Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence (AAAI-16)

Understanding Emerging Spatial Entities

Jinyoung Yeo, Jin-woo Park
Pohang University of Science and Technology (POSTECH)
{jinyeo, jwpark85}@postech.ac.kr

Seung-won Hwang
Yonsei University
seungwonh@yonsei.ac.kr

Abstract

In Foursquare or Google+ Local, emerging spatial entities, such as new business or venue, are reported to grow by 1% every day. As information on such spatial entities is initially limited (*e.g.*, only name), we need to quickly harvest related information from social media such as Flickr photos. Especially, achieving high-recall in photo population is essential for emerging spatial entities, which suffer from data sparseness (*e.g.*, 71% restaurants of TripAdvisor in Seattle do not have any photo, as of Sep 03, 2015). Our goal is thus to address this limitation by identifying effective linking techniques for emerging spatial entities and photos. Compared with state-of-the-art baselines, our proposed approach improves recall and F1 score by up to 24% and 18%, respectively. To show the effectiveness and robustness of our approach, we have conducted extensive experiments in three different cities, Seattle, Washington D.C., and Taipei, of varying characteristics such as geographical density and language.

Introduction

New business and venue open daily, as reported to be 1% of knowledge bases (KBs) on spatial entities (SEs) such as Foursquare and Google+ Local. However, KB pages of a new place are often non-existent or near-empty (only name and location are known). Thus, users cannot obtain information about new places, until expert or non-expert volunteers visit the places and annotate their KB pages, which may take from a few days to a few years. Meanwhile, such new arrivals create attention rather immediately in social media that users post image and text to share their experiences and interests, which can be considered as a passive annotation.

Our research question is thus to automatically populate empty KB pages by such photos uploaded on social media sites. A naive idea is querying SE names (*e.g.*, 'McCaw Hall Prelude') to Flickr (Wang et al. 2012), using location as an additional feature to disambiguate (Packer et al. 2012). This method is highly precise but suffers from low recall. In contrast, we propose high-recall approaches, tackling the following two causes of low recall.

- **[C1] Tag sparsity**: Web photos may be poorly tagged or often have no "identifying" tag such as SE name.
- **[C2] Vocabulary mismatch**: People tend to use various "synonymous" tags, which refer to the same SE but cannot be matched with its queried official SE name.

An existing solution for improving recall is clustering duplicate or near-duplicate photos to collect missing or synonymous tags (which we use as a baseline in this work). We propose a more systematic way to tackle the aforementioned two challenges and achieve significantly higher recall over baselines (up to 24.3% in Table 2). This recall gain is critical in many applications– F1 score of SE type categorization (Srihari, Niu, and Li 2000; Wang et al. 2012) using photos is improved by 7.7% (Table 7).

In particular, we address **C1** and **C2** in the following three novel ways:

First, we infer missing tags by propagating tags among duplicate/near-duplicate "photo clusters" referring to the same SE. To infer such relations, existing related methods adopt geo-spatial (Crandall et al. 2009), visual (Zheng et al. 2009), and/or textual (Belém et al. 2011) signals among single photos. We later emprically show that our systematic aggregation of these three signals is significantly superior to a naive combination baseline (Figure 1).

Second, we mine a large Flickr corpus to obtain high-quality synonyms. Existing related methods (Cheng, Lauw, and Paparizos 2012; Jiang et al. 2013) leverage term co-occurrence patterns in text documents, as an evidence of synonymous term pairs. In contrast, we use image and location signals, which can be mined from photos shooting the same SE, to extract higher-quality synonyms in user-generated tags. We show that aggregating these signals is effective to expand SE names for photo population (Table 3 and Figure 2).

Lastly, we observe the mutually reinforcing nature of the above two approaches to combine them into an iterative framework for photo population. The key contributions of this framework is to improve trade-off between precision and recall (Figure 3) and show consistent performance in cities of varying geographic density of SEs, language, and popularity of SEs (Table 4-6 and Figure 4).

Preliminaries

Problem Statement

Our goal is to link photos with their referent SE to populate KB pages. In this work, SEs refer to places, such as attractions or restaurants, in a city. Coarser granular SEs, such as cities, are not considered because they are rarely emerging and can often be trivially linked due to broad coverage.

Our problem formulation focuses on supporting emerging SE. An emerging SE can be formed as a tuple that has no value for every attribute, such as name, type, location, and background text, because emerging SEs are generally defined as SEs out of KBs (Hoffart, Altun, and Weikum 2014). However, in this paper, we make a minimal assumption on an emerging SE– only an official name n_e and (optional) location $l_e = \{lat, lon\}$ of an SE e are known. We argue this assumption is realistic based on the following two observations: We could easily discover such emerging SE tuples by adopting a simple heuristics of (Gao et al. 2010). This approach achieves 85% precision and 77% recall in discovering SE names of TripAdvisor in Seattle. Alternatively, online map services cover a lot of emerging SEs $\langle n_e, l_e \rangle$, which can be considered as a new KB. For example, in the case of park in Seattle, Google map covers hundreds of parks while Wikipedia covers only around 80 parks.

