Entity Linking within a Social Media Platform: A Case Study on Yelp

Hongliang Dai[1], Yangqiu Song[1], Liwei Qiu[2] and Rijia Liu[2]
[1]Department of CSE, HKUST
[2]Tencent Technology (SZ) Co., Ltd.
[1]{hdai,yqsong}@cse.ust.hk
[2]{drolcaqiu,rijialiu}@tencent.com

Abstract

In this paper, we study a new entity linking problem where both the entity mentions and the target entities are within a same social media platform. Compared with traditional entity linking problems that link mentions to a knowledge base, this new problem have less information about the target entities. However, if we can successfully link mentions to entities within a social media platform, we can improve a lot of applications such as comparative study in business intelligence and opinion leader finding. To study this problem, we constructed a dataset called Yelp-EL, where the business mentions in Yelp reviews are linked to their corresponding businesses on the platform. We conducted comprehensive experiments and analysis on this dataset with a *learning to rank* model that takes different types of features as input, as well as a few state-of-the-art entity linking approaches. Our experimental results show that two types of features that are not available in traditional entity linking: social features and location features, can be very helpful for this task.

1 Introduction

Entity linking is the task of determining the identities of entities mentioned in texts. Most existing studies on entity linking have focused on linking entity mentions to their referred entities in a knowledge base (Cucerzan, 2007; Liu et al., 2013; Ling et al., 2015). However, on social media platforms such as Twitter, Instagram, Yelp, Facebook, etc., the texts produced on them may often mention entities that cannot be found in a knowledge base, but can be found on the platform itself. For example, consider Yelp, a platform where users can write reviews about businesses such as restaurants, hotels, etc., a restaurant review on Yelp may mention another restaurant to compare, which is also likely to be on Yelp but cannot be found in a knowledge base such as Wikipedia. As another example, when people post a photo on a social media platform, their friends may be mentioned in this post if they are also in the photo. Usually, their friends are not included in a knowledge base but may also have accounts on the same platform. Thus for such entity mentions, linking them to an account that is also on the platform is more practical than linking them to a knowledge base.

Performing this kind of entity linking can benefit many applications. For example, on Yelp, we can perform analysis on the comparative sentences in reviews after linking the business mentions in them. The results can be directly used to either provide recommendations for users or suggestions for business owners.

Thus, in this paper, we focus on a new entity linking problem where both the entity mentions and the target entities are within a social media platform. Specifically, the entity mentions are from the texts (which we will refer to as *mention texts*) produced by the users on a social media platform; and these mentions are linked to the accounts on this platform.

It is not straightforward to apply existing entity linking systems that link to a knowledge base to this problem, because they usually take advantage of the rich information knowledge bases provide for the entities. For example, they can use detailed text descriptions, varies kinds of attributes, etc., as features (Francis-Landau et al., 2016; Gupta et al., 2017; Tan et al., 2017), or even additional signals such as the anchor texts in Wikipedia articles (Guo and Barbosa, 2014; Globerson et al., 2016; Ganea et al., 2016). However, on social media platforms, most of these resources or information are either unavailable or of poor quality.

On the other hand, social media platforms also have some unique resources that can be exploited. One that commonly exists on all of them is social

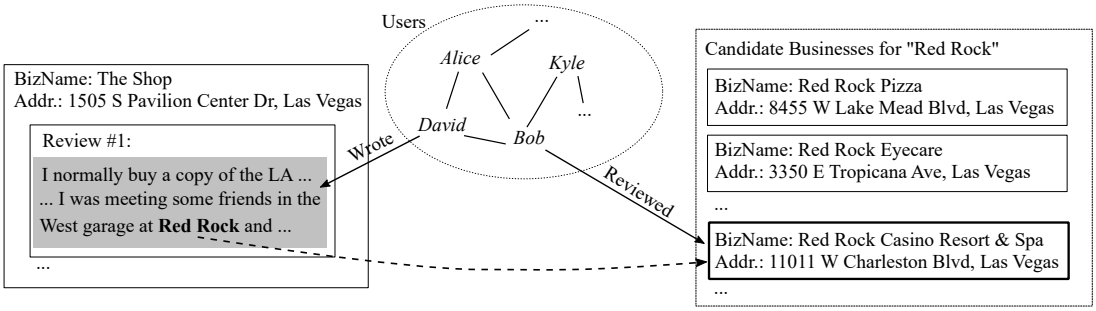

Figure 1: An example of entity linking within the Yelp social media platform. On Yelp, users can have friends which makes it a social network. Users can also write reviews about a business and compare with other businesses.

information, which can be intuitively used in our problem where *mention texts* and target entities may be directly connected by users and their social activities. Other than this, for location-based social media platforms such as Yelp and Foursquare, location information can also be helpful since people are more likely to mention and compare places close to each other.

