning frameworks, we noticed another unique challenge that tweets pose to wikification due to their informal writing style, shortness and noisiness. The context of a single tweet usually cannot provide enough information for prominent mention detection and similarity computing for disambiguation. Therefore, a collective inference model over multiple tweets in the semi-supervised setting is desirable. For instance, the four tweets in Figure 1 are posted by the same author within a short time period. If we perform collective inference over them we can reliably link ambiguous mentions such as "*Gators*", "*Hawks*", and "*Bucks*" to basketball teams instead of other concepts such as the county *Bucks County*.

In order to address these unique challenges for wikification for the short tweets, we employ graph-based semi-supervised learning algorithms (Zhu et al., 2003; Smola and Kondor, 2003; Blum et al., 2004; Zhou et al., 2004; Talukdar and Crammer, 2009) for collective inference by exploiting the manifold (cluster) structure in both unlabeled and labeled data. These approaches normally assume label smoothness over a defined graph, where the nodes represent a set of labeled and unlabeled instances, and the weighted edges reflect the closeness of each pair of instances. In order to construct a semantic-rich graph capturing the similarity between mentions and concepts for the model, we introduce three novel fine-grained relations based on a set of local features, social networks and meta paths.

The main contributions of this paper are summarized as follows:

- To the best of our knowledge, this is the first effort to explore graph-based semi-supervised learning algorithms for the wikification task.
- We propose a novel semi-supervised graph regularization model performing collective inference for joint mention detection and disambiguation. Our approach takes advantage of three proposed principles to incorporate both local and global evidence from multiple tweets.
- We propose a meta path-based unified framework to detect both explicitly and implicitly relevant mentions.

2 Preliminaries

Concept and Concept Mention We define a *concept* c as a Wikipedia article (e.g., *Atlanta Hawks*), and a *concept mention* m as an n-gram from a specific tweet. Each concept has a set of textual representation fields (Meij et al., 2012), including *title* (the title of the article), *sentence* (the first sentence of the article), *paragraph* (the first paragraph of the article), *content* (the entire content of the article), and *anchor* (the set of all anchor texts with incoming links to the article).

Wikipedia Lexicon Construction We first construct an offline lexicon with each entry as $\langle m, \{c_1, ..., c_k\}\rangle$, where $\{c_1, ..., c_k\}$ is the set of possible referent concepts for the mention m. Following the previous work (Bunescu, 2006; Cucerzan, 2007; Hachey et al., 2013), we extract the possible mentions for a given concept c using the following resources: the title of c; the aliases appearing in the introduction and infoboxes of c (e.g., *The Evergreen State* is an alias of *Washington state*); the titles of pages redirecting to c (e.g., *State of Washington* is a redirecting page of *Washington (state)*); the titles of the disambigua-

tion pages containing c; and all the anchor texts appearing in at least 5 pages with hyperlinks to c (e.g., *WA* is a mention for the concept *Washington (state)* in the text "*401 5th Ave N [[Seattle]], [[Washington (state)—WA]] 98109 USA*". We also propose three heuristic rules to extract mentions (i.e., different combinations of the family name and given name for a person, the headquarters of an organization, and the city name for a sports team).

Concept Mention Extraction Based on the constructed lexicon, we then consider all n-grams of size $\leq n$ (n=7 in this paper) as concept mention candidates if their entries in the lexicon are not empty. We first segment @usernames and #hashtags into regular tokens (e.g., @amandapalmer is segmented as *amanda palmer* and #WorldWaterDay is split as *World Water Day*) using the approach proposed by (Wang et al., 2011). Segmentation assists finding concept candidates for these non-regular mentions.

3 Principles and Approach Overview

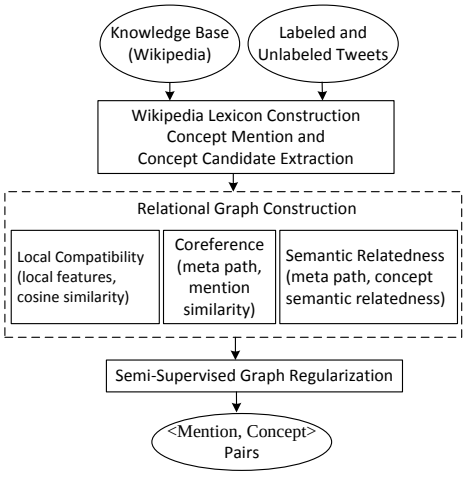

Figure 2: Approach Overview.

3.1 Principles

A single tweet may not provide enough evidence to identify prominent mentions and infer their correct referent concepts due to the lack of contextual information. To tackle this problem, we propose to incorporate global evidence from multiple tweets and performing collective inference for both mention identification and disambiguation. We first introduce the following three principles that our approach relies on.

Principle 1 (Local compatibility): *Two pairs of $\langle m, c\rangle$ with strong local compatibility tend to have similar labels.* Mentions and their correct referent concepts usually tend to share a set of characteristics such as string similarity between m and c (e.g., $\langle$*Chicago*, *Chicago*$\rangle$ and $\langle$*Facebook*, *Facebook*$\rangle$). We define the *local compatibility* to model such set of characteristics.