Formally, as an output, we augment an emerging SE $e = \langle n_e, l_e \rangle$ into $\langle N_e, l_e, I_e \rangle$, where N_e is a set of synonymous names of e and I_e is a set of relevant photos of e.

Solution Framework

We introduce the framework of our photo population system. First, we collect photos and their metadata ⟨user ID, image, location (*lat* and *lon*), tags (only if exists)⟩, which are generally available in most photo-sharing sites such as Flickr. These photos are both geographically and visually grouped into duplicate/near-duplicate photo clusters (clustering precision is over 98%) (Zheng et al. 2009). Given a photo cluster p located in a city, finding its most relevant SE e^* in all SEs E of the city is to find a photo set I_{e^*} of e^* as follows:

$$\begin{aligned} e^* &= argmax_{e \in E} P(e, p) \\ &= argmax_{e \in E} P(n_e, l_e, p) \end{aligned} \tag{1}$$

Intuitively, the expression (n_e, l_e) of a given SE e is proper, as mentioned in problem statement, to aim at emerging SEs.

C1: Spatio-textual Photo Linking

To find the most relevant SE e^* of a given photo cluster p, the goal of the first component is to compute $P(n_e, l_e, p)$. We decompose the model below, to simplify the estimation of probability values.

$$\begin{aligned} P(n_e, l_e, p) &= P(n_e|l_e, p) P(l_e, p) \\ &= P(n_e|p) P(l_e, p) \end{aligned} \tag{2}$$

Similar to (Fang and Chang 2014), we assume that, given an SE e, how e is expressed (n_e) and where e is located (l_e) are conditionally independent. In other words, we have $P(n_e|l_e, p) = P(n_e|p)$. Thus, computing $P(n_e, l_e, p)$ boils down to:

- $P(l_e, p)$ of representing the geographical proximity between e and p. $P(l_e, p)$ can be computed as distance-based probability between a single location (l_e if exists) and a set of locations (of photos in p) by using Gaussian mixture model (Li and King 1999). In case that every l_e is unknown, we set $P(l_e, p)$ as 1.
- $P(n_e|p)$ of representing the textual proximity between e and p. We now discuss the estimation of $P(n_e|p)$ in detail.

Textual Signal Estimation

Given a photo cluster p, we compute $P(n_e|p)$ which represents how relevant p is to n_e. Intuitively, the probability is estimated by matching an aggregated tag set of p with n_e. Such approach is likely to exclude photo clusters not annotated with SE names. To loosen it, a photo cluster p can be matched with an SE e, even though it is not annotated with its name, if p can be matched with another photo cluster p_i with such annotation. To decide whether two clusters p and p_i refer to the same SE, $P(n_e|p)$ is obtained from a pseudo-generative model using Bayes' Rule. That is, given two photo cluster p and p_i, we combine the two clues, $P(p, p_i)$ representing their textual similarity and $P(n_e|p_i)$ of reliability of p_i for representing an SE e as follows:

$$P(n_e|p) = \sum_{\forall i} P(p_i|p) P(n_e|p_i) \tag{3}$$

Strictly speaking, neither the generative process from p to p_i nor the generative model from p_i to n_e are known or defined precisely, hence the above conditional probabilities cannot be known exactly. However, we are not interested in probabilities "per-se", but rather in probability values as indicators used eventually for linking decision in Eq. 1. For this reason, we can use proxy quantities – respectively $P(p_i|p)$ and $P(n_e|p_i)$ – which are presented as below.

The term $P(p_i|p)$ represents the probability of generating contents (*i.e.*, textual tags) of a photo cluster p_i from contents of a given photo cluster p. To estimate $P(p_i|p)$, we compute the cosine similarity of the cluster pair based on the Bag-of-Words model:

$$Sim(p_i, p) = \frac{T_{p_i} \cdot T_p}{|T_{p_i}||T_p|} \tag{4}$$

where T_p is a set of tags collected from a photo cluster p. To denoise, all tags (geographically and visually aggregated) in T_p are weighted by term frequency-inverse document frequency (TFIDF), as we shall discuss later. Now a proxy of $P(p_i|p)$ can be obtained by normalizing the content similarity between p_i and p:

$$P(p_i|p) = \frac{Sim(p_i, p)}{\sum_{\forall j} Sim(p_j, p)} \tag{5}$$

The term $P(n_e|p_i)$ can be interpreted as an indicator to how reliably a photo cluster p_i represents an SE e. We directly derive the proxy value for this term using a simple frequency-based approach as follows:

$$P(n_e|p_i) = \frac{|n_e \cap T_{p_i}|}{\sum_{e' \in E} |n_{e'} \cap T_{p_i}|} \tag{6}$$