To study this problem, we construct a dataset based on Yelp, which we name as Yelp-EL. As shown in Figure 1, on Yelp, users can write reviews for businesses and friend other users, and the reviews they write may mention businesses other than the reviewed ones. Thus, reviews, users, and businesses are connected and form a network through users' activities on the platform. In Yelp-EL, we link the business mentions in reviews to their corresponding businesses on the platform. We choose Yelp because other social media platforms such as Facebook and Instagram do not provide open dataset and there can be privacy issues related.

We then study the roles of three types of features in our entity linking problem: social features, location features, as well as conventional features that are also frequently used in traditional entity linking problems. We implemented a *learning to rank* model that takes the above features as input. We conducted comprehensive experiments and analysis on Yelp-EL with this model and also a few state-of-the-art entity linking approaches that we tailored to meet the requirements of Yelp-EL. Experimental results show that both social and location features can improve performance significantly.

Our contributions are summarized as follows.

- We are the first attempt to study the new entity linking problem where both entity mentions and target entities are within a same social media platform.
- We created a dataset based on Yelp to illustrate the usefulness of this problem and use it as a benchmark to compare different approaches.
- We studied both traditional entity linking features and social/location based features that are available from the social media platform, and show that they are indeed helpful for improving the entity linking performance.

The code and data are available at https://github.com/HKUST-KnowComp/ELWSMP.

2 Yelp-EL Dataset Construction

In this section we introduce how we create the dataset Yelp-EL based on the Yelp social media platform. We used the Round 9 version of the Yelp challenge dataset[1] to build Yelp-EL. There are 4,153,150 reviews, 144,072 businesses, and 1,029,432 users in this dataset. In order to build Yelp-EL, we first find possible entity mentions in Yelp reviews, and then ask people to manually link these mentions to Yelp businesses if possible.

Ideally, the mentions we need to extract from the reviews should be only those that refer the businesses in Yelp. Unfortunately, there is no existing method or tool that can accomplish this task. In fact, this problem itself is worth studying. Nonetheless, since we focus on entity linking in this paper, we only try to find as many mentions that may refer to Yelp businesses as we can, and then let the annotators decide whether to link this mention to a business. Thus, we use the following two ways to find mentions and then merge their results.

[1] https://www.yelp.com/dataset/challenge

#Mentions	#Linked	#NIL	#Disagreement1	#Disagreement2	Agreement%
7,731	1,749	5,117	842	23	88.8%

Table 1: Annotation statistics. "Linked" means the mentions that both annotators link to a same business. "NIL" means the mentions that both annotators think are "unlinkable." "Disagreement1" means the mentions that are labeled by one annotator as "unlinkable," but are linked to a business by the other annotator. "Disagreement2" means the mentions that are linked by two annotators to two different businesses.

(1) We use the Standford NER tool (Finkel et al., 2005) to find ordinary entity mentions and filter those that are unlikely to refer to businesses. To do the filtering, we first construct a dictionary which contains entity names that may occur in Yelp reviews frequently but are unlikely to refer to businesses, e.g., city names, country names, etc. Then we run through the mentions found with the NER tool and remove those whose mention strings matches one of the names in the dictionary.

(2) We find all the words/multi-word expressions in reviews that match the name of a business, and output them as mentions.

After extracting the mentions, we obtain the ground-truth by asking annotators to label them. Each time, we show the annotator one review with the mentions in this review highlighted, the annotator then needs to label each of the highlighted mentions. For each mention, we show several candidate businesses whose names match the mention string well. The annotator can also search the business by querying its name and/or location, in case the referred business is not included in the given candidates. We also ask the annotators to label the mention as "unlinkable" when its referred entity is not a Yelp business or it is not an entity mention.

An important issue to note is franchises. There are some mentions that refer to a franchise as a whole, e.g., the mention "Panda Express" in the sentence "If you want something different than the usual Panda Express this is the place to come." There are also some mentions that refer to a specific location of a franchise. For example, the mention "Best Buy" in "Every store you could possibly need is no further than 3 miles from here, which at that distance is Best Buy" refers to a specific "Best Buy" shop. As a location based social network platform, Yelp only contains businesses for different locations of franchises, not franchises themselves. Thus in these cases, we ask the annotators to link the mentions when they refer to a specific location of a franchise, but label them as "unlinkable" when they refer to a franchise as a whole.