Principle 2 (Coreference): *Two coreferential mentions should be linked to the same concept.* For example, if we know "*nc*" and "*North Carolina*" are coreferential, then they should both be linked to *North Carolina*.

Principle 3 (Semantic Relatedness): *Two highly semantically-related mentions are more likely to be linked to two highly semantically-related concepts.* For instance, when "*Sweet 16*" and "*Hawks*" often appear together within relevant contexts, they can be reliably linked to two baseketball-related concepts *NCAA Men's Division I Basketball Championship* and *Atlanta Hawks*, respectively.

3.2 Approach Overview

Given a set of tweets $\langle t_1, ..., t_{|T|}\rangle$, our system first generates a set of candidate concept mentions, and then extracts a set of candidate concept referents for each mention based on the Wikipedia lexicon. Given a pair of mention and its candidate referent concept $\langle m, c\rangle$, the remaining task of wikification is to assign either a positive label if m should be selected as a prominently linkable mention and c is its correct referent concept, or otherwise a negative label. The label assignment is obtained by our semi-supervised graph regularization framework based on a relational graph, which is constructed from *local compatibility*, *coreference*, and *semantic relatedness* relations. The overview of our approach is as illustrated in Figure 2.

4 Relational Graph Construction

We first construct the relational graph $G = \langle V, E\rangle$, where $V = \{v_1, ..., v_n\}$ is a set of nodes and $E = \{e_1, ..., e_m\}$ is a set of edges. Each $v_i = \langle m_i, c_i\rangle$ represents a tuple of mention m_i and its referent concept candidate c_i. An edge is added between two nodes v_i and v_j if there is a proposed relation based on the three principles described in section 3.1.

4.1 Local Compatibility

We first compute local compatibility (Principle 1) by considering a set of novel local features to cap-

ture the importance and relevance of a mention m to a tweet t, as well as the correctness of its linkage to a concept c. We have designed a number of features which are similar to those commonly used in wikification and entity linking work (Meij et al., 2012; Guo et al., 2013; Mihalcea and Csomai, 2007).

Mention Features We define the following features based on information from mentions.

- $IDF_f(m) = \log(\frac{|C|}{df(m)})$, where $|C|$ is the total number of concepts in Wikipedia and $df(m)$ is the total number of concepts in which m occurs, and f indicates the field property, including *title*, *content*, and *anchor*.
- $Keyphraseness(m) = \frac{|C_a(m)|}{df(m)}$ to measure how likely m is used as an anchor in Wikipedia, where $C_a(m)$ is the set of concepts where m appears as an anchor.
- $LinkProb(m) = \frac{\sum_{c \in C_a(m)} count(m,c)}{\sum_{c \in C} count(m,c)}$, where $count(m, c)$ indicates the number of occurrence of m in c.
- $SNIL(m)$ and $SNCL(m)$ to count the number of concepts that are equal to or contain a sub-n-gram of m, respectively (Meij et al., 2012).

Concept Features The concept features are solely based on Wikipedia, including the number of incoming and outgoing links for c, and the number of words and characters in c.

Mention + Concept Features This set of features considers information from both mentions and concepts:

- **prior popularity** $prior(m, c) = \frac{count(m,c)}{\sum_{c'} count(m,c')}$, where $count(m, c)$ measures the frequency of the anchor links from m to c in Wikipedia.
- $TF_f(m, c) = \frac{count_f(m,c)}{|f|}$ to measure the relative frequency of m in each field representation f of c, normalized by the length of f. The fields include *title*, *sentence*, *paragraph*, *content* and *anchor*.
- $NCT(m, c)$, $TCN(m, c)$, and $TEN(m, c)$ to measure whether m contains the title of c, whether the title of c contains m, and whether m equals to the title of c, respectively.

Context Features This set of features include (i) Context Capitalization features, which indicate whether the current mention, the token before, and the token after are capitalized. (ii) tf-idf based features, which include the dot product of two word vectors v_c and v_t, and the average tf-idf value of common items in v_c and v_t, where v_c and v_t are the top 100 tf-idf word vectors in c and t.

Local Compatibility Computation For each node $v_i = \langle m_i, c_i \rangle$, we collect its local features as a feature vector $F_i = \langle f_1, f_2, ..., f_d \rangle$. To avoid features with large numerical values that dominate other features, the value of each feature is re-scaled using feature standardization approach. The cosine similarity is then adopted to compute the local compatibility of two nodes and construct a k *nearest neighbor* (kNN) graph, where each node is connected to its k nearest neighboring nodes. We compute the weight matrix that represents the local compatibility relation as:

$$W_{ij}^{loc} = \begin{cases} cosine(F_i, F_j) & j \in kNN(i) \\ 0 & \text{Otherwise} \end{cases}$$

4.2 Meta Path

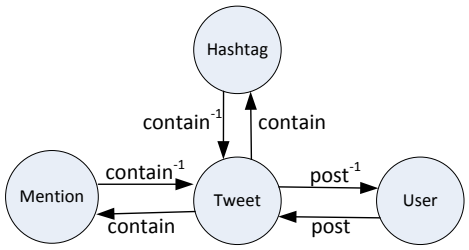

Figure 3: Schema of the Twitter network.

