Map Translation Using Geo-tagged Social Media

Sunyou Lee, Taesung Lee, Seung-won Hwang
POSTECH, Korea
{sylque,elca4u,swhwang}@postech.edu

Abstract

This paper discusses the problem of map translation, of servicing spatial entities in multiple languages. Existing work on entity translation harvests translation evidence from text resources, not considering spatial locality in translation. In contrast, we mine geo-tagged sources for multilingual tags to improve recall, and consider spatial properties of tags for translation to improve precision. Our approach empirically improves accuracy from 0.562 to 0.746 using Taiwanese spatial entities.

1 Introduction

A map is becoming an essential online service for mobile devices, providing a current location and generating directions to spatial entities (SEs). Although major map services aim to support a map in more than 100 local languages, their current support is often biased either to English or local maps. For example, Figure 1 contrasts richly populated Taiwanese entities (in the local language) whereas only some of those entities are translated in English version. Our goal is to translate richly populated SEs into another language, in the finer granularity such as restaurants.

A baseline approach would be adopting existing work on entity transliteration work, which uses phonetic similarity, such as translating 'Barack Obama' into '贝拉克·奥巴马' [Beilake·Aobama]. Another approach is using automatically-harvested or manually-built translation resources, such as multilingual Gazetteer (or, SE dictionary[1]). However, these resources are often limited to well-known or large SEs, which leads to translation with near-perfect precision but low recall.

[1]For example, http://tgnis.ascc.net provides SE translation pairs.

Moreover, blindly applying existing entity translation methods to SE translation leads to extremely low accuracy. For example, an SE '十分車站' should be translated into 'Shifen station', where '十分' is transliterated to [Shifen], whereas '車站' is semantically translated based on its meaning 'station'. However, due to this complex nature often observed in SE translation, an off-the-shelf translation service (e.g., Google Translate) returns 'very station'[2] as an output. In addition, SE names are frequently abbreviated so that we cannot infer the meanings to semantically translate them. For instance, 'United Nations' is often abbreviated into 'UN' and its translation is also often abbreviated. As a result, the abbreviation in the two languages, (UN, 联合国), shares neither phonetic nor semantic similarity.

To overcome these limitations, we propose to extract and leverage properties of SEs from a social media, namely Flickr. Especially, we exploit co-occurrence of names in two different languages. For example, '台北' co-occurs with its English translation 'Taipei' as tags on the same photo. This is strong evidence that they are translations of each other. In addition to co-occurrence, we leverage spatial properties of SEs. For example, among tags that frequently co-occur with '台北', such as 'Taipei' and 'Canon', 'Taipei' is

[2]As of Dec 26, 2013.

Figure 1: A map of Taipei in English. Google Maps, as of Oct 14, 2013

Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, pages 59–63,
Gothenburg, Sweden, April 26-30 2014.

Symbols	Description
$\mathbb{C}$	A set of all Chinese spatial entities
c	A Chinese spatial entity, $c \in \mathbb{C}$
e	An English entity
p	A photo
D	Photos
D_c	Photos with c
D_e	Photos with e
E_c	a set of English tags from D_c
G_c	a set of GPS coordinates from D_c
G_e	a set of GPS coordinates from D_e

Table 1: Overview of symbols

more likely to be its correct translation because the spatial distributions of the two tags are similarly skewed in the same area. Our approach significantly improves the F1-score (0.562 to 0.746), compared to an off-the-shelf translators.

2 Overall Framework

We provide the framework of our proposed method using predefined symbols (Table 1). We consider a scenario of translating each SE c in a set of all SEs $\mathbb{C}$ in a Chinese map into English so that we obtain an English map[3].

STEP 1. Finding a set D_c: We crawl a photo set D with tags from Flickr. We consider each of the tags as an entity. Given an SE $c \in \mathbb{C}$, we find a set $D_c \subseteq D$. For each photo in D_c, we obtain a set of tags in multiple languages and GPS coordinates of the photo as translation evidence (Table 2).