We asked 14 annotators who are all undergraduate or graduate students in an English environment university to perform the annotation. They were given a tutorial before starting to annotate, and the annotation supervisor answered questions during the procedure to ensure the annotation quality. Each review is assigned to two annotators.

The statistics of the annotation results are shown in Table 1. The total agree rate, calculated as $(\#\text{Linked} + \#\text{NIL})/\#\text{Mentions}$, is 88.8%. Most disagreements are on whether to link a mention or not. We checked the data and find that this happens mostly when: they disagree on whether the mention refers to a franchise as a whole or just one specific location; one of the annotators fails to find the referred business. However, when both annotators think the mention should be linked to a business, the disagree rate, calculated as $\#\text{Disagreement2}/(\#\text{Linked} + \#\text{Disagreement2})$, is very low (only 1.3%).

We only use the mentions that both annotators give the same labeling results to build the dataset. As a result, we obtain 1,749 mentions that are linked to a business. These mentions refer to 1,134 different businesses (mentioned businesses) and are from 1,110 reviews. The reviews that contain these mentions are for 967 different businesses (reviewed businesses).

The reviewed businesses are located in 96 different cities and belong to 419 different categories. Note that a business can only locate in one city but may have several different categories. The mentioned businesses are located in 98 different cities and belong to 425 different categories. Figure 2 shows the numbers of reviewed businesses and mentioned businesses in the most popular cities and categories, from where we can see that these mentions have an acceptable level of diversity.

The mentions that can be linked are our focus, but we also include the 5,117 unlinkable mentions in our dataset since they can be helpful for building a complete entity discovery and linking system (Ji et al., 2016).

(a) Top Cities

(b) Top Categories

Figure 2: Statistics of the related businesses in Yelp-EL. (a) The number of businesses in the six most popular cities. (b) The number of businesses in the five most popular categories. Here, "popular" means having the largest number of businesses in the dataset.

3 Entity Linking Algorithm

In this section, we introduce *LinkYelp*, an entity linking approach we design for Yelp-EL to investigate the new proposed problem. LinkYelp contains two main steps: candidate generation and candidate ranking. The candidate generation step finds a set of businesses that are plausible to be the target of a mention based on the mention string. Afterwards, the candidate ranking step ranks all the candidates and chooses the top ranked one as the target business.

3.1 Candidate Generation

For the first step, candidate generation, we score each business b with $g(m,b) = g_c(m,b) \cdot g_n(s_m, s_b)$ for a mention m, where s_m is the mention string of m, s_b is the name of b. $g_c(m,b)$ equals to a constant value that is larger than 1 (it is set to 1.3 in practice) when the review that contains m is for a business that is located in the same city with b; Otherwise, it equals to 0. g_n is defined as

$$g_n(s_m, s_b) = \begin{cases} 1 & \text{if } s_m \in A(s_b) \\ \text{sim}(s_m, s_b) & \text{Otherwise,} \end{cases} \quad (1)$$

where $A(s_b)$ is the set of possible acronyms for s_b, $\text{sim}(s_m, s_b)$ is the cosine similarity between the TF-IDF representations of s_m and s_b. In practice, $A(s_b)$ is empty when s_b contains less than two words; Otherwise, it contains one string: the concatenation of the first letter of each word in s_b. Then, we find the top 30 highest scored businesses as candidates. This approach has a recall of 0.955 on Yelp-EL.

3.2 Candidate Ranking

Let m be a mention and b be a candidate business of m. We use the following function to score how likely b is the correct business that m refers to:

$$f(m,b) = \boldsymbol{w} \cdot \boldsymbol{\phi}(m,b), \quad (2)$$

where $\boldsymbol{\phi}(m,b)$ is the feature vector for mention-candidate pair m and b, Section 4 describes how to obtain it in detail; $\boldsymbol{w}$ is a parameter vector.

We use a max-margin based loss function to train $\boldsymbol{w}$:

$$J = \frac{1}{|T|} \sum_{<m,b_t,b_c> \in T} \max[0, 1 - f(m,b_t) + f(m,b_c)] + \lambda \|\boldsymbol{w}\|^2, \quad (3)$$

where b_t is the true business mention m refers to; $b_c \neq b_t$ is a corrupted business sample randomly picked from the candidates of m; T is the set of training samples; $\|\cdot\|$ is the l_2-norm; λ is a hyper-parameter that controls the regularization strength. We use stochastic gradient descent to train this model.