In this subsection, we introduce the concept meta path which will be used to detect coreference (section 4.3) and semantic relatedness relations (section 4.4).

A meta-path is a path defined over a network and composed of a sequence of relations between different object types (Sun et al., 2011b). In our experimental setting, we can construct a natural Twitter network summarized by the network schema in Figure 3. The network contains four types of objects: *Mention* (M), *Tweet* (T), *User* (U), and *Hashtag* (H). Tweets and mentions are connected by links "contain" and "contained by" (denoted as "$contain^{-1}$"); and other linked relationships can be described similarly.

We then define the following five types of meta paths to connect two mentions as:

- "M - T - M",
- "M - T - U - T - M",
- "M - T - H - T - M",
- "M - T - U - T - M - T - H - T - M",
- "M - T - H - T - M - T - U - T - M".

Each meta path represents one particular semantic relation. For instance, the first three paths are *basic* ones expressing the explicit relations that two mentions are from the same tweet, posted by the same user, and share the same #hashtag, respectively. The last two paths are *concatenated* ones which are constructed by concatenating the first three simple paths to express the implicit relations that two mentions co-occur with a third mention sharing either the same authorship or #hashtag. Such complicated paths can be exploited to detect more semantically-related mentions from wider contexts. For example, the relational link between "*narita airport*" and "*Japan*" would be missed without using the path "*narita airport* - t_1 - u_1 - t_2 - *american* - t_3 - h_1 - t_4 - *Japan*" since they don't directly share any authorships or #hashtags.

4.3 Coreference

A coreference relation (Principle 2) usually occurs across multiple tweets due to the highly redundant information in Twitter. To ensure high precision, we propose a simple yet effective approach utilizing the rich social network relations in Twitter.

We consider two mentions m_i and m_j coreferential if m_i and m_j share the same surface form or one is an abbreviation of the other, and at least one meta path exists between m_i and m_j. Then we define the weight matrix representing the coreferential relation as:

$$W_{ij}^{coref} = \begin{cases} 1.0 & \text{if } m_i \text{ and } m_j \text{ are coreferential,} \\ & \text{and } c_i = c_j \\ 0 & \text{Otherwise} \end{cases}$$

4.4 Semantic Relatedness

Ensuring topical coherence (Principle 3) has been beneficial for wikification on formal texts (e.g., News) by linking a set of semantically-related mentions to a set of semantically-related concepts simultaneously (Han et al., 2011; Ratinov et al., 2011; Cheng and Roth, 2013). However, the shortness of a single tweet means that it may not provide enough topical clues. Therefore, it is important to extend this evidence to capture semantic relatedness information from multiple tweets.

We define the semantic relatedness score between two mentions as $SR(m_i, m_j) = 1.0$ if at least one meta path exists between m_i and m_j, otherwise $SR(m_i, m_j) = 0$. In order to compute the semantic relatedness of two concepts c_i and c_j, we adopt the approach proposed by (Milne and Witten, 2008a):

$$SR(c_i, c_j) = 1 - \frac{\log \max(|C_i|, |C_j|) - \log |C_i \cap C_j|}{\log(|C|) - \log \min(|C_i|, |C_j|)},$$

where $|C|$ is the total number of concepts in Wikipedia, and C_i and C_j are the set of concepts that have links to c_i and c_j, respectively.

Then we compute a weight matrix representing the semantic relatedness relation as:

$$W_{ij}^{rel} = \begin{cases} SR(N_i, N_j) & \text{if } SR(N_i, N_j) \geq \delta \\ 0 & \text{Otherwise} \end{cases}$$

where $SR(N_i, N_j) = SR(m_i, m_j) \times SR(c_i, c_j)$ and $\delta = 0.3$, which is optimized from a development set.

4.5 The Combined Relational Graph

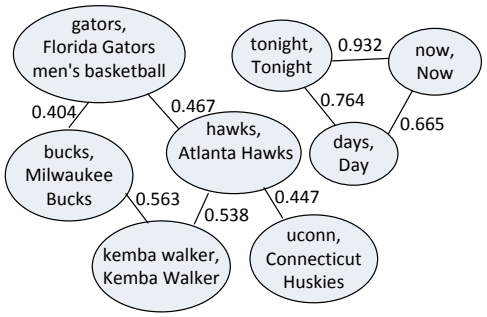

Figure 4: A example of the relational graph constructed for the example tweets in Figure 1. Each node represents a pair of $\langle m, c \rangle$, separated by a comma. The edge weight is obtained from the linear combination of the weights of the three proposed relations. Not all mentions are included due to the space limitations.