STEP 2. Collecting candidate English tags: To obtain translation candidates of c, we build a set E_c of English tags that co-occur with c, and a set $D_e \subseteq D$ of photos for each $e \in E_c$.

STEP 3. Calculating matching score $w(c, e)$: For an English candidate $e \in E_c$, we calculate the matching score between c and e, and translate c into e with the highest $w(c, e)$ score. We describe the details of computing $w(c, e)$ in Section 3.

[3]We use an example of translating from Chinese to English for illustration, but we stress that our work straightforwardly extends if multilingual tags of these two languages are sufficient.

Photos	Chinese tag	English tag
p_1	女王頭	Taipei, The Queen's Head, food
p_2	愛河	love river, food, park, dog
p_3	野柳, 女王頭	Yehliu, Taipei, food
p_4	台北, 東北角, 女王頭	The Queen's Head, Taipei, restaurant
p_5	淡水河	Taipei, Tamsui river, dog, food

Table 2: Structure of crawled photos $D = \{p_1, p_2, p_3, p_4, p_5\}$

e	The Queen's Head	Taipei
D_e	$\{p_1, p_4\}$	$\{p_1, p_3, p_4, p_5\}$
$CF(c, e)$ **(FB)**	2	3
$TS(c, e)$	0	-0.3
$w(c, e)$ **(SB)**	0	-0.9

Table 3: **SB** vs. **FB**: Translating c = 女王頭 into $e \in E_{\text{女王頭}}$ where $D_{\text{女王頭}} = \{p_1, p_3, p_4\}$

3 Matching Score

3.1 Naive Approach: Frequency-based Translation (FB)

A naive solution for map translation is to use co-occurrence of multilingual tags. For example, if a Chinese tag '女王頭' frequently co-occurs with an English tag 'The Queen's Head', we can translate '女王頭' into 'The Queen's Head'. Specifically, for a given Chinese SE c and a candidate English tag e, we define *co-occurring frequency* $CF(c, e)$.

Definition. *Co-occurring Frequency* $CF(c, e)$. Co-occurring frequency $CF(c, e)$ is the number of photos in which c and e are co-tagged,

$$CF(c, e) = |D_c \cap D_e|, \quad (1)$$

where D_c and D_e are photos with a Chinese SE c and an English tag e, respectively.

We compute $CF(c, e)$ for all candidates in $e \in E_c$ and rank them. Then, **FB** translates c into e with the highest $CF(c, e)$ score. However, **FB** cannot address the following two challenges that occur due to tag sparseness.

- C1 : Large regions such as 'Taiwan', 'Taipei' (Section 3.2)
- C2 : Non-SEs such as 'dog', 'food' (Section 3.3)

3.2 Overcoming C1: Scarcity-biased Translation (SB)

Users tend to tag photos with both a specific SE and large administrative regions such as 'Taiwan' and 'Taipei', which makes **FB** score of large regions higher than the proper one. For example, 'Taipei' is tagged in most photos in D (Table 2); therefore, $CF(\text{女王頭}, \text{Taipei})$ larger than $CF(\text{女王頭}, \text{The Queen's Head})$ (Table 3).

To reduce the effect of large regions, we introduce a new feature to give high scores for specific SEs (e.g., 'The Queen's Head'). We observe that a large region's tag is associated with many photos in $D - D_c$, whereas a scarce but useful tag is particularly tagged in D_c. We consider $\frac{|D_e|}{|D-D_c|}$ to measure how many photos have e without c. Therefore, $\frac{|D_e|}{|D-D_c|}$ increases as e frequently appears where c does not. In contrast, if e appears mostly with c, the ratio decreases. Taking inverse of the ratio to give higher score when e appears mostly with c, we define *tag scarcity* $TS(c, e)$ and apply it to the candidate ranking function.