4 Feature Engineering

We study the effectiveness of three types of features: conventional features, social features, and location features. Among them, conventional features are those that can also be use in traditional entity linking tasks; social features and location features are unique in our problem.

4.1 Conventional Features

Lots of information used in traditional entity linking cannot be found for Yelp businesses, but we try our best to include all such features that can be used in our problem.

For Yelp-EL, we use the following conventional features for a mention m and its candidate business b:

u_1 : The cosine similarity between the TF-IDF representations of the mention string of m and the name of b.

u_2 : Whether the mention string of m is a possible acronym of b's name (i.e., whether it is an element of the set $A(s_b)$ in Equation 1).

u_3 : The popularity of b. Let the number of reviews received by b be n. Then this feature value equals to n/C if n is smaller than a parameter C that's used for normalization, otherwise it equals to 1.

u_4 : The cosine similarity between the TF-IDF representations of the review that contains m and combination of all reviews of b. This feature evaluates how well b fits m semantically.

u_5 : Whether b is the same as the reviewed business. This feature is actually not available in traditional EL, and it is usually not available on other social media platforms either. But it is obviously useful on Yelp-EL. Including it here helps us to see how beneficial social features and location features truly are.

4.2 Social Features

Through the activities of the users on the platform, the users, mentions, reviews and businesses in Yelp-EL form a network where there are different types of nodes and edges. Thus we use Heterogeneous Information Networks (HIN) to model it, and then design meta-path based features to capture the relations between mentions and their candidate businesses. We skip the formal definitions of HIN and meta-path here, readers can refer to (Sun et al., 2011) for detailed introduction. The HIN schema for Yelp-EL is shown in Figure 3.

The following meta-paths are used:

$$P1: \quad M - R - U - R - B$$
$$P2: \quad M - R - U - U - R - B$$
$$P3: \quad M - R - U - R - B - R - U - R - B$$

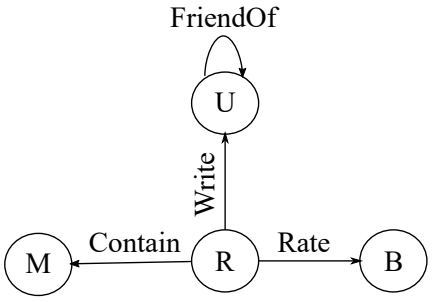

Figure 3: HIN schema of Yelp-EL. M: mention; R: Review; U: user; B: business.

where we denote M for mention, R for Review, U for user, and B for business.

Different meta-paths above capture different kinds of relations between a mention and its candidate entities that are induced by users' social activities. For example, if an instance of $P1$ exists between a mention m and a business b, then m is contained in a review that is written by a user who also reviewed business b. If many such instances of $P1$ exist, then we may assume that m and b are related, which makes it more possible for m to be referring to b.

With the meta-paths above, we use the *Path Count* feature defined in (Sun et al., 2011) to feed into the entity linking model described in Section 3. Given a meta-path P, for mention m and business b, *Path Count* is the number of path instances of P that start from m and end with b. In practice, we normalize this value based on global statistics before feeding it to a model.

4.3 Location Features

Location information commonly exists in location-based social media platforms such as Yelp and Foursquare. Users on platforms such as Twitter and Instagram may also be willing to provide their locations.

Here, we use the following two features for a mention m and its candidate business b:

v_1 : Whether the reviewed business is in the same city as b.

v_2 : The geographical distance between the reviewed business and b. This value is calculated based on the longitude and latitude coordinates of the businesses.

There are still some other location features that can be designed. For example, we can also consider the locations of the other businesses that are

reviewed by the user. We only use the above two since we find in our experiments that including them already provides high performance boost.

5 Experiments

5.1 Compared Methods

We compare with a baseline method we name as DirectLink, as well as two existing entity linking methods including the method proposed by (Liu et al., 2013) (which we refer to as ELT) and SSRegu proposed by (Huang et al., 2014).

DirectLink simply links each mention to the corresponding reviewed business. Many business mentions in Yelp reviews actually refer to the business that is being reviewed. This baseline method tells us how many of these mentions there are in Yelp-EL.

