Resolving Entity Morphs in Censored Data

Hongzhao Huang[1], Zhen Wen[2], Dian Yu[1], Heng Ji[1], Yizhou Sun[3], Jiawei Han[4], He Li[5]

[1]Computer Science Department and Linguistics Department,
Queens College and Graduate Center, City University of New York, New York, NY, USA
[2]IBM T. J. Watson Research Center, Hawthorne, NY, USA
[3]College of Computer and Information Science, Northeastern University, Boston, MA, USA
[4]Computer Science Department, Univerisity of Illinois at Urbana-Champaign, Urbana, IL, USA
[5]Admaster Inc., China
{hongzhaohuang[1],yudiandoris[1],hengjicuny[1], liheact[5]}@gmail.com,
zhenwen@us.ibm.com[2], yzsun@ccs.neu.edu[3], hanj@illinois.edu[4]

Abstract

In some societies, internet users have to create information morphs (e.g. "*Peace West King*" to refer to "*Bo Xilai*") to avoid active censorship or achieve other communication goals. In this paper we aim to solve a new problem of resolving entity morphs to their real targets. We exploit temporal constraints to collect cross-source comparable corpora relevant to any given morph query and identify target candidates. Then we propose various novel similarity measurements including surface features, meta-path based semantic features and social correlation features and combine them in a learning-to-rank framework. Experimental results on Chinese Sina Weibo data demonstrate that our approach is promising and significantly outperforms baseline methods[1].

1 Introduction

Language constantly evolves to maximize communicative success and expressive power in daily social interactions. The proliferation of online social media significantly expedites this evolution, as new phrases triggered by social events may be disseminated rapidly in social media. To automatically analyze such fast evolving language in social media, new computational models are demanded.

In this paper, we focus on one particular language evolution that creates new ways to communicate sensitive subjects because of the existence of internet information censorship. We call this phenomenon *information morph*. For example, when Chinese online users talk about the former politician "*Bo Xilai*", they use a morph "*Peace West King*" instead, a historical figure four hundreds years ago who governed the same region as Bo. Morph can be considered as a special case of alias used for hiding true entities in malicious environment (Hsiung et al., 2005; Pantel, 2006). However, social network plays an important role in generating morphs. Usually morphs are generated by harvesting the collective wisdom of the crowd to achieve certain communication goals. Aside from the purpose of avoiding censorship, other motivations for using morph include expressing sarcasm/irony, positive/negative sentiment or making descriptions more vivid toward some entities or events. Table 1 presents the wide range of cases that are used to create the morphs. We can see that a morph can be either a regular term with new meaning or a newly created term.

Morph	Target	Motivation
Peace West King	Bo Xilai	Sensitive
Blind Man	Chen Guangcheng	Sensitive
Miracle Brother	Wang Yongping	Irony
Kim Fat	Kim Joing-il	Negative
Kimchi Country	South Korea	Vivid

Table 1: Morph Examples and Motivations.

We believe that successful resolution of morphs is a crucial step for automated understanding of the fast evolving social media language, which is important for social media marketing (Barwise and Meehan, 2010). Another application is to help common users without enough background/cultural knowledge to understand internet language for their daily use. Furthermore, our approaches can also be applied for satire or other implicit meaning recognition, as well as information extraction (Bollegala et al., 2011).

[1]Some of the resources and open source programs developed in this work are made freely available for research purpose at http://nlp.cs.qc.cuny.edu/Morphing.tar.gz

Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, pages 1083–1093,
Sofia, Bulgaria, August 4-9 2013. ©2013 Association for Computational Linguistics

However, morph resolution in social media is challenging due to the following reasons. First, the sensitive real targets that exist in the same data source under active censorship are often automatically filtered. Table 2 presents the distributions of some examples of morphs and their targets in English Twitter and Chinese Sina Weibo. For example, the target "*Chen Guangcheng*" only appears once in Weibo. Thus, the co-occurrence of a morph and its target is quite low in the vast amount of information in social media. Second, most morphs were not created based on pronunciations, spellings or other encryptions of their original targets. Instead, they were created according to semantically related entities in historical and cultural narratives (e.g. "*Peace West King*" as morph of "*Bo Xilai*") and thus very difficult to capture based on typical lexical features. Third, tweets from Twitter/Chinese Weibo are short (only up to 140 characters) and noisy, resulting in difficult extraction of rich and accurate evidences due to the lack of enough contexts.

Morph	Target	Frequency in Twitter		Frequency in Weibo	
		Morph	Target	Morph	Target
Hu Ji	Hu Jintao	1	3,864	2,611	71
Blind Man	Chen Guangcheng	18	2,743	20,941	1
Baby	Wen Jiabao	2238	2021	26,279	8

Table 2: Distributions of Morph Examples

To the best of our knowledge, this is the first work to use NLP and social network analysis techniques to automatically resolve morphed information. To address the above challenges, our paper offers the following novel contributions.