Based on the above three weight matrices W^{loc}, W^{coref}, and W^{rel}, we first obtain their corresponding transition matrices P^{loc}, P^{coref}, and P^{rel}, respectively. The entry P_{ij} of the transition matrix P for a weight matrix W is computed as $P_{ij} = \frac{W_{ij}}{\sum_k W_{ik}}$ such that $\sum_k P_{ik} = 1$. Then we obtain the combined graph G with weight matrix W, where $W_{ij} = \alpha P_{ij}^{loc} + \beta P_{ij}^{coref} + \gamma P_{ij}^{rel}$. α, β, and γ are three coefficients between 0 and 1 with the constraint that $\alpha + \beta + \gamma = 1$. They control the contributions of these three relations in our semi-supervised graph regularization model. We choose transition matrix to avoid the domination of one relation over others. An example graph of G is shown in Figure 4. Compared to the *referent graph* which considers each mention or concept as a node in previous graph-based re-ranking approaches (Han et al., 2011; Shen et al., 2013), our

novel graph representation has two advantages: (i) It can easily incorporate more features related to both mentions and concepts. (ii) It is more appropriate for our graph-based semi-supervised model since it is difficult to assign labels to a pair of mention and concept in the referent graph.

5 Semi-supervised Graph Regularization

Given the constructed relational graph with the weighted matrix W and the label vector Y of all nodes, we assume the first l nodes are labeled as Y_l and the remaining u nodes ($u = n - l$) are initialized with labels Y_u^0. Then our goal is to refine Y_u^0 and obtain the final label vector Y_u.

Intuitively, if two nodes are strongly connected, they tend to hold the same label. We propose a novel semi-supervised graph regularization framework based on the graph-based semi-supervised learning algorithm (Zhu et al., 2003):

$$\mathcal{Q}(\mathcal{Y}) = \mu \sum_{i=l+1}^{n} (y_i - y_i^0)^2 + \frac{1}{2} \sum_{i,j} W_{ij}(y_i - y_j)^2.$$

The first term is a loss function that incorporates the initial labels of unlabeled examples into the model. In our method, we adopt *prior popularity* (section 4.1) to initialize the labels of the unlabeled examples. The second term is a regularizer that smoothes the refined labels over the constructed graph. μ is a regularization parameter that controls the trade-off between initial labels and the consistency of labels on the graph. The goal of the proposed framework is to ensure that the refined labels of unlabeled nodes are consistent with their strongly connected nodes, as well as not too far away from their initial labels.

The above optimization problem can be solved directly since $\mathcal{Q}(\mathcal{Y})$ is convex (Zhu et al., 2003; Zhou et al., 2004). Let I be an identity matrix and D_W be a diagonal matrix with entries $D_{ii} = \sum_j W_{ij}$. We can split the weighted matrix W into four blocks as $W = \begin{bmatrix} W_{ll} & W_{lu} \\ W_{ul} & W_{uu} \end{bmatrix}$, where W_{mn} is an $m \times n$ matrix. D_w is split similarly. We assume that the vector of the labeled examples Y_l is fixed, so we only need to infer the refined label vector of the unlabeled examples Y_u. In order to minimize $\mathcal{Q}(\mathcal{Y})$, we need to find Y_u^* such that

$$\left.\frac{\partial Q}{\partial Y_u}\right|_{Y_u = Y_u^*} = (D_{uu} + \mu I_{uu})Y_u - W_{uu}Y_u - W_{ul}Y_l - \mu Y_u^0 = 0.$$

Therefore, a closed form solution can be derived as $Y_u^* = (D_{uu} + \mu I_{uu} - W_{uu})^{-1}(W_{ul}Y_l + \mu Y_u^0)$.

However, for practical application to a large-scale data set, an iterative solution would be more efficient to solve the optimization problem. Let Y_u^t be the refined labels after the t^{th} iteration, the iterative solution can be derived as:

$$Y_u^{t+1} = (D_{uu} + \mu I_{uu})^{-1}(W_{uu}Y_u^t + W_{ul}Y_l + \mu Y_u^0).$$

The iterative solution is more efficient since $(D_{uu} + \mu I_{uu})$ is a diagonal matrix and its inverse is very easy to compute.

6 Experiments

In this section we compare our approach with state-of-the-art methods as shown in Table 1.

6.1 Data and Scoring Metric

For our experiments we use a public data set (Meij et al., 2012) including 502 tweets posted by 28 verified users. The data set was annotated by two annotators. We randomly sample 102 tweets for development and the remaining for evaluation. We use a Wikipedia dump on May 3, 2013 as our knowledge base, which includes 30 million pages. For computational efficiency, we also filter some mention candidates by applying the preprocessing approach proposed in (Ferragina and Scaiella, 2010), and remove all the concepts with prior popularity less than 2% from an mention's concept set for each mention, similar to (Guo et al., 2013).

A mention and concept pair $\langle m, c \rangle$ is judged as correct if and only if m is linkable and c is the correct referent concept for m. To evaluate the performance of a wikification system, we use the standard precision, recall and F1 measures.