ELT collectively links a set of mentions with an objective to maximize local compatibility and global consistence. It achieves this by integrating three types of similarities: mention-entity similarity, entity-entity similarity, and mention-mention similarity. To apply ELT to Yelp-EL, we use the conventional features introduced in Section 4.1 for mention-entity similarities. The path count feature of meta-path B–R–U–R–B is used as entity-entity similarity. For mention-mention similarity, we use two features that are both TF-IDF based cosine similarities, with one between the two mention strings and the other between the reviews that the two mentions belong to.

SSRegu is also a collective approach. It is a graph regularization model that incorporates both local and global evidence through three principals: local compatibility, coreference, and semantic relatedness. SSRegu computes a weight matrix for each of these three principals, and then forms a graph based on the weight matrices and performs graph regularization to rank candidate entities. To apply SSRegu, we need to compute three weight matrices. The weight matrix for local compatibility is based on features extracted from the mention and the candidate entity. In our case, the conventional features are used for computing this matrix. Computing the coreference weight matrix requires to determine whether two corresponding mentions are coreferential. Huang et al. (2014) assume two mentions to be coreferential if their mention strings are the same and there exists at least one meta-path instance of specific patterns between them. In our case, the meta-paths used are M–R–M and M–R–U–R–M. To compute the semantic relatedness weight matrix, we apply the entity-entity similarity used for ELT. Note that SSRegu is a semi-supervised approach and is capable of using unsupervised data, but for fair comparison we do not use this feature here.

Method	Accuracy (mean±std)
DirectLink	0.6684±0.008
ELT	0.8451±0.012
SSRegu	0.7970±0.013
LinkYelp	**0.9034±0.014**

Table 2: Entity linking performance of different methods on Yelp-EL.

ELT and SSRegu are originally proposed to tackle the problem of entity linking for tweets, but their linking target is Wikipedia. Evaluating the performance of these two methods on Yelp-EL shows the difference between their problem and ours.

5.2 Experimental Settings

Throughout our experiments, the hyperparameter λ in Equation (3) is set to 0.001. For each mention, three corrupted samples are random selected for the training with Equation (3). For ELT and SSRegu, the hyperparameters are tuned on the validation set with grid search. The candidate businesses for ELT and SSRegu are also obtained with the method describe in Section 3.1. We run five trials of random split of the linked mentions in the dataset, where each trail uses 60% of the linked mentions as training data and 40% as test data. In each of the training set, we further select 20% as validation set.

Note that we only use linked mentions to evaluate different methods since NIL detection is not our focus, but NIL mentions are utilized in Section 5.6 to build a complete entity linking system.

5.3 Comparison Results

Table 2 shows the entity linking performance of different methods on Yelp-EL. Here, all three types of features described in Section 4 are fed into LinkYelp. Within the compared methods, LinkYelp performs substantially better. This shows that methods carefully designed for traditional entity linking problems may not work so well when applied to entity linking within a social media platform, and this new problem we pro-

Features	All	Restaurants	Nightlife	Shopping	Food	A & E	Bars	E & S	H & T
C	84.05	86.80	79.18	83.15	81.15	71.67	83.49	68.79	63.70
S	79.65	82.92	75.09	82.05	77.69	66.95	80.66	72.25	67.41
L	80.05	85.71	72.70	81.32	85.77	50.21	82.55	49.13	37.78
C+S	86.45	89.29	84.98	85.71	85.00	78.54	87.26	**77.46**	**71.85**
C+L	89.42	93.32	84.98	90.48	88.46	76.39	88.21	72.83	66.67
S+L	85.19	89.29	79.86	87.91	87.31	72.10	85.38	72.25	67.41
C+S+L	**90.34**	**93.79**	**85.67**	**91.94**	**90.00**	**78.97**	**88.68**	76.30	71.11

Table 3: Entity linking accuracy (%) on different categories of businesses with different types of features as input. On the "Features" column, "C," "S," and "L" means conventional, social, and location features respectively. "All" means all the categories combined, i.e., the whole test set; "A & E" means Arts & Entertainment; "E & S" means Event Planning & Services; "H & T" means Hotels & Travel.

pose is worth studying differently from the traditional entity linking problem. The accuracy of DirectLink means that many mentions (about 67%) in Yelp-EL simply refer to the corresponding reviewed businesses. However, this does not mean that our problem is less challenging than traditional entity linking, since simply using the popularity measure of entities can achieve an accuracy of about 82% in the latter task (Pan et al., 2015).