- We detect target candidates by exploiting the dynamics of the social media to extract temporal distribution of entities, based on the assumption that the popularity of an individual is correlated between censored and uncensored text within a certain time window.
- Our approach builds and analyzes heterogeneous information networks from multiple sources, such as Twitter, Sina Weibo and web documents in formal genre (e.g. news) because a morph and its target tend to appear in similar contexts.
- We propose two new similarity measures, as well as integrating temporal information into the similarity measures to generate global semantic features.
- We model social user behaviors and use social correlation to assist in measuring semantic similarities because the users who posted a morph and its corresponding target tend to share similar interests and opinions.

Our experiments demonstrate that the proposed approach significantly outperforms traditional alias detection methods (Hsiung et al., 2005).

2 Approach Overview

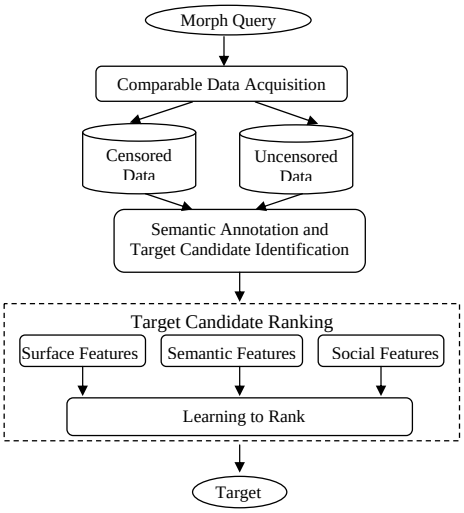

Figure 1: Overview of Morph Resolution

Given a morph query m, the goal of morph resolution is to find its real target. Figure 1 depicts the general procedure of our approach. It consists of two main sub-tasks:

- ***Target Candidate Identification***: For each m, discover a list of target candidates $E = \{e_1, e_2, ..., e_N\}$. First, relevant comparable data sets that include m are retrieved. In this paper we collect comparable censored data from Weibo and uncensored data from Twitter and Web documents such as news articles. We then apply various annotations such as word segmentation, part-of-speech tagging, noun phrase chunking, name tagging and event extraction to these data sets.
- ***Target Candidate Ranking***: Rank the target candidates in E. We explore various features including surface, semantic and social features, and incorporate them into a learning to

rank framework. Finally, the top ranked candidate is produced as the resolved target.

3 Target Candidate Identification

The general goal of the first step is to identify a list of target candidates for each morph query from the comparable corpora including Sina Weibo, Chinese News websites and English Twitter. However, obviously we cannot consider all of the named entities in these sources as target candidates due to the sheer volume of information. In addition, morphs are not limited to named entity forms. In order to narrow down the scope of target candidates, we propose a *Temporal Distribution Assumption* as follows. The intuition is that a morph m and its real target e should have similar temporal distributions in terms of their occurrences. Suppose the data sets are separated into Z temporal slots (e.g. by day), the assumption can be stated as:

Let $T_m = \{t_{m1}, t_{m2}, ..., t_{mZ_m}\}$ be the set of temporal slots each morph m occurs, and $T_e = \{t_{e1}, t_{e2}, ..., t_{eZ_e}\}$ be the set of slots a target candidate e occurs. Then e is considered as a target candidate of m if and only if, for each $t_{mi} \in T_m$ $(i = 1, 2, ..., Z_m)$, there exist a $j \in \{1, 2, ..., Z_e\}$ such that $t_{mi} - t_{ej} \leq \delta$, where δ is a threshold value (in this paper we set the threshold to 7 days, which is optimized from a development set). For comparison we also attempted topic modeling approach to detect target candidates, as shown in section 5.3.

4 Target Candidate Ranking

Next, we propose a learning-to-rank framework to rank target candidates based on various levels of novel features based on surface, semantic and social analysis.

4.1 Surface Features

We first extract surface features between the morph and the candidate based on measuring orthographic similarity measures which were commonly used in entity coreference resolution (e.g. (Ng, 2010; Hsiung et al., 2005)). The measures we use include "string edit distance", "normalized string edit distance" (Wagner and Fischer, 1974) and "longest common subsequence" (Hirschberg, 1977).

4.2 Semantic Features

4.2.1 Motivations

Fortunately, although a morph and its target may have very different orthographic forms, they tend to be embedded in *similar semantic contexts* which involve similar topics and events. Figure 2 presents some example messages under censorship (Weibo) and not under censorship (Twitter and Chinese Daily). We can see that they include similar topics, events (e.g., "*fell from power*", "*gang crackdown*", "*sing red songs*"), and semantic relations (e.g., family relations with "*Bo Guagua*"). Therefore if we can automatically extract and exploit these indicative semantic contexts, we can narrow down the real targets effectively.

Figure 2: Cross-source Comparable Data Example (each morph and target pair is shown in the same color)

4.2.2 Information Network Construction

We define an information network as a directed graph $G = (\mathcal{V}, \mathcal{E})$ with an object type mapping function $\tau : \mathcal{V} \rightarrow \mathcal{A}$ and a link type mapping function $\phi : \mathcal{E} \rightarrow \mathcal{R}$, where each object $v \in \mathcal{V}$ belongs to one particular object type $\tau(v) \in \mathcal{A}$, each link $e \in \mathcal{E}$ belongs to a particular relation $\phi(e) \in \mathcal{R}$. If two links belong to the same relation type, then they share the same starting object type as well as the same ending object type. An information network is *homogeneous* if and only if there is only one type for both objects and links, and an information network is *heterogeneous* when the objects are from multiple distinct types or there exist more than one type of links.