C2: SE Name Expansion

We now discuss how to obtain a high-quality name set N_e of semantically equivalent terms (*i.e.*, synonyms) of a given SE e. We consider all photo tags in I_e as synonym candidates. To estimate the likelihood that a candidate tag t is an SE synonym of n_e, we measure how specific t is to a location and how relevant t is to n_e based on geo-spatial and image signals, respectively, that can be captured from photo corpus. As such a synonym scoring measure, we output $Score(e,t)$ (further normalized in the interval [0,1]) as follows:

$$Score(e,t) = \lambda \cdot \vec{w}_{geo} \cdot \vec{f}_{geo}(t) + (1-\lambda) \cdot \vec{w}_{img} \cdot \vec{f}_{img}(e,t) \quad (7)$$

where:

- $\vec{f}_{geo}(t)$ is the feature vector that models the geo-specificity of tag t;
- $\vec{f}_{img}(e,t)$ is the feature vector that models the image relevance between tag t and SE e;
- $\vec{w}_{geo}$ and $\vec{w}_{img}$ are the feature weight vector related to $\vec{f}_{geo}(t)$ and $\vec{f}_{img}(e,t)$, respectively, which are trained using structural SVM on the training data set;
- $\lambda \in (0,1)$ is a systematic parameter, which is determined using the quality function (F1 score in Eq. 9) on the training data set; it is used to adjust to tradeoff between $\vec{f}_{geo}(t)$ and $\vec{f}_{img}(e,t)$. It is experimentally set to 0.5 in our work.

Geo-spatial Features

To quantify the spatiality of candidate tags, an external geo-database identifies geo-terms in candidate tags, and cooperatively the identified tags are rated according to the geographic distribution of tagged photos.

- Geoterm Database: In Microsoft Bing services, Geocode Dataflow API (GDA) can query large numbers of geo-terms and their representative locations. Querying a tag t to GDA, we set $f^1_{geo}(t)$ as 1 if its location is returned, and 0 otherwise.
- Geographic Variation: Not all geo-terms found by GDA are desirable candidates for SE synonyms. Not only place-level geo-terms such as 'Space Needle', there are also country- or city-level geo-terms such as 'USA' and 'Seattle', which are far more widely distributed. Thus, we compute the variance $Var(t)$ of coordinates set of photos including a tag t, then set $f^2_{geo}(t) = exp(-Var(t))$. We set $Var(t)$ as infinite if there exists only one photo with t.

Image Features

To quantify the image relevance between a tag and an SE, we extend the intuition of TFIDF. However, our unique contribution is to define a "document" as an estimated set I_e (*e.g.*, by a query n_e) of all photos on the same SE e. Using this document, the following frequency features convey the synonym evidence in both TF and IDF.

- Photo Frequency: A candidate tag is likely to be one of SE names as more photos of an SE have the tag. Given a tag t and an SE e, we set $f^1_{img}(e,t)$ as the number of photos that have t in I_e.
- User Frequency: When uploading a photo collection, "lazy users" often tend to copy the same tag set to all photos, which may cause overestimation of photo frequency. To prevent this, we set $f^2_{img}(e,t)$ as the number of users that assign t to photos in I_e.
- Inverse SE Frequency: We need to filter popular non-name tags, such as 'Travel', which have high photo and user frequencies. As a penalty for such tags, we use *SE frequency*$(e \in E, t)$, which is defined as the number of SEs $E' \subseteq E$ having any photo with t in $I_{e' \in E'}$. Thus, we set $f^3_{img}(e,t) = \frac{|E|}{SE\, frequency(e,t)}$.

Combining C1 and C2

We embed two components, C1 of computing $P(n_e, l_e, p)$ and C2 of extracting N_e, into a unified framework for populating I_e. Two components are mutually dependent and can reinforce each other. First, C1 collects a new photo set I_e, which updates synonym candidates and image feature values for name expansion. Reversely, C2 extracts a new synonym set N_e, which updates SE names (initially n_e) and textual probability for photo linking. Thus, we consider an iterative process of alternating C1 and C2.