5.4 Ablation Study

We further investigate how the three different types of features described in Section 4 contribute to the final performance of LinkYelp, and how they perform differently in linking mentions that refer to a specific category of Yelp businesses. The results are listed in Table 3. The categories in Table 3 are those that include the largest numbers of businesses in the dataset. Entries in the "All" column in Table 3 are the accuracies on all the categories combined. We can see from this column that both social features and location features are able to improve the performance when combined with conventional features. Location features are relatively more effective than social features, this is because people's activities are mainly restricted to a certain area, so they are more likely to mention businesses that are within this area. But social features are still helpful even when both conventional features and location features are already used, as the best performance is achieved with all the three types of features combined. Moreover, social features can become more important for other social media platforms that do not have location information available.

There are also some interesting findings if we consider the performance on different categories. For example, compared with only using conventional features (row C), incorporating social features (row C+S) provides the largest improvement for Event Planing & Services (e.g., wedding planning, party planning). This matches our intuition because for these kinds of businesses, people tend to be influenced more by their friends and make choices that are socially related. Table 3 also shows that on the categories Event Planning & Services and Hotels & Travel, incorporating location features is not that helpful as it does on other categories. We manually checked the mentions under these two categories that are linked correctly by C+S but incorrectly by C+L. We find that the reasons why incorporating location features fails on these mentions vary from case to case. Two possible reasons are: location information is not helpful to disambiguate a hotel and the shops in this hotel; it also does not work well in disambiguating different locations of a hotel chain that are all not far away from the reviewed business.

5.5 Error Examples

We also manually checked some of the errors made by LinkYelp with all the three types of features as input. A few examples are shown in Table 4. In the first case, since the reviewed business "Jean Philippe Patisserie" is a restaurant, our system tends to find a similar business instead of a hotel. Location features do not help here because Cafe Bellagio has the same location as Bellagio Hotel. The system is also incapable of identifying that "stay at" should be probably followed by a hotel instead of a Cafe. In the second case, the algorithm outputs the reviewed business because it is unable to understand what "the other Second Sole in Rock River" means. The above two examples show that there are still some errors caused by the failure of natural language understanding.

Reviewed Biz:	*Name*: Jean Philippe Patisserie *Addr.*: 3600 S Las Vegas Blvd, Las Vegas
Review:	... Even if you are not staying at the Bellagio, you have to stop by anyway to ...
True Referent:	*Name*: Bellagio Hotel *Addr.*: 3600 S Las Vegas Blvd, Las Vegas
System Prediction:	*Name*: Cafe Bellagio *Addr.*: 3600 S Las Vegas Blvd, Las Vegas
Reviewed Biz:	*Name*: Second Sole Athletic Footwear *Addr.*: 5114 Mayfield Rd, Cleveland
Review:	I did a review of the other Second Sole in Rocky River. This one is in Lyndhurst...
True Referent:	*Name*: Second Sole *Addr.*: 19341 Detroit Rd, Rocky River
System Prediction:	*Name*: Second Sole Athletic Footwear *Addr.*: 5114 Mayfield Rd, Cleveland
Reviewed Biz:	*Name*: Hoot Owl *Addr.*: 4361 W Bell Rd, Phoenix
Review:	This place is a really fun neighborhood bar ... Its tucked away in the Frys parking lot...
True Referent:	*Name*: Fry's Food and Drug *Addr.*: 4315 W Bell Road, Phoenix
System Prediction:	*Name*: Frys *Addr.*: 2626 S 83rd Ave, Phoenix

Table 4: Examples of errors made by LinkYelp. Business mentions are underlined.

Reviewed Biz	Sentence	Mentioned Biz
Name: Burger King *Addr.*: 1194 King St W, Toronto	When you compare it to the McDonald's across the street, the service is way better.	*Name*: McDonald's *Addr.*: 1221 King Street W, Toronto
Name: The Turf Public House *Addr.*: 705 N 1st St, Phoenix	I have to say I like the atmosphere and surroundings of the Turf better than Seamus.	*Name*: Seamus McCaffrey's *Addr.*: 18 W Monroe St, Phoenix

Table 5: Examples of comparative sentences and linked mentions.

Name	City	Stars	#Better
Bacchanal Buffet	Las Vegas	4.0	33
Wicked Spoon	Las Vegas	3.5	24
Cibo	Phoenix	4.5	9
Pizzeria Bianco	Phoenix	4.0	2
XS Nightclub	Las Vegas	4.0	8
Marquee	Las Vegas	3.5	1

Table 6: Comparative study using texts and average ratings. Each row is a pair of two frequently compared businesses. #Better means the number of sentences that claim the corresponding business to be better than the other one.