In order to construct the information networks for morphs, we apply the Standford Chinese word

segmenter with Chinese Penn Treebank segmentation standard (Chang et al., 2008) and Stanford part-of-speech tagger (Toutanova et al., 2003) to process each sentence in the comparable data sets. Then we apply a hierarchical Hidden Markov Model (HMM) based Chinese lexical analyzer ICTCLAS (Zhang et al., 2003) to extract named entities, noun phrases and events.

We have also attempted using the results from Dependency Parsing, Relation Extraction and Event Extraction tools (Ji and Grishman, 2008) to enrich the link types. Unfortunately the state-of-the-art techniques for these tasks still perform poorly on social media in terms of both accuracy and coverage of important information, these sophisticated semantic links all produced negative impact on the target ranking performance. Therefore we limited the types of vertices into: *Morph* (M), *Entity*(E), which includes target candidates, *Event* (EV), and *Non-Entity Noun Phrases* (NP); and used *co-occurrence* as the edge type. We extract entities, events, and non-entity noun phrases that occur in more than one tweet as neighbors. And for two vertices x_i and x_j, the weight w_{ij} of their edge is the frequency they co-occur together within the tweets. A network schema of such networks is shown in Figure 3. Figure 4

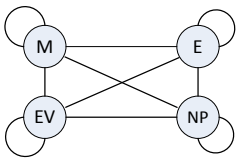

Figure 3: Network Schema of Morph-Related Heterogeneous Information Network

presents an example of a heterogeneous information network from the motivation examples following the above network schema, which connects the morphs "*Peace West King*", "*Buhou*" and their corresponding target "*Bo Xilai*".

4.2.3 Meta-Path-Based Semantic Similarity Measurements

Given the constructed network, a straightforward solution for finding the target for a morph is to use link-based similarity search. However, now objects are linked to different types of neighbors, if all neighbors are treated as the same, it may cause information loss problems. For example, the entity "重庆 (Chongqing)" is a very important aspect characterizing the politician "薄熙来(Bo Xilai)"

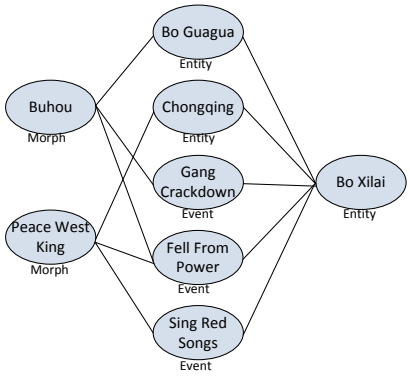

Figure 4: Example of Morph-Related Heterogeneous Information Network

since he governed it, and if a morph m which is also highly correlated with "重庆 (Chongqing)", it is very likely that "Bo Xilai" is the real target of m. Therefore, the semantic features generated from neighbors such as the entity "重庆 (Chongqing)" should be treated differently from other types of neighbors such as "人才(talented people)" .

In this work, we propose to measure the similarity of two nodes over heterogeneous networks as shown in Figure 3, by distinguishing neighbors into three types according to the network schema (i.e. entities, events, non-entity noun phrases). We then adopt meta-path-based similarity measures (Sun et al., 2011a; Sun et al., 2011b), which are defined over heterogeneous networks to extract semantic features. A meta-path is a path defined over a network, and composed of a sequence of relations between different object types. For example, as shown in Figure 3, a morph and its target candidate can be connected by three meta-paths, including "M - E - E", "M - EV - E", and "M - NP - E". Intuitively, each meta-path provides a unique angle to measure how similar two objects are.

For the determined meta-paths, we extract semantic features using the similarity measures proposed in (Sun et al., 2011a; Hsiung et al., 2005). We denote the neighbor sets of certain type for a morph m and a target candidate e as $\Gamma(m)$ and $\Gamma(e)$, and a meta-path as $\mathcal{P}$. We now list several meta-path-based similarity measures below.

Common neighbors (CN). It measures the number of common neighbors that m and e share as $|\Gamma(m) \cap \Gamma(e)|$.

Path count (PC). It measures the number of path instances between m and e following meta-path $\mathcal{P}$.

Pairwise random walk (PRW). For a meta-path $\mathcal{P}$ that can be decomposed into two shorter

meta-paths with the same length $\mathcal{P} = (\mathcal{P}_1\mathcal{P}_2)$, pairwise random walk measures the probability of the pairwise random walk starting from both m and e and reaching the same middle object. More formally, it is computed as $\sum_{(p_1p_2)\in(\mathcal{P}_1\mathcal{P}_2)} prob(p_1)prob(p_2^{-1})$, where p_2^{-1} is the inverse of p_2.

Kullback-Leibler distance (KLD). For m and e, the pairwise random walk probability of their neighbors can be represented as two probability vectors, then Kullback-Leibler distance (Hsiung et al., 2005) can be used to compute $sim(m, e)$.