6.2 Experimental Results

The overall performance of various approaches is shown in Table 2. The results of the supervised method proposed by (Meij et al., 2012) are obtained from 5-fold cross validation. For our semi-supervised setting, we experimentally sample 200 tweets for training and use the remaining set as unlabeled and testing sets. In our semi-supervised regularization model, the matrix W^{loc} is constructed by a kNN graph ($k = 20$). The regularization parameter μ is empirically set to 0.1, and the coefficients α, β, and γ are learnt from the development set by considering all the combina-

Methods	Descriptions
TagMe	The same approach that is described in (Ferragina and Scaiella, 2010), which aims to annotate short texts based on prior popularity and semantic relatedness of concepts. It is basically an unsupervised approach, except that it needs a development set to tune the probability threshold for linkable mentions.
Meij	A state-of-the-art system described in (Meij et al., 2012), which is a supervised approach based on the random forest model. It performs mention detection and disambiguation jointly, and it is trained from 400 labeled tweets.
SSRegu$_1$	Our proposed model based on Principle 1, using 200 labeled tweets.
SSRegu$_{12}$	Our proposed model based on Principle 1 and 2, using 200 labeled tweets.
SSRegu$_{13}$	Our proposed model based on Principle 1 and 3, using 200 labeled tweets.
SSRegu$_{123}$	Our proposed full model based on Principle 1, 2 and 3, using 200 labeled tweets.

Table 1: Description of Methods.

Methods	Precision	Recall	F1
TagMe	0.329	0.423	0.370
Meij	0.393	0.598	0.475
SSRegu$_1$	0.538	0.435	0.481
SSRegu$_{12}$	0.638	0.438	0.520
SSRegu$_{13}$	0.541	0.457	0.495
SSRegu$_{123}$	0.650	0.441	**0.525**

Table 2: Overall Performance.

tions of values from 0 to 1 at 0.1 intervals[1]. In order to randomize the experiments and make the comparison fair, we conduct 20 test runs for each method and report the average scores across the 20 trials.

The relatively low performance of the baseline system *TagMe* demonstrates that only relying on prior popularity and topical information within a single tweet is not enough for an end-to-end wikification system for the short tweets. As an example, it is difficult to obtain topical clues in order to link the mention "*Clinton*" to *Hillary Rodham Clinton* by relying on the single tweet "*wolfblitzercnn: Behind the scenes on Clinton's Mideast trip #cnn*". Therefore, the system mistakenly links it to the most popular concept *Bill Clinton*.

In comparision with the supervised baseline proposed by (Meij et al., 2012), our model *SSRegu*$_1$ relying on *local compatibility* already achieves comparable performance with 50% of labeled data. This is because that our model performs collective inference by making use of the manifold (cluster) structure of both labeled and unlabeled data, and that the local compatibility relation is detected with high precision[2] (89.4%). For example, the following three pairs of mentions and concepts ⟨*pelosi, Nancy Pelosi*⟩, ⟨*obama, Barack Obama*⟩, and ⟨*gaddafi, Muammar Gaddafi*⟩ have strong local compatibility with each other since they share many similar characteristics captured by the local features such as string similarity between the mention and the concept. Suppose the first pair is labeled, then its positive label will be propagated to other unlabeled nodes through the local compatibility relation, and correctly predict the labels of other nodes.

Incorporating coreferential or semantic relatedness relation into *SSRegu*$_1$ provides further gains, demonstrating the effectiveness of these two relations. For instance, "*wh*" is correctly linked to *White House* by incorporating evidence from its coreferential mention "*white house*". The coreferential relation (Principle 2) is demonstrated to be more beneficial than the semantic relatedness relation (Principle 3) because the former is detected with much higher precision (99.7%) than the latter (65.4%).

Our full model *SSRegu*$_{123}$ achieves significant improvement over the supervised baseline (5% absolute F1 gain with 95.0% confidence level by the Wilcoxon Matched-Pairs Signed-Ranks Test), showing that incorporating global evidence from multiple tweets with fine-grained relations is beneficial. For instance, the supervised baseline fails to link "*UCONN*" and "*Bucks*" in our examples to *Connecticut Huskies* and *Milwaukee Bucks*, respectively. Our full model corrects these two wrong links by propagating evidence through the semantic links as shown in Figure 4 to obtain mutual ranking improvement. The best performance of our full model also illustrates that the three relations complement each other.

We also study the disambiguation performance for the annotated mentions, as shown in Table 3. We can easily see that our proposed approach using 50% labeled data achieves similar performance with the state-of-the-art supervised model with 100% labeled data. When the mentions are given, the unpervised approach *TagMe* has already

[1]These three coefficients are slightly different with different training data, a sample of them is: $\alpha = 0.4$, $\beta = 0.5$, and $\gamma = 0.1$

[2]Here we define precision as the percentage of links that holds the same label.

Methods	TagMe	Meij	SSRegu_{123}
Accuracy	0.710	0.779	0.772

Table 3: Disambiguation Performance.

Methods	Precision	Recall	F1
$\textbf{SSRegu}_{12}$	0.644	0.423	0.510
$\textbf{SSRegu}_{13}$	0.543	0.441	0.486
$\textbf{SSRegu}_{123}$	0.657	0.419	0.512

Table 4: The Performance of Systems Without Using Concatenated Meta Paths.

achieved reasonable performance. In fact, mention detection actually is the performance bottleneck of a tweet wikification system (Guo et al., 2013). Our system performs better in identifying the prominent mention.