Formally, we reinforce the linking probability at each time ϕ, *i.e.*, $P({N_e}^{\phi}, l_e, p)^{\phi} \rightarrow P({N_e}^{\phi+1}, l_e, p)^{\phi+1}$, being repeated until convergence $(*)$:

$$P(e,p) = P(n_e, l_e, p) \approx P({N_e}^{*}, l_e, p)^{*} \quad (8)$$

In the view of SE tuples, one iteration updates an SE tuple $\langle N_e, I_e \rangle^{\phi}$ at time ϕ to an SE tuple $\langle N_e, I_e \rangle^{\phi+1}$ at time $\phi+1$. This procedure is repeated until all SE tuples at time ϕ are equivalent to the SE tuples at time $\phi+1$.

Experimental Evaluation

Settings

Datasets and ground truth To validate the robustness of our proposed linking system, we select three cities of varying characteristics– Washington D.C. with skewed SEs (near National Mall area), Seattle with less skewed SEs, and Taipei with tags in Chinese. We use Flickr API to collect photo tuples, which amount to 21,676 for Seattle; 19,995 for Washington D.C.; and 165,057 for Taipei. Note that such photos are collected only by using the bounding box of city area, which is a publicly available approach.

For experiments on emerging SEs defined in problem statement, we limit information on SE tuples to only an name n_e and representative geo-coordinates l_e. To find such $\langle n_e, l_e \rangle$ pairs, we first collect all SEs in the three cities from TripAdvisor, which amount to 4,417 for Seattle; 3,513 for Washington D.C.; and 10,840 for Taipei. Among the SEs, we select all SEs appearing in our photo datasets, such that their official names are tagged to photos by at least two different users.

Table 1 presents statistics on the discovered SEs with our photo datasets. As ground truth, two volunteers manually label I_e from our photo datasets and find N_e from tags assigned in I_e, and then cross-check the labeling result. We can observe SE distributions are more skewed in some area

Table 1: Statistics on ground truth

	Seattle	Washington D.C.	Taipei
Name language	English	English	Chinese
#SEs	80	56	64
#SE-labeled photos	6,038	8,412	12,463
Avg.#SE synonyms	2.9	3.7	3.8
Avg. SE pair distance	5.7 km	2.1 km	7.5 km

in Washington D.C. (average SE pair distance is smaller as 2.1 km), compared to Seattle (5.7 km) or Taipei (7.5 km).

Evaluation measure As evaluation metrics, we adopt precision, recall, and their F1 score. Let I_e^{GT} be a photo set of the ground truth for an SE e, I_e^{Sys} be a photo set estimated by a linking system (Sys) for e, and E be a set of SEs in a city, then these evaluation metrics are computed as:

$$\begin{aligned} Precision =& \frac{1}{|E|} \sum_{e \in E} \frac{|I_e^{GT} \cap I_e^{Sys}|}{|I_e^{Sys}|} \\ Recall =& \frac{1}{|E|} \sum_{e \in E} \frac{|I_e^{GT} \cap I_e^{Sys}|}{|I_e^{GT}|} \end{aligned} \tag{9}$$

We adopt a 3-fold cross validation by randomly partitioning the ground truth into three similarly sized groups.

Baselines We consider the following one high-precision and two high-recall baseline linking systems to evaluate our linking systems (denoted as **Ours**, **Ours1**, and **Ours2**).

- **CLU**: A state-of-the-art of linking photos to KB entry (Wang et al. 2012). We cluster duplicate or near-duplicate photos for SEs, as a preprocessing, by using geo-spatial and visual clustering techniques (Zheng et al. 2009). In the same manner as Eq. 1, CLU links a photo cluster to its most relevant SE to be matched with aggregated tag and location sets of the photo cluster.
- **CLU+REC**: To overcome tag sparsity, REC (Belém et al. 2011), a state-of-the-art tag recommender, finds relevant tags to be added to photo tuples before performing CLU. The key evidence of tag recommendation is tag co-occurrence patterns to find textually similar photos of a given query photo.
- **CLU+SYN**: SYN (Cheng, Lauw, and Paparizos 2012), a state-of-the-art synonym extractor, addresses vocabulary mismatch by expanding SE tuples with synonyms before performing CLU. SYN uses its original scoring measures but adopts photos having any tag as text documents.

Results and Discussion

We present the empirical findings for the following research questions:

Q1: Does our system outperform baseline systems?
Q2: How does our system balance precision and recall?
Q3: Is our system robust in various factors?
Q4: Does our system indeed help machine understanding?

Overall performance (Q1). First, we compare the performance of linking systems (Ours and three baselines). In our target application of populating KB pages, we argue recall gain is crucial for emerging SEs suffering from extreme sparsity. That is, it is important to improve recall, maintaining high-precision. Thus, in the training step of all systems, we set the objective function to maximize F1 score while setting the minimum allowable precision to 0.8.