Beyond the above similarity measures, we also propose to use cosine-similarity-style normalization method to modify common neighbor and pairwise random walk measures so that we can ensure the morph node and the target candidate node are strongly connected and also have similar popularity. The modified algorithms penalize features involved with the highly popular objects, since they are more likely to have accidental interactions with each other.

Normalized common neighbors (NCN). Normalized common neighbors can be measured as $sim(m, e) = \frac{|\Gamma(m)\cap\Gamma(e)|}{\sqrt{|\Gamma(m)|}\sqrt{|\Gamma(e)|}}$. It refines the simple counting of common neighbors by avoiding bias to highly visible or concentrated objects.

Pairwise random walk/cosine (PRW/cosine). Pairwise random walk measures linkage weights disproportionately with their visibility to their neighbors, which may be too strong. Instead, we propose to use a tamer normalization method as $\sum_{(p_1p_2)\in(\mathcal{P}_1\mathcal{P}_2)} f(p_1)f(p_2^{-1})$, where.

$$f(p_1) = \frac{count(m, x)}{\sqrt{\sum_{x\in\Omega} count(m, x)}},$$
$$f(p_2) = \frac{count(e, x)}{\sqrt{\sum_{x\in\Omega} count(e, x)}},$$

and Ω is the set of middle objects connecting the decomposed meta-paths p_1 and p_2^{-1}, $count(y, x)$ is the total number of paths between y and the middle object x, y could be m or e.

The above similarity measures can also be applied to homogeneous networks that do not differentiate the neighbor types.

4.2.4 Global Semantic Feafure Generation

A morph tends to have higher temporal correlation with its real target, and share more similar topics compared to other irrelevant targets. Therefore, we propose to incorporate temporal information into similarity measures to generate global semantic features.

Let $T = t_1 \cup t_2 \cup ... \cup t_N$ be a set of temporal slots (i.e. by day), E be the set of target candidates for each morph m. Then for each $t_i \in T$, and each $e \in E$, the local semantic features $sim_{t_i}(m, e)$ is extracted based only on the information posted within t_i using one of the similarity measures introduced in Section 4.2.3. Then we propose two approaches to generate global semantic features. The first approach is adding the similarity score between m and e in each temporal slot to attain the first set of global features:

$$sim_{global\_sum}(m, e) = \sum_{t_i\in T} sim_{t_i}(m, e).$$

The second method first normalizes the similarity score in each temporal slot t_i, them sum the normalized scores to generate the second set of global features, which can be calculated as

$$sim_{global\_norm}(m, e) = \sum_{t_i\in T} norm_{t_i}(m, e).$$

where $norm_{t_i}(m, e) = \frac{sim_{t_i}(m,e)}{\sum_{e\in E} sim_{t_i}(m,e)}$.

4.2.5 Integrate Cross Source/Cross Genre Information

Due to internet information censorship or surveillance, users may need to use morphs to post sensitive information. For example, the Chinese Weibo message "都进去了,还要贡着不厚吗 *(Already put in prison, still need to serve Buhou?*" include a morph 不厚 *(Buhou)*. In contrast, users are less restricted in some other uncensored social media such as Twitter. For example, the tweet from Twitter "...把薄熙来称作"平西王"或者"不厚"... *(...call Bo Xilai"peace west king" or "buhou"...)*" contains both the morph and the real target 薄熙来 *(Bo Xilai)*. Therefore, we propose to integrate information from another source (e.g. Twitter) to help resolution of sensitive morphs in Weibo.

Another difficulty from morph resolution in micro-blogging is that tweets are only allowed to contain maximum 140 characters with a lot of noise and diverse topics. The shortness and diversity of tweets may limit the power of content analysis for semantic feature extraction. However, formal genres such as web documents are cleaner and contain richer contexts, thus can provide more topically related information. In this work, we also exploit the background web documents from the

embedded URLs in tweets to enrich information network construction. After applying the same annotation techniques as tweets for uncensored data sets, sentence-level co-occurrence relations are extracted and integrated into the network as shown in Figure 3.

4.3 Social Features

It has been shown that there exist correlation between neighbors in social networks (Anagnostopoulos et al., 2008; Wen and Lin, 2010). Because of such social correlation, close social neighbors in social media such as Twitter and Weibo may post similar information, or share similar opinion. Therefore, we can utilize social correlation to assist in resolving morphs.

As social correlation can be defined as a function of social distance between a pair of users, we use social distance as a proxy to social correlation in our approach. The social distance between user i and j is defined by considering the degree of separation in their interaction (*e.g.* retweeting and mentioning) and the amount of the interaction. Similar definition has been shown effective in characterizing social distance in social networks extracted from communication data (Lin et al., 2012; Wen and Lin, 2010). Specifically, it is $dist(i,j) = \sum_{k=1}^{K-1} \frac{1}{strength(v_k, v_{k+1})}$, where $v_1, ..., v_k$ are the nodes on the shortest path from user i to user j, and $strength(v_k, v_{k+1})$ measures the strength of interactions between v_k and v_{k+1} as: $strength(i,j) = \frac{log(X_{ij})}{\max_j log(X_{ij})}$, where X_{ij} is the total interactions between user i and j, including both retweeting and mentioning (If $X_{ij} < 10$, we set $strength(i,j) = 0$).