6.3 Effect of Concatenated Meta Paths

In this work, we propose a unified framework utilizing meta path-based semantic relations to explore richer relevant context. Beyond the *basic* meta paths, we introduce *concatenated* ones by concatenating the basic ones. The performance of the system without using the concatenated meta paths is shown in Table 4. In comparison with the system based on all defined meta paths, we can clearly see that the systems using concatenated ones outperform those relying on the simple ones. This is because the concatenated meta paths can incorporate more relevant information with implicit relations into the models by increasing 1.6% coreference links and 9.3% semantic relatedness links. For example, the mention "*narita airport*" is correctly disambiguated to the concept "*Narita International Airport*" with higher confidence since its semantic relatedness relation with "*Japan*" is detected with the concatenated meta path as described in section 4.2.

6.4 Effect of Labeled Data Size

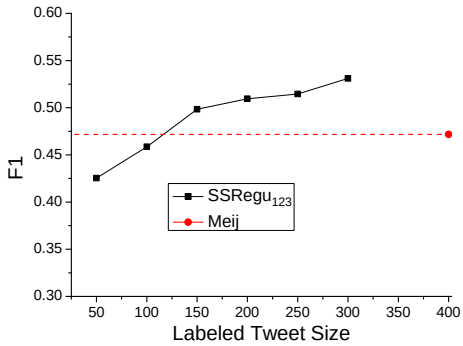

Figure 5: The effect of Labeled Tweet Size.

In previous experiments, we experimentally set the number of labeled tweets to be 200 for overall performance comparision with the baselines. In this subsection, we study the effect of labeled data size on our full model. We randomly sample 100 tweets as testing data, and randomly select 50, 100, 150, 200, 250, and 300 tweets as labeled data. 20 test runs are conducted and the average results are reported across the 20 trials, as shown in Figure 5. We find that as the size of the labeled data increases, our proposed model achieves better performance. It is encouraging to see that our approach, with only $\mathbf{31.3}\%$ labeled tweets (125 out of 400), already achieves a performance that is comparable to the state-of-the-art supervised model trained from 100% labeled tweets.

6.5 Parameter Analysis

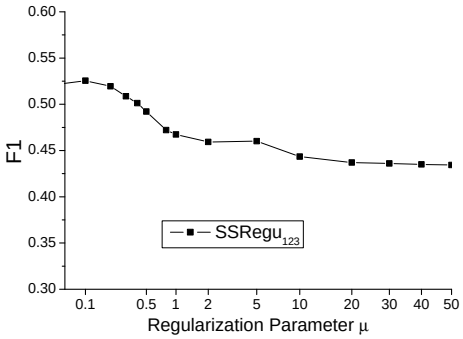

Figure 6: The effect of parameter μ.

In previous experiments, we empirically set the parameter $\mu = 0.1$. μ is the regularization parameter that controls the trade-off between initial labels and the consistency of labels on the graph. When μ increases, the model tends to trust more in the initial labels. Figure 6 shows the performance of our models by varying μ from 0.02 to 50. We can easily see that the system performce is stable when $\mu < 0.4$. However, when $\mu \geq 0.4$, the system performance dramatically decreases, showing that prior popularity is not enough for an end-to-end wikification system.

7 Related Work

The task of linking concept mentions to a knowledge base has received increased attentions over the past several years, from the linking of concept mentions in a single text (Mihalcea and Csomai, 2007; Milne and Witten, 2008b; Milne and Witten, 2008a; Kulkarni et al., 2009; He et al., 2011; Ratinov et al., 2011; Cassidy et al., 2012; Cheng and Roth, 2013), to the linking of a cluster of corefer-

ent named entity mentions spread throughout different documents (Entity Linking) (McNamee and Dang, 2009; Ji et al., 2010; Zhang et al., 2010; Ji et al., 2011; Zhang et al., 2011; Han and Sun, 2011; Han et al., 2011; Gottipati and Jiang, 2011; He et al., 2013; Li et al., 2013; Guo et al., 2013; Shen et al., 2013; Liu et al., 2013).

A significant portion of recent work considers the two sub-problems *mention detection* and *mention disambiguation* separately and focus on the latter by first defining candidate concepts for a deemed mention based on anchor links. Mention disambiguation is then formulated as a ranking problem, either by resolving one mention at each time (non-collective approaches), or by disambiguating a set of relevant mentions simultaneously (collective approaches). Non-collective methods usually rely on prior popularity and context similarity with supervised models (Mihalcea and Csomai, 2007; Milne and Witten, 2008b; Han and Sun, 2011), while collective approaches further leverage the global coherence between concepts normally through supervised or graph-based re-ranking models (Cucerzan, 2007; Milne and Witten, 2008b; Han and Zhao, 2009; Kulkarni et al., 2009; Pennacchiotti and Pantel, 2009; Ferragina and Scaiella, 2010; Fernandez et al., 2010; Radford et al., 2010; Cucerzan, 2011; Guo et al., 2011; Han and Sun, 2011; Han et al., 2011; Ratinov et al., 2011; Chen and Ji, 2011; Kozareva et al., 2011; Cassidy et al., 2012; Shen et al., 2013; Liu et al., 2013). Especially note that when applying the collective methods to short messages from social media, evidence from other messages usually needs to be considered (Cassidy et al., 2012; Shen et al., 2013; Liu et al., 2013). Our method is a collective approach with the following novel advancements: (i) A novel graph representation with fine-grained relations, (ii) A unified framework based on meta paths to explore richer relevant context, (iii) Joint identification and linking of mentions under semi-supervised setting.