Definition. *Tag scarcity* $TS(c, e)$. Given an SE c and a candidate English tag $e \in E_c$, the tag scarcity is defined as

$$TS(c, e) = \log |D - D_c|/|D_e|. \quad (2)$$

Definition. *Scarcity-biased Matching Score* $w(c, e)$. Given an SE c and a candidate English tag $e \in E_c$, the matching score between c and e is

$$w(c, e) = CF(c, e) \times TS(c, e). \quad (3)$$

To illustrate the effect of **SB** with our running example (Table 2), we compare 'The Queen's Head' to 'Taipei' for translating '女王頭' (Table 3). **FB** gives a higher score to 'Taipei' than to the correct translation 'The Queen's Head'. In contrast, by reflecting TS, **SB** correctly concludes that 'The Queen's Head' is the best match.

Apart from **SB**, we can also leverage an additional resource such as an *administrative hierarchy*, if exists, to blacklist some large regions' names from E_c. By first translating larger regions and excluding them, the precision for translating small SEs can increase. For instance, we translate a country '台灣 (Taiwan)' earlier than a city '台北 (Taipei)'. Then, when translating '台北', even though $CF(\text{台北}, \text{Taiwan})$ is higher than $CF(\text{台北}, \text{Taipei})$, we ignore 'Taiwan' in $E_{\text{台北}}$ because it is already matched with '台灣'.

3.3 Overcoming C2: Pruning Non-SEs (PN)

We prune non-SEs such as 'food' based on spatial locality of a tag. We observe that the GPS coordinates G_e of photos with an SE tag e tend to be more concentrated in a specific region than those of photos with a non-SE. For instance, comparing a non-SE 'food' and an SE 'The Queen's Head', the GPS coordinates in G_{food} are more widespread all over Taiwan than those in $G_{\text{The Queen's Head}}$.

We leverage the coordinates of a *distant SE pair*. For example, two spatially far SEs '台北 (Taipei)' and '台東 (Taitung)' compose a distant SE pair. Because both SEs are unlikely to be tagged in a single photo, an English tag that co-occurs with both of them would be a non-SE.

Formally, we define two Chinese SEs c_1 and c_2 as a distant SE pair if $G_{c_1} \cap G_{c_2} = \emptyset$, and M as a set of all distant SE pairs among $\mathbb{C} \times \mathbb{C}$. We judge that an English tag e is a non-SE if G_e intersects with both G_{c_1} and G_{c_2} for a distant pair c_1 and c_2. Formally, an English tag e is non-SE if the following equation $PN(e)$ is nonzero.

$$PN(e) = \sum_{(c_1,c_2)\in M} |G_{c_1} \cap G_e| \times |G_{c_2} \cap G_e|. \quad (4)$$

4 Evaluation

4.1 Experimental Setting

Photo Data and Ground Truth: We crawled 227,669 photos taken in Taipei from Flickr, which also provided GPS coordinates of photos. We took a set D of 148,141 photos containing both Chinese and English tags and manually labelled 200 gold standard Chinese-English SE pairs whose names appeared together in at least one photo in D.

Administrative Hierarchy: An administrative hierarchy was obtained from *Taiwan Geographical Names Information System* [4].

Baselines: We chose baselines available for many languages except for the gazetteer and excluded methods that used specific textual corpora.

- Phonetic Similarity (PH) (Kim et al., 2013)
- Off-the-shelf Translator: Google Translate[5], Bing Translator[6]
- Taiwanese-English Gazetteer (official SE translation[4])

[4] http://tgnis.ascc.net/. Its latest modification has done on August 23, 2013.

[5] http://translate.google.co.kr/

[6] http://www.bing.com/translator

Chinese SE [Transliteration]	SB+PN	PH	Google Translate	Bing Translator	Gazetteer
兔子餐廳 [Tuzi Canting]	**To House**	Astrid	Rabbit Restaurant	Hare House	ø
典華旗艦館 [Dianhua Gijianguan]	**Denwell Restaurant**	Taipei Restaurants	Dianhua Flagship Museum	Classic China Flagship Center	ø

Table 4: Example translation from our method and the baselines (Correct translations are boldfaced.)

Method	P	