We integrate social correlation and temporal information to define our social features. The intuition is that when a morph is used by an user, the real target may also in the posts by the user or his/her close friends within a certain time period. Let T be the set of temporal slots a morph m occurs, U_t be the set of users whose posts include m in slot t where $t \in T$, and U_c be the set of close friends (*i.e.*, social distance < 0.5) for U_t. The social features are defined as

$$s(m,e) = \frac{\sum_{t \in T} f(e,t,U_t,U_c)}{|T|}.$$

where $f(e,t,U_t,U_c)$ is a indicator function which return 1 if one of the users in U_t or U_c posts tweets include the target candidate e within 7 days before t.

4.4 Learning-to-Rank

Similar to (Hsiung et al., 2005; Sun et al., 2011a), we then model the probability of linkage prediction between a morph m and its target candidate e as a function incorporating the surface, semantic and social features. Given a training pair $\langle m, e \rangle$, we choose the standard logistic regression model to learn weights for the features defined above. The learnt model is used to predict the probability of linking an unseen morph and its target candidate. Based on the descending ranking order of the probability, we select top k candidates as the final answers based on the answer size k.

5 Experiments

Next, we present the experiment under various settings shown in Table 3, and the impacts of cross source and cross genre information.

5.1 Data and Evaluation Metric

We collected $1,553,347$ tweets from Chinese Sina Weibo from May 1 to June 30 to construct the censored data set, and retrieved $66,559$ web documents from the embedded URLs in tweets as the initial uncensored data set. Retweets and redundant web documents are filtered to ensure more reliable frequency counting of co-occurrence relations. We asked two native Chinese annotators to analyze the data, and construct a test set consisted of 107 morph entities (81 persons and 26 locations) and their real targets as our references. We verified the references by Web resources including the summary of popular morphs in Wikipedia [2]. In addition, we used 23 sensitive morphs and the entities that appear in the tweets as queries and retrieved $25,128$ Chinese tweets from 10% Twitter feeds within the same time period, as well as $7,473$ web documents from the embedded URLs and added them into the uncensored data set.

To evaluate the system performance, we use leave-one-out cross validation by computing accuracy as $Acc@k = \frac{C_k}{Q}$, where C_k is the total number of correctly resolved morphs at top k ranked answers, and Q is the total number of morph queries. We consider a morph as correctly resolved at the top k answers if the top k answer set contains the real target of the morph.

[2] http://zh.wikipedia.org/wiki/中国大陆网络语言列表

Feature sets	Descriptions
Surf	Surface features
HomB	Semantic features extracted from homogeneous CN, PC, PRW, and KLD
HomE	HomB + semantic features extracted from homogeneous NCN and PRW/cosine
HetB	Semantic features extracted from heterogeneous CN, PC, PRW and KLD
HetE	HetB + Semantic features extracted from heterogeneous NCN and PRW/cosine
Glob*	Global semantic features
Social	Social network features

Table 3: Description of feature sets. * Glob only uses the same set of similarity measures when combined with other semantic features.

5.2 Resolution Performance

5.2.1 Single Genre Information

We first study the contributions of each set of surface and semantic features, as shown in the first five rows in Table 4. The poor performance based on surface features shows that morph resolution task is very challenging since 70% of morphs are not orthographically similar to their real targets. Thus, capturing a morph's semantic meaning is crucial. Overall, the results demonstrate the effectiveness of our proposed methods. Specifically, comparing "HomB" and "HetB", "HomE" and "HetE", we can see that the semantic features based on heterogeneous networks have advantages over those based on homogeneous networks. This corroborates that different neighbor sets contribute differently, and such discrepancies should be captured. And comparisions of "HomB" and "HomE", "HetB" and "HetE"demonstrate the effectiveness of our two new proposed measures. To evaluate the importance of each similarity measures, we delete the semantic features obtained from each measure in "HetE" and re-evaluate the system. We find that NCN is the most effective measure, while KLD is the least important one. Further adding the global semantic features significantly improves the performance. This indicates that capturing both temporal correlations and semantics of morphing simultaneously are important for morph resolution.

Table 5 shows that combination of surface and semantic features further improves the performance, showing that they are complementary. For example, using only surface features, the real target "乔布斯 （Steve Jobs）" of the morph "乔帮主 (Qiao Boss)" is not top ranked since some other candidates such as "乔治 (George)" are more orthographically similar. However, "Steve Jobs" is ranked top when combined with semantic features.

Features	Surf	HomB	HomE	HetB	HetE
Acc@1	0.028	0.201	0.192	0.224	**0.252**
Acc@5	0.159	0.313	0.369	0.393	**0.421**
Acc@10	0.243	0.346	0.407	0.439	**0.467**
Acc@20	0.313	0.411	0.467	0.50	**0.523**
Features		+ Glob	+ Glob	+ Glob	+ Glob
Acc@1		0.230	**0.285**	0.257	**0.285**
Acc@5		0.402	0.407	0.449	**0.458**
Acc@10		0.435	0.458	**0.50**	0.495
Acc@20		0.486	0.523	**0.565**	0.542

Table 4: The System Performance Based on Each Single Feature Set.