Table 2 shows precision, recall, and F1 score of the linking systems. We can see that Ours achieves a significant recall gain with marginal loss of precision, compared to the high-precision method, CLU, and thus shows higher F1 scores. Although other high-recall baselines, CLU+REC and CLU+SYN, also improve recall and F1 score, recall gain is marginal and often smaller than precision loss. In contrast, recall gain of Ours is (around three times) higher than precision loss of that consistently in three cities.

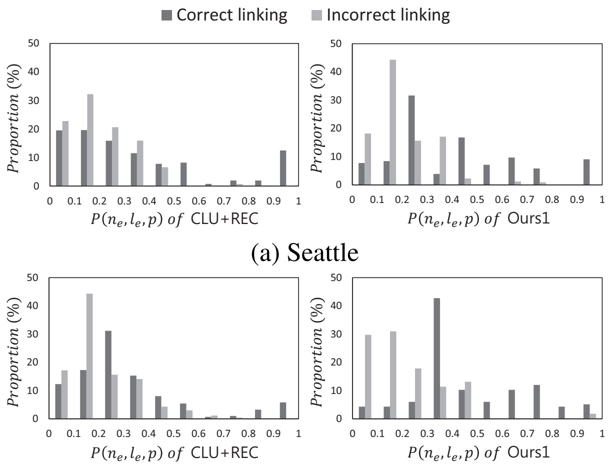
Component study (Q2). To see how our system achieves such improvement, this section breaks down Ours into two linking systems, Ours1 and Ours2, leveraging only a subcomponent, C1 or C2, respectively. We then discuss the effectiveness of each system and their combination. Due to lack of space, we cover only Seattle and Washington D.C.

First, to discuss how tag sparsity is addressed, we compare CLU+REC and Ours1, which considers only C1 but not C2. The difference between CLU+REC and Ours1 is how geo-spatial, visual, and textual signals among photos are aggregated to compute $P(n_e, l_e, p)$. CLU+REC performs tag recommendation (with textual signal) and photo clustering (with spatio-visual signal), but two signals are considered independent and thus performed relatively poorly in Figure 1: In this figure, we can see that Ours shows clearer distinction between correct linking (frequent in high-scoring region) and incorrect linking (frequent in low-scoring region) than CLU+REC. By cooperating multiple signals, Ours1 is more reliable in addressing tag sparsity than CLU+REC.

Second, to discuss how vocabulary mismatch is addressed, we compare CLU+SYN and Ours2, which is implemented considering only C2 but not C1. Specifically, Ours2 first adopts a naive approximation of I_e as a set of photo clusters having n_e to find N_e, then re-approximate I_e as

Table 2: Comparison of linking systems. The performance gap with CLU is presented in parentheses.

	Seattle			Washington D.C.			Taipei		
	Precision	Recall	F1 score	Precision	Recall	F1 score	Precision	Recall	F1 score
CLU	**.929**	.524	.670	**.945**	.384	.546	**.939**	.414	.575
CLU+REC	.915 (-.014)	.549 (+.025)	.687 (+.017)	.920 (-.025)	.435 (+.051)	.591 (+.045)	**.939** (-)	.425 (+.011)	.585 (+.010)
CLU+SYN	.899 (-.030)	.535 (+.011)	.671 (+.001)	.902 (-.043)	.414 (+.030)	.567 (+.021)	.929 (-.006)	.483 (+.069)	.636 (+.061)
Ours	.891 (-.038)	**.636** (+.112)	**.742** (+.072)	.864 (-.081)	**.627** (+.243)	**.726** (+.180)	.862 (-.073)	**.621** (+.207)	**.722** (+.147)



(a) Seattle

(b) Washington D.C.

Figure 1: Score distribution of correct and incorrect linking

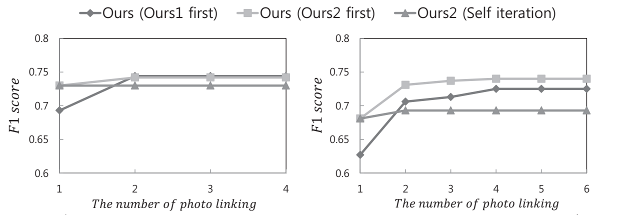
photo clusters having a tag $t \in N_e$ (such that $Score(e, t) \geq 0.8$). As a result, Table 3 reports that Ours2 outperforms naive fuzzy matching (Edit distance and common N-gram) and SYN in terms of synonym precision[1]. Figure 2 shows the linking performance of Ours2 over varying λ (linear combination weight for f_{geo} and f_{img}). While using geo-spatial features achieves high-precision due to accurate geo-term DB, using image features improves recall by discovering SE synonyms out of geo-term DB (*e.g.*, 'UW information school'). In this figure, around 0.5 is optimal for λ. This explains the complementary nature such that our combined approach, C2, outperforms using either geo-spatial features or image features ($\lambda = 1$ or 0).