Two most similar methods to ours were proposed by (Meij et al., 2012; Guo et al., 2013) by performing joint detection and disambiguation of mentions. (Meij et al., 2012) studied several supervised machine learning models, but without considering any global evidence either from a single tweet or other relevant tweets. (Guo et al., 2013) explored second order entity-to-entity relations but did not incorporate evidence from multiple tweets.

This work is also related to graph-based semi-supervised learning (Zhu et al., 2003; Smola and Kondor, 2003; Zhou et al., 2004; Talukdar and Crammer, 2009), which has been successfully applied in many Natural Language Processing tasks (Niu et al., 2005; Chen et al., 2006). We introduce a novel graph that incorporates three fine-grained relations. Our work is further related to meta path-based heterogeneous information network analysis (Sun et al., 2011b; Sun et al., 2011a; Kong et al., 2012; Huang et al., 2013), which has demonstrated advantages over homogeneous information network analysis without differentiating object types and relational links.

8 Conclusions

We have introduced a novel semi-supervised graph regularization framework for wikification to simultaneously tackle the unique challenges of annotation and information shortage in short tweets. To the best of our knowledge, this is the first work to explore the semi-supervised collective inference model to jointly perform mention detection and disambiguation. By studying three novel fine-grained relations, detecting semantically-related information with semantic meta paths, and exploiting the data manifolds in both unlabeled and labeled data for collective inference, our work can dramatically save annotation cost and achieve better performance, thus shed light on the challenging wikification task for tweets.

Acknowledgments

This work was supported by the U.S. Army Research Laboratory under Cooperative Agreement No. W911NF-09-2-0053 (NS-CTA), U.S. NSF CAREER Award under Grant IIS-0953149, U.S. DARPA Award No. FA8750-13-2-0041 in the Deep Exploration and Filtering of Text (DEFT) Program, IBM Faculty Award, Google Research Award and RPI faculty start-up grant. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation here on.

References

A. Blum, J. Lafferty, M. Rwebangira, and R. Reddy. 2004. Semi-supervised learning using randomized mincuts. In *Proceedings of the Twenty-first International Conference on Machine Learning*, ICML '04.

Razvan Bunescu. 2006. Using encyclopedic knowledge for named entity disambiguation. In *EACL*, pages 9–16.

T. Cassidy, H. Ji, L. Ratinov, A. Zubiaga, and H. Huang. 2012. Analysis and enhancement of wikification for microblogs with context expansion. In *Proceedings of COLING 2012*.

Z. Chen and H. Ji. 2011. Collaborative ranking: A case study on entity linking. In *Proc. EMNLP2011*.

J. Chen, D. Ji, C Tan, and Z. Niu. 2006. Relation extraction using label propagation based semi-supervised learning. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*.

X. Cheng and D. Roth. 2013. Relational inference for wikification. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*.

Silviu Cucerzan. 2007. Large-scale named entity disambiguation based on wikipedia data. In *EMNLP-CoNLL 2007*.

S. Cucerzan. 2011. Tac entity linking by performing full-document entity extraction and disambiguation. In *Proc. TAC 2011 Workshop*.

N. Fernandez, J. A. Fisteus, L. Sanchez, and E. Martin. 2010. Webtlab: A cooccurence-based approach to kbp 2010 entity-linking task. In *Proc. TAC 2010 Workshop*.

P. Ferragina and U. Scaiella. 2010. Tagme: on-the-fly annotation of short text fragments (by wikipedia entities). In *Proceedings of the 19th ACM international conference on Information and knowledge management*, CIKM '10.

S. Gottipati and J. Jiang. 2011. Linking entities to a knowledge base with query expansion. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*.

Y. Guo, W. Che, T. Liu, and S. Li. 2011. A graph-based method for entity linking. In *Proc. IJCNLP2011*.

S. Guo, M. Chang, and E. Kiciman. 2013. To link or not to link? a study on end-to-end tweet entity linking. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

B. Hachey, W. Radford, J. Nothman, M. Honnibal, and J. Curran. 2013. Evaluating entity linking with wikipedia. *Artif. Intell.*

X. Han and L. Sun. 2011. A generative entity-mention model for linking entities with knowledge base. In *Proc. ACL2011*.

X. Han and J. Zhao. 2009. Named entity disambiguation by leveraging wikipedia semantic knowledge. In *Proceedings of the 18th ACM conference on Information and knowledge management*, CIKM 2009.