Features	Surf + HomB	Surf + HomE	Surf + HetB	Surf + HetE
Acc@1	0.234	0.238	0.262	**0.276**
Acc@5	0.416	0.444	0.481	**0.519**
Acc@10	0.477	0.505	0.533	**0.570**
Acc@20	0.519	0.561	0.565	**0.598**
Features	+ Glob	+ Glob	+ Glob	+ Glob
Acc@1	0.290	0.341	0.322	**0.346**
Acc@5	0.505	0.495	0.528	**0.533**
Acc@10	0.551	0.551	0.579	**0.584**
Acc@20	0.594	0.603	**0.636**	0.631

Table 5: The System Performance Based on Combinations of Surface and Semantic Features.

5.2.2 Cross Source and Cross Genre Information

We integrate the cross source information from Twitter, and the cross genre information from web documents into Weibo tweets for information network construction, and extract a new set of semantic features. Table 6 shows that further gains can be achieved. Notice that integrating tweets from Twitter mainly improves the ranking for top k where $k > 1$. This is because Weibo dominates our dataset, and in Weibo many of these sensitive morphs are mostly used with their traditional meanings instead of the morph senses. Further performance improvement is achieved by integrating information from background formal web documents which can provide richer context and relations.

Features	Surf + HomB + Glob	Surf + HomE + Glob	Surf + HetB + Glob	Surf + HetE + Glob
Acc@1	0.290	0.341	0.322	**0.346**
Acc@5	0.505	0.495	0.528	**0.533**
Acc@10	0.551	0.551	0.579	**0.584**
Acc@20	0.594	0.603	**0.636**	0.631
Features	+ Twitter	+ Twitter	+ Twitter	+ Twitter
Acc@1	0.308	0.336	0.336	**0.346**
Acc@5	0.514	0.519	0.547	**0.565**
Acc@10	0.579	0.594	0.594	**0.636**
Acc@20	0.631	0.640	**0.668**	**0.668**
Features	+ Web	+ Web	+ Web	+ Web
Acc@1	0.327	0.360	0.341	**0.379**
Acc@5	0.528	0.519	0.565	**0.575**
Acc@10	0.594	0.589	0.622	**0.645**
Acc@20	0.631	0.650	**0.678**	**0.678**

Table 6: The System Performance of Integrating Cross Source and Cross Genre Information.

5.2.3 Effects of Social Features

Table 7 shows that adding social features can improve the best performance achieved so far. This is because a group of people with close relationships may share similar opinion. As an example, two tweets "*...of course the reputation of Buhou is a little too high! //@User1: //@User2: Chongqing event tells us...)*" and "*...do not follow Bo Xilai...@User1...)* are from two users in the same social group.One includes a morph "Buhou" and the other includes its target "Bo Xilai".

Features	Surf + HomB + Glob + Twitter + Web	Surf + HomE + Glob + Twitter + Web	Surf + HetB + Glob + Twitter + Web	Surf + HetE + Glob + Twitter + Web
Acc@1	0.327	0.360	0.341	**0.379**
Acc@5	0.528	0.519	0.565	**0.575**
Acc@10	0.594	0.589	0.622	**0.645**
Acc@20	0.631	0.650	**0.678**	**0.678**
Features	+ Social	+ Social	+ Social	+ Social
Acc@1	0.336	0.369	0.365	**0.379**
Acc@5	0.537	0.547	0.589	**0.594**
Acc@10	0.594	0.601	0.645	**0.659**
Acc@20	0.645	0.664	**0.701**	**0.701**

Table 7: The Effects of Social Features.

5.3 Effects of Candidate Detection

The performance with and without candidate detection step (using all features) is shown in Table 8. The gain is small since the combination of all features in the learning to rank framework can already well capture the relationship between a morph and a target candidate. Nevertheless, the temporal distribution assumption is effective. It helps to filter out 80% of unrelated targets and speed up the system 5 times, while retain 98.5% of the morph candidates that can be detected.

System	Acc@1	Acc@5	Acc@10	Acc@20
Without	0.365	0.579	0.645	0.696
With	**0.379**	**0.594**	**0.659**	**0.701**

Table 8: The Effects of Temporal Constraint

We also attempted using topic modeling approach to detect target candidates. Due to the large amount of data, we first split the data set on a daily basis, then applied Probabilistic Latent Semantic Analysis (PLSA) (Hofmann, 1999). Named entities which co-occur at least δ times with a morph query in the same topic are selected as its target candidates. As shown in Table9 (K is the number of predefined topics), PLSA is not quite effective mainly because traditional topic modeling approaches do not perform well on short texts from social media. Therefore, in this paper we choose a simple method based on temporal distribution to detect target candidates.

Method	All	Temporal	PLSA($K=5$ $\delta=1$)	PLSA($K=5$ $\delta=2$)
Acc	0.935	0.921	0.935	0.925
No.	$8,111$	$1,964$	$6,380$	$4,776$
Method	PLSA($K=10$ $\delta=1$)	PLSA($K=10$ $\delta=2$)	PLSA($K=20$ $\delta=1$)	PLSA($K=20$ $\delta=2$)
Acc	0.935	0.907	0.888	0.757
No.	$5,117$	$3,138$	$3,702$	$1,664$

Table 9: Accuracy of Target Candidate Detection

5.4 Discussions

Compared with the standard alias detection ("Surf+HomB") approach (Hsiung et al., 2005), our proposed approach achieves significantly better performance (99.9% confidence level by the Wilcoxon Matched-Pairs Signed-Ranks Test for Acc@1). We further explore two types of factors which may affect the system performance as follows.