Lastly, as shown in Figure 3, the F1 score of Ours is improved by alternating its two components C1 and C2 (*i.e.*, 1:Our1→ 2:Our2→ 3:Our1→...) and is higher than that of a self-iteration of Ours2 (*i.e.*, 1:Our2→ 2:Our2→...). This means that C1 and C2 are mutually reinforced to address tag sparsity and vocabulary mismatch together. As a result, Ours is converged to F1 score of Table 2 within the small number of SE-photo linking iterations.

Robustness analysis (Q3). This section presents the three factors influencing the performance over different datasets: Geographic density, language, and popularity.

As shown in Table 1, SEs in Washington D.C. are more skewed to a certain area than those in Seattle. Because skewed SEs cause ambiguous photo linking, as shown in Table 2, baselines in Washington D.C. thus suffer lower F1 score compared to those in Seattle. In contrast, Table 4 shows that performance gap between the two cities decreases in our approaches.

Another factor is dependence on the linguistic characteristics of tags. For example, synonym extraction using Edit distance (denoted as EditDist) is more effective in finding English synonyms. Table 5 shows the significant score gap differentiating English synonym pairs from other name pairs in Seattle and Washington D.C. Meanwhile, such difference is marginal in Taipei dataset of another language. This inef-



(a) Seattle

(b) Washington D.C.

Figure 3: Convergence of Ours

Table 3: Comparison of SE synonym extraction

	Seattle			Washington D.C.		
	SP@1	SP@2	SP@3	SP@1	SP@2	SP3
Fuzzy matching	0.33	0.27	0.22	0.58	0.45	0.36
SYN	0.91	0.55	0.38	**0.98**	0.50	0.41
Ours2	**0.94**	**0.69**	**0.55**	**0.98**	**0.87**	**0.72**

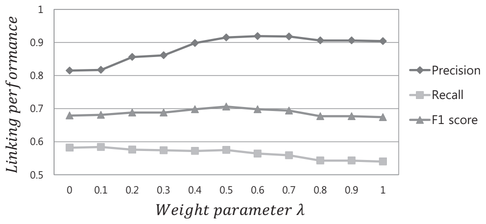


Figure 2: Influence of a parameter λ in Ours2. Performance values are averaged between Seattle and Washington D.C.

Table 4: Comparison of linking performance gaps between Seattle and Washington D.C.

	F1 score		
	Seattle	Washington D.C.	Gap
CLU	.670	.546	±.124
CLU+REC	.687	.591	±.096
CLU+SYN	.671	.567	±.104
Ours1	.693	.627	±.066
Ours2	.730	.681	±.049
Ours (converged)	.742	.726	±**.016**

Table 5: Average edit distance between synonyms of the same SE and between synonyms of different SEs

	Average Edit distance		
	Synonym pair	Other pair	Gap
Seattle	8.6	15.4	±6.8
Washington D.C.	9.5	18.6	±9.1
Taipei	3.5	5.9	±2.4

[1] Synonym precision (SP@K) represents ratio of ranking results such that the correct synonym is contained within Top-K tags.

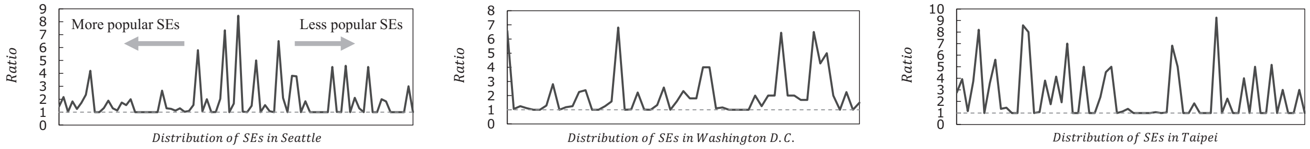


Figure 4: Ratio of the number of correctly linked photos between Ours and CLU. The dotted lines indicate where $Ratio = 1$.

Table 6: Gap of normalized synonym precision at Top-1 (SP@1) between English- and Chinese-named SEs

	Normalized SP@1		
	U.S. cities	Taipei	Gap
EditDist	.147	.104	±.043
Ours2	.331	.352	±**.021**

fectiveness affects synonym mining. Table 6 reports SP@1 normalized by the average number of synonyms of each set, of English- and Chinese-named SEs, respectively.[2] We can observe that normalized SP@1 of Chinese-named SEs suffers when EditDist is used. In contrast, Ours2 achieves comparable SP@1 for English- and Chinese-named SE.