X. Han, L. Sun, and J. Zhao. 2011. Collective entity linking in web text: A graph-based method. In *Proc. SIGIR2011*.

J. He, M. de Rijke, M. Sevenster, R. van Ommering, and Y. Qian. 2011. Generating links to background knowledge: A case study using narrative radiology reports. In *Proceedings of the 20th ACM international conference on Information and knowledge management*. ACM.

Z. He, S. Liu, Y. Song, M. Li, M. Zhou, and H. Wang. 2013. Efficient collective entity linking with stacking. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*.

H. Huang, Z. Wen, D. Yu, H. Ji, Y. Sun, J. Han, and H. Li. 2013. Resolving entity morphs in censored data. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

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) 2010*.

H. Ji, R. Grishman, and H.T. Dang. 2011. Overview of the tac 2011 knowledge base population track. In *Text Analysis Conference (TAC) 2011*.

X. Kong, P. Yu, Y. Ding, and J. Wild. 2012. Meta path-based collective classification in heterogeneous information networks. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, CIKM '12.

Z. Kozareva, K. Voevodski, and S. Teng. 2011. Class label enhancement via related instances. In *Proc. EMNLP2011*.

S. Kulkarni, A. Singh, G. Ramakrishnan, and S. Chakrabarti. 2009. Collective annotation of wikipedia entities in web text. In *KDD*.

Y. Li, C. Wang, F. Han, J. Han, D. Roth, and X. Yan. 2013. Mining evidences for named entity disambiguation. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '13.

X. Liu, Y. Li, H. Wu, M. Zhou, F. Wei, and Y. Lu. 2013. Entity linking for tweets. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

P. McNamee and H.T. Dang. 2009. Overview of the tac 2009 knowledge base population track. In *Text Analysis Conference (TAC) 2009*.

E. Meij, W. Weerkamp, and M. de Rijke. 2012. Adding semantics to microblog posts. In *Proceedings of the fifth ACM international conference on Web search and data mining*, WSDM '12.

R. Mihalcea and A. Csomai. 2007. Wikify!: linking documents to encyclopedic knowledge. In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, CIKM '07.

D. Milne and I.H. Witten. 2008a. Learning to link with wikipedia. In *An effective, low-cost measure of semantic relatedness obtained from wikipedia links*. the Wikipedia and AI Workshop of AAAI.

D. Milne and I.H. Witten. 2008b. Learning to link with wikipedia. In *Proceeding of the 17th ACM conference on Information and knowledge management*, pages 509–518. ACM.

Z. Niu, D. Ji, and C. Tan. 2005. Word sense disambiguation using label propagation based semi-supervised learning. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*.

M. Pennacchiotti and P. Pantel. 2009. Entity extraction via ensemble semantics. In *Proc. EMNLP2009*.

W. Radford, B. Hachey, J. Nothman, M. Honnibal, and J. R. Curran. 2010. Cmcrc at tac10: Document-level entity linking with graph-based re-ranking. In *Proc. TAC 2010 Workshop*.

L. Ratinov and D. Roth. 2012. Learning-based multi-sieve co-reference resolution with knowledge. In *EMNLP*.

L. Ratinov, D. Roth, D. Downey, and M. Anderson. 2011. Local and global algorithms for disambiguation to wikipedia. In *Proc. of the Annual Meeting of the Association of Computational Linguistics (ACL)*.

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 the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '13.

A. Smola and R. Kondor. 2003. Kernels and regularization on graphs. *COLT*.

Y. Sun, R. Barber, M. Gupta, C. Aggarwal, and J. Han. 2011a. Co-author relationship prediction in heterogeneous bibliographic networks. In *Proceedings of the 2011 International Conference on Advances in Social Networks Analysis and Mining*, ASONAM '11.

Y. Sun, J. Han, X. Yan, P. Yu, and T. Wu. 2011b. Pathsim: Meta path-based top-k similarity search in heterogeneous information networks. *PVLDB*, 4(11).

P. Talukdar and K. Crammer. 2009. New regularized algorithms for transductive learning. In *Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases: Part II*, ECML PKDD '09.

D. Vitale, P. Ferragina, and U. Scaiella. 2012. Classification of short texts by deploying topical annotations. In *ECIR*, pages 376–387.

K. Wang, C. Thrasher, and B. Hsu. 2011. Web scale nlp: A case study on url word breaking. In *Proceedings of the 20th International Conference on World Wide Web*, WWW '11.

W. Zhang, J. Su, C. Tan, and W. Wang. 2010. Entity linking leveraging automatically generated annotation. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*.

W. Zhang, J. Su, and C. L. Tan. 2011. A wikipedia-lda model for entity linking with batch size changing. In *Proc. IJCNLP2011*.

D. Zhou, O. Bousquet, T. Lal, J. Weston, and B. Schölkopf. 2004. Learning with local and global consistency. In *Advances in Neural Information Processing Systems 16*.

X. Zhu, Z. Ghahramani, and J. Lafferty. 2003. Semi-supervised learning using gaussian fields and harmonic functions. In *ICML*.