One important aspect affecting the resolution performance is the morph & non-morph ambiguity. We categorize a morph query as "Unique" if the string is mainly used as a morph when it occurs, such as "薄督 (Bodu)" which is used to refer to "Bo Xilai"; otherwise as "Common" (e.g. "宝宝 (Baby)" ,"校长 (President)"). Table 10 presents the separate scores for these two categories. We can see that the morphs in "Unique"

category have much better resolution performance than those in "Common" category.

Category	Number	Acc@1	Acc@5	Acc@10	Acc@20
Unique	72	0.479	0.715	0.771	0.819
Common	35	0.171	0.343	0.40	0.429

Table 10: Performance of Two Categories

We also investigate the effects of popularity of morphs on the resolution performance. We split the queries into 5 bins with equal size based on the non-descending frequency, and evaluate Acc@1 separately. As shown in Table11, we can see that the popularity is not highly correlated with the performance.

Rank	0 ∼ 20%	20% ∼ 40%	40% ∼ 60%	60% ∼ 80%	80% ∼ 100%
All	0.333	0.476	0.341	0.429	0.318
Unique	0.321	0.679	0.379	0.571	0.483
Common	0.214	0.214	0.071	0.071	0.286

Table 11: Effects of Popularity of Morphs

6 Related Work

To analyze social media behavior under active censorship, (Bamman et al., 2012) automatically discovered politically sensitive terms from Chinese tweets based on message deletion analysis. In contrast, our work goes beyond target idendification by resolving implicit morphs to their real targets.

Our work is closely related to alias detection (Hsiung et al., 2005; Pantel, 2006; Bollegala et al., 2011; Holzer et al., 2005). We demonstrated that state-of-the-art alias detection methods did not perform well on morph resolution. In this paper we exploit cross-genre information and social correlation to measure semantic similarity. (Yang et al., 2011; Huang et al., 2012) also showed the effectiveness of exploiting information from formal web documents to enhance tweet summarization and tweet ranking.

Other similar research lines are the TAC-KBP Entity Linking (EL) (Ji et al., 2010; Ji et al., 2011), which links a named entity in news and web documents to an appropriate knowledge base (KB) entry, the task of mining name translation pairs from comparable corpora (Udupa et al., 2009; Ji, 2009; Fung and Yee, 1998; Rapp, 1999; Shao and Ng, 2004; Hassan et al., 2007) and the link prediction problem (Adamic and Adar, 2001; Liben-Nowell and Kleinberg, 2003; Sun et al., 2011b; Hasan et al., 2006; Wang et al., 2007; Sun et al., 2011a). Most of the work focused on unstructured or structured data with clean and rich relations (e.g. DBLP). In contrast, our work constructs heterogeneous information networks from unstructured, noisy multi-genre text without explicit entity attributes.

7 Conclusion and Future Work

To the best of our knowledge, this is the first work of resolving implicit information morphs from the data under active censorship. Our promising results can well serve as a benchmark for this new problem. Both of the Meta-path based and social correlation based semantic similarity measurements are proven powerful and complementary.

In this paper we have focused on entity morphs. In the future we will extend our method to discover other types of information morphs, such as events and nominal mentions. In addition, automatic identification of candidate morphs is another challenging task, especially when the mentions are ambiguous and can also refer to other real entities. Our ongoing work includes identifying candidate morphs from scratch, as well as discovering morphs for a given target based on anomaly analysis and textual coherence modeling.

Acknowledgments

Thanks to the three anonymous reviewers for their insightful comments. This work was supported by the U.S. Army Research Laboratory under Cooperative Agreement No. W911NF- 09-2-0053 (NS-CTA), the U.S. NSF CAREER Award under Grant IIS-0953149, the U.S. NSF EAGER Award under Grant No. IIS-1144111, the U.S. DARPA FA8750-13-2-0041 - Deep Exploration and Filtering of Text (DEFT) Program, the U.S. DARPA under Agreement No. W911NF-12-C-0028, CUNY Junior Faculty Award, NSF IIS-0905215, CNS-0931975, CCF-0905014, and MIAS, a DHS-IDS Center for Multimodal Information Access and Synthesis at UIUC. 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

Lada A. Adamic and Eytan Adar. 2001. Friends and neighbors on the web. *SOCIAL NETWORKS*, 25:211–230.

Aris Anagnostopoulos, Ravi Kumar, and Mohammad Mahdian. 2008. Influence and correlation in social networks. In *KDD*, pages 7–15.

David Bamman, Brendan O'Connor, and Noah A. Smith. 2012. Censorship and deletion practices in chinese social media. *First Monday*, 17(3).

Patrick Barwise and Seán Meehan. 2010. The one thing you must get right when building a brand. *Harvard Business Review*, 88(12):80–84.

D. Bollegala, Y. Matsuo, and M. Ishizuka. 2011. Automatic discovery of personal name aliases from the web. *Knowledge and Data Engineering, IEEE Transactions on*, 23(6):831–844.