Lastly, we should see the impact of photo population over SE popularity, as a desirable solution for emerging SEs should consistently outperform both before and after these SEs get noticed (*i.e.*, low and high populairy). We thus report how many more photos are linked by Ours compared to CLU according to the size of ground truth. In Figure 4, x-axis represents SEs sorted by $|I_e^{GT}|$ and y-axis represents the ratio[3] between the number of photos linked by Ours and CLU. That is, a dotted line ($Ratio = 1$) represents the performance of CLU and the solid line shows how much Ours outperforms– 2.1 times better than CLU on average in all cities. We stress that Ours consistently outperforms in all ranges– up to 4.5 (=$\frac{45}{10}$), 6.5 (=$\frac{13}{2}$), and, 5 (=$\frac{30}{6}$) times better than CLU in the tail range of each city, respectively.

Knowledge mining (Q4). This section discusses how our system contributes to mining other knowledge on emerging SEs. For example, we can consider a task of inferring SE type, such as Park, Cafe, and Restaurant. For each SE, we collect photos using either Ours or CLU, from which we compare the accuracy of predicted SE type. Formally, we collect a set J of 16 SE types frequently used in TripAdvisor and collect a set I_j of photos having each SE type term $j \in J$. Then, given an SE e and its populated photos I_e, we identify its SE type $e.j$ by maximizing cosine similarity $\cos(T_{I_j}, T_{I_e})$ where T_I is an TFIDF-weighted tag vector obtained from a photo set I (Wang et al. 2012).

Table 7 shows the performance of type inference for 80 SEs in Seattle. We can see that the results of Ours are better than those of CLU. This improvement is resulted from the better photo population (significantly increasing recall with marginal loss of precision) in Table 2.

Table 7: Performance of SE type inference in Seattle

	Type inference		
	Precision	Recall	F1 score
CLU	.774 (41/53)	.513 (41/80)	.617
Ours	**.781** (50/64)	**.625** (50/80)	**.694**

Related Work

Entity linking aims either to understand meanings of mentions (*e.g.*, 'Apple') by linking to KB (Liu et al. 2013; Shen et al. 2012; Han, Sun, and Zhao 2011) or to populate KB pages with information around mentions, such as photos (Wang et al. 2012; García-Silva et al. 2011; Taneva, Kacimi, and Weikum 2010). Existing photo linking systems for well-known entities leverage rich information on KB, such as entity type (Wang et al. 2012), context word (García-Silva et al. 2011), and synonym (Taneva, Kacimi, and Weikum 2010). However, it is difficult to apply this work to emerging entities, which have no such information on KBs (Hoffart, Altun, and Weikum 2014; Jin et al. 2014; Li and Sun 2014; Gao et al. 2010).

Recent projects propose to mine such information from plain text on the Web. For example, empty (or near empty) KB pages of emerging entities can be annotated by discovering entity-related context words (Taneva, Kacimi, and Weikum 2011), synonyms (Jiang et al. 2013; Cheng, Lauw, and Paparizos 2012; Chakrabarti et al. 2012), and acronyms (Zhang et al. 2011) from arbitrary Web documents. Similar techniques also have been applied to mine tags for photos without entity name, which is key evidence for linking (Belém et al. 2011; Sigurbjörnsson and Van Zwol 2008) – We observe that over 50% of photos are not tagged with relevant SE names.

To better apply above solutions to emerging SEs and their photos, location feature should be considered (Packer et al. 2012). We thus aggregate geographically and visually close photos (Crandall et al. 2009; Zheng et al. 2009) and their tags (Qian et al. 2013; Sergieh et al. 2012; Silva and Martins 2011) into one cluster to enrich linking evidence for SE. However, our experimental results show that various existing features as above are less effective and less robust in populating photos on emerging SEs, while our systematic aggregation significantly outperforms such baselines.

Conclusion

In this paper, we have presented how to harvest information from social media on emerging SEs. By aggregating location, image, and text evidences in a systematic way, our

[2] We observe that city with more synonyms is favored in SP@1.

[3] $Ratio = |I_e^{GT} \cap I_e^{Ours}| / |I_e^{GT} \cap I_e^{CLU}|$.

linking approach was effective in improving 24% in recall and 18% in F1 score from state-of-the-arts, and was robust in different cities of varying characteristics. We empirically showed that photos harvested by our approach are more helpful to understand emerging SEs by improving SE type inference. Mining more sophisticated knowledge beyond SE type is a promising future direction.