In the third case, "Fry's Food and Drug" is located at "4315 W Bell Road, Phoenix" which is nearer to the reviewed business "Hoot Owl" located at "4361 W Bell Rd, Phoenix." However, although location information favors the correct business, the others features may contribute more for the system output "Frys" since "Frys" has an exact match of the candidate mention name.

5.6 Comparative Study

In this study, we provide some insight on the possible applications of our task by checking the comparative sentences in Yelp reviews.

First, we find comparative sentences from the whole Yelp review dataset with a simple pattern matching method: we retrieve the sentences that contain one of eight predefined comparison phrases such as "is better," "not as good as," etc. Then we extract the named entity mentions within these sentences and link them to Yelp businesses. A threshold based approach is used to detect NIL mentions (Dalton and Dietz, 2013).

As a result, we get 12,149 comparative sentences from the total 4,153,150 reviews that contains at least one linked mention. Some of the results are shown in Table 5. We can successfully identify both the entity names and their locations on Yelp. We also selected the top three frequently compared pairs and compare with the stars provided by Yelp dataset. From Table 6 we can see that the text comparison is consistent with star ratings.

6 Related Work

The traditional entity linking task of mapping mentions in articles to their corresponding entities in a knowledge base has been studied extensively (Shen et al., 2015; Ling et al., 2015). Various kinds of methods have been studied, e.g., neural network models (Sun et al., 2015; He et al., 2013), generative models (Li et al., 2013), etc. A large group of the existing entity linking approaches are called collective approaches, which are based on the observation that the entities mentioned in a same context are usually related with each other. Thus they usually form entity linking as an optimization problem that tries to maximizes both local mention-entity compatibility and global entity-entity coherence (Han et al., 2011; Nguyen et al., 2016). LinkYelp does not consider global

entity-entity coherence as it is not the focus of this paper, but it can be applied to our problem too.

The prevalence of on-line social networks has also motivated researchers to study entity linking in such environments. (Huang et al., 2014; Shen et al., 2013; Liu et al., 2013) proposed methods that are specially designed for linking named entities in tweets. They mainly address the problem that tweets are usually short and informal, while taking advantage of some of the extra information that tweets may provide. For example, (Shen et al., 2013) assumed that each user's tweets have an underlying interest distribution and proposed a graph based interest propagation algorithm to rank the entities. (Huang et al., 2014) also used meta-path on HIN in their entity linking approach, but they only used it to get an indication of whether two mentions are related. Finally, although these studies focused on entity linking for tweets, they still use entities in knowledge bases as the target.

There are a few entity linking studies that do not link mentions to knowledge bases. (Shen et al., 2017) proposed to link entity mentions to an HIN such as DBLP and IMDB. However, their articles are collected from the Internet through searching and thus are not related to the target entities. They also used an HIN based method, but their use is restricted to get the relatedness between different entities. (Lin et al., 2017) studied the entity linking problem where the entities are included in different lists and entities of the same type belong to the same list. They only used this information along with the name of each entity to perform entity linking. Thus their focus is very different from ours.

7 Conclusions

In this paper, we propose a new entity linking problem where both entity mentions and target entities are in a same social media platform. To study this problem, we first create a dataset called Yelp-EL, and then conduct extensive experiments and analysis on it with a learning to rank model that takes three different types of features as input. Through the experimental results, we find that traditional entity linking approaches may not work so well on our problem. The two types of features that are usually not available for traditional entity linking tasks – social features and location features – can both improve the performance significantly on Yelp-EL. Our work can also motivate and enable a lot of downstream applications such as comparative analysis of location based businesses. In the future, we plan to extract more patterns to obtain more comparative sentences, so that we may more accurately demonstrate how useful performing comparative analysis after linking the business mentions can be.

Acknowledgments

This paper was supported by the Early Career Scheme (ECS, No. 26206717) from Research Grants Council in Hong Kong. The experiments and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the affiliated company.

References

Silviu Cucerzan. 2007. Large-scale named entity disambiguation based on wikipedia data. *Proceedings of EMNLP-CoNLL*, page 708.

Jeffrey Dalton and Laura Dietz. 2013. A neighborhood relevance model for entity linking. In *Proceedings of the 10th Conference on Open Research Areas in Information Retrieval*, pages 149–156.

Jenny Rose Finkel, Trond Grenager, and Christopher Manning. 2005. Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of ACL*, pages 363–370.