Pi-Chuan Chang, Michel Galley, and Christopher D. Manning. 2008. Optimizing chinese word segmentation for machine translation performance. In *Proceedings of the Third Workshop on Statistical Machine Translation*, StatMT '08, pages 224–232.

Pascale Fung and Lo Yuen Yee. 1998. An ir approach for translating new words from nonparallel, comparable texts. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics - Volume 1*, ACL '98, pages 414–420.

Mohammad Al Hasan, Vineet Chaoji, Saeed Salem, and Mohammed Zaki. 2006. Link prediction using supervised learning. In *In Proc. of SDM 06 workshop on Link Analysis, Counterterrorism and Security*.

Ahmed Hassan, Haytham Fahmy, and Hany Hassan. 2007. Improving named entity translation by exploiting comparable and parallel corpora. In *RANLP*.

Daniel S. Hirschberg. 1977. Algorithms for the longest common subsequence problem. *J. ACM*, 24(4):664–675.

Thomas Hofmann. 1999. Probabilistic latent semantic indexing. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '99, pages 50–57.

Ralf Holzer, Bradley Malin, and Latanya Sweeney. 2005. Email alias detection using social network analysis. In *Conference on Knowledge Discovery in Data: Proceedings of the 3 rd international workshop on Link discovery*, volume 21, pages 52–57.

Paul Hsiung, Andrew Moore, Daniel Neill, and Jeff Schneider. 2005. Alias detection in link data sets. In *Proceedings of the International Conference on Intelligence Analysis*, May.

Hongzhao Huang, Arkaitz Zubiaga, Heng Ji, Hongbo Deng, Dong Wang, Hieu Khac Le, Tarek F. Abdelzaher, Jiawei Han, Alice Leung, John Hancock, and Clare R. Voss. 2012. Tweet ranking based on heterogeneous networks. In *COLING*, pages 1239–1256.

Heng Ji and Ralph Grishman. 2008. Refining event extraction through cross-document inference. In *Proceedings of ACL*, pages 254–262.

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*.

Heng Ji. 2009. Mining name translations from comparable corpora by creating bilingual information networks. In *Proceedings of the 2nd Workshop on Building and Using Comparable Corpora: from Parallel to Non-parallel Corpora*, BUCC '09, pages 34–37.

David Liben-Nowell and Jon Kleinberg. 2003. The link prediction problem for social networks. In *Proceedings of the twelfth international conference on Information and knowledge management*, CIKM '03, pages 556–559.

Ching-Yung Lin, Lynn Wu, Zhen Wen, Hanghang Tong, Vicky Griffiths-Fisher, Lei Shi, and David Lubensky. 2012. Social network analysis in enterprise. *Proceedings of the IEEE*, 100(9):2759–2776.

Vincent Ng. 2010. Supervised noun phrase coreference research: the first fifteen years. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, ACL '10, pages 1396–1411.

Patrick Pantel. 2006. Alias detection in malicious environments. In *AAAI Fall Symposium on Capturing and Using Patterns for Evidence Detection*, pages 14–20.

Reinhard Rapp. 1999. Automatic identification of word translations from unrelated english and german corpora. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, ACL '99, pages 519–526.

Li Shao and Hwee Tou Ng. 2004. Mining new word translations from comparable corpora. In *Proceedings of the 20th international conference on Computational Linguistics*, COLING '04.

Yizhou Sun, Rick Barber, Manish Gupta, Charu C. Aggarwal, and Han Jiawei. 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, pages 121–128.

Yizhou Sun, Jiawei Han, Xifeng Yan, Philip S. Yu, and Tianyi Wu. 2011b. Pathsim: Meta path-based top-k similarity search in heterogeneous information networks. *PVLDB*, 4(11):992–1003.

Kristina Toutanova, Dan Klein, Christopher D. Manning, and Yoram Singer. 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*, NAACL '03, pages 173–180.

Raghavendra Udupa, K. Saravanan, A. Kumaran, and Jagadeesh Jagarlamudi. 2009. Mint: a method for effective and scalable mining of named entity transliterations from large comparable corpora. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '09, pages 799–807.

Robert A. Wagner and Michael J. Fischer. 1974. The string-to-string correction problem. *J. ACM*, 21(1):168–173.

Chao Wang, Venu Satuluri, and Srinivasan Parthasarathy. 2007. Local probabilistic models for link prediction. In *Proceedings of the 2007 Seventh IEEE International Conference on Data Mining*, ICDM '07, pages 322–331.

Zhen Wen and Ching-Yung Lin. 2010. On the quality of inferring interests from social neighbors. In *KDD*, pages 373–382.

Zi Yang, Keke Cai, Jie Tang, Li Zhang, Zhong Su, and Juanzi Li. 2011. Social context summarization. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*, SIGIR '11, pages 255–264.

Hua-Ping Zhang, Hong-Kui Yu, De-Yi Xiong, and Qun Liu. 2003. Hhmm-based chinese lexical analyzer ictclas. In *Proceedings of the second SIGHAN workshop on Chinese language processing - Volume 17*, SIGHAN '03, pages 184–187.