Acknowledgement

This work was supported by Institute for Information & communications Technology Promotion (IITP) grant funded by the Korea government (MSIP) (No. B0101-15-0307, Basic Software Research in Human-level Lifelong Machine Learning (Machine Learning Center)), and also by the MSIP (The Ministry of Science, ICT and Future Planning), Korea and Microsoft Research, under ICT/SW Creative research program supervised by the IITP (Institute for Information & Communications Technology Promotion) (IITP-2015-R2215-15-1009). S. Hwang is a corresponding author.

References

Belém, F.; Martins, E.; Pontes, T.; Almeida, J.; and Gonçalves, M. 2011. Associative tag recommendation exploiting multiple textual features. In *SIGIR*.

Chakrabarti, K.; Chaudhuri, S.; Cheng, T.; and Xin, D. 2012. A framework for robust discovery of entity synonyms. In *SIGKDD*.

Cheng, T.; Lauw, H. W.; and Paparizos, S. 2012. Entity synonyms for structured web search. *In TKDE*.

Crandall, D. J.; Backstrom, L.; Huttenlocher, D.; and Kleinberg, J. 2009. Mapping the world's photos. In *WWW*.

Fang, Y., and Chang, M.-W. 2014. Entity linking on microblogs with spatial and temporal signals. *In TACL*.

Gao, Y.; Tang, J.; Hong, R.; Dai, Q.; Chua, T.-S.; and Jain, R. 2010. W2go: a travel guidance system by automatic landmark ranking. In *MM*.

García-Silva, A.; Jakob, M.; Mendes, P. N.; and Bizer, C. 2011. Multipedia: enriching dbpedia with multimedia information. In *Proceedings of the sixth international conference on Knowledge capture*.

Han, X.; Sun, L.; and Zhao, J. 2011. Collective entity linking in web text: a graph-based method. In *SIGIR*.

Hoffart, J.; Altun, Y.; and Weikum, G. 2014. Discovering emerging entities with ambiguous names. In *WWW*.

Jiang, L.; Luo, P.; Wang, J.; Xiong, Y.; Lin, B.; Wang, M.; and An, N. 2013. Grias: an entity-relation graph based framework for discovering entity aliases. In *ICDM*.

Jin, Y.; Kıcıman, E.; Wang, K.; and Loynd, R. 2014. Entity linking at the tail: sparse signals, unknown entities, and phrase models. In *WSDM*.

Li, X., and King, I. 1999. Gaussian mixture distance for information retrieval. In *Neural Networks, 1999. IJCNN'99. International Joint Conference on*, volume 4, 2544–2549. IEEE.

Li, C., and Sun, A. 2014. Fine-grained location extraction from tweets with temporal awareness. In *SIGIR*.

Liu, X.; Li, Y.; Wu, H.; Zhou, M.; Wei, F.; and Lu, Y. 2013. Entity linking for tweets. In *ACL*.

Packer, H. S.; Hare, J. S.; Samangooei, S.; and Lewis, P. 2012. Semantically tagging images of landmarks. In *KECSM*.

Qian, X.; Liu, X.; Zheng, C.; Du, Y.; and Hou, X. 2013. Tagging photos using users' vocabularies. *Neurocomputing*.

Sergieh, H. M.; Gianini, G.; Döller, M.; Kosch, H.; Egyed-Zsigmond, E.; and Pinon, J.-M. 2012. Geo-based automatic image annotation. In *ICMR*.

Shen, W.; Wang, J.; Luo, P.; and Wang, M. 2012. Linden: linking named entities with knowledge base via semantic knowledge. In *WWW*.

Sigurbjörnsson, B., and Van Zwol, R. 2008. Flickr tag recommendation based on collective knowledge. In *WWW*.

Silva, A., and Martins, B. 2011. Tag recommendation for georeferenced photos. In *LBSN*.

Srihari, R.; Niu, C.; and Li, W. 2000. A hybrid approach for named entity and sub-type tagging. In *ACL*.

Taneva, B.; Kacimi, M.; and Weikum, G. 2010. Gathering and ranking photos of named entities with high precision, high recall, and diversity. In *WSDM*.

Taneva, B.; Kacimi, M.; and Weikum, G. 2011. Finding images of difficult entities in the long tail. In *CIKM*.

Wang, X.-J.; Xu, Z.; Zhang, L.; Liu, C.; and Rui, Y. 2012. Towards indexing representative images on the web. In *MM*.

Zhang, W.; Sim, Y. C.; Su, J.; and Tan, C. L. 2011. Entity linking with effective acronym expansion, instance selection, and topic modeling. In *IJCAI*.

Zheng, Y.-T.; Zhao, M.; Song, Y.; Adam, H.; Buddemeier, U.; Bissacco, A.; Brucher, F.; Chua, T.-S.; and Neven, H. 2009. Tour the world: building a web-scale landmark recognition engine. In *CVPR*.