Matthew Francis-Landau, Greg Durrett, and Dan Klein. 2016. Capturing semantic similarity for entity linking with convolutional neural networks. In *Proceedings of NAACL*, pages 1256–1261.

Octavian-Eugen Ganea, Marina Ganea, Aurelien Lucchi, Carsten Eickhoff, and Thomas Hofmann. 2016. Probabilistic bag-of-hyperlinks model for entity linking. In *Proceedings of WWW*, pages 927–938.

Amir Globerson, Nevena Lazic, Soumen Chakrabarti, Amarnag Subramanya, Michael Ringaard, and Fernando Pereira. 2016. Collective entity resolution with multi-focal attention. In *Proceedings of ACL*, volume 1, pages 621–631.

Zhaochen Guo and Denilson Barbosa. 2014. Robust entity linking via random walks. In *Proceedings of CIKM*, pages 499–508.

Nitish Gupta, Sameer Singh, and Dan Roth. 2017. Entity linking via joint encoding of types, descriptions, and context. In *Proceedings of EMNLP*, pages 2681–2690.

Xianpei Han, Le Sun, and Jun Zhao. 2011. Collective entity linking in web text: a graph-based method. In *Proceedings of SIGIR*, pages 765–774.

Zhengyan He, Shujie Liu, Mu Li, Ming Zhou, Longkai Zhang, and Houfeng Wang. 2013. Learning entity representation for entity disambiguation. In *Proceedings of ACL*, pages 30–34.

Hongzhao Huang, Yunbo Cao, Xiaojiang Huang, Heng Ji, and Chin-Yew Lin. 2014. Collective tweet wikification based on semi-supervised graph regularization. In *Proceedings of ACL*, pages 380–390.

Heng Ji, Joel Nothman, Hoa Trang Dang, and Sydney Informatics Hub. 2016. Overview of tac-kbp2016 tri-lingual edl and its impact on end-to-end cold-start kbp. *Proceedings of of TAC*.

Yang Li, Chi Wang, Fangqiu Han, Jiawei Han, Dan Roth, and Xifeng Yan. 2013. Mining evidences for named entity disambiguation. In *Proceedings of KDD*, pages 1070–1078.

Ying Lin, Chin-Yew Lin, and Heng Ji. 2017. List-only entity linking. In *Proceedings of ACL*, volume 2, pages 536–541.

Xiao Ling, Sameer Singh, and Daniel S Weld. 2015. Design challenges for entity linking. *TACL*, 3:315–328.

Xiaohua Liu, Yitong Li, Haocheng Wu, Ming Zhou, Furu Wei, and Yi Lu. 2013. Entity linking for tweets. In *Proceedings of ACL*, pages 1304–1311.

Thien Huu Nguyen, Nicolas Fauceglia, Mariano Rodriguez Muro, Oktie Hassanzadeh, Alfio Massimiliano Gliozzo, and Mohammad Sadoghi. 2016. Joint learning of local and global features for entity linking via neural networks. In *Proceedings of COLING*, pages 2310–2320.

Xiaoman Pan, Taylor Cassidy, Ulf Hermjakob, Heng Ji, and Kevin Knight. 2015. Unsupervised entity linking with abstract meaning representation. In *Proceedings of NAACL-HLT*, pages 1130–1139.

Wei Shen, Jiawei Han, Jianyong Wang, Xiaojie Yuan, and Zhenglu Yang. 2017. Shine+: A general framework for domain-specific entity linking with heterogeneous information networks. *IEEE Trans. on Knowl. and Data Eng.*

Wei Shen, Jianyong Wang, and Jiawei Han. 2015. Entity linking with a knowledge base: Issues, techniques, and solutions. *IEEE Trans. Knowl. Data Eng.*, 27(2):443–460.

Wei Shen, Jianyong Wang, Ping Luo, and Min Wang. 2013. Linking named entities in tweets with knowledge base via user interest modeling. In *Proceedings of KDD*, pages 68–76.

Yaming Sun, Lei Lin, Duyu Tang, Nan Yang, Zhenzhou Ji, and Xiaolong Wang. 2015. Modeling mention, context and entity with neural networks for entity disambiguation. In *Proceedings of IJCAI*, pages 1333–1339.

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

Chuanqi Tan, Furu Wei, Pengjie Ren, Weifeng Lv, and Ming Zhou. 2017. Entity linking for queries by searching wikipedia sentences. In *Proceedings of EMNLP*, pages 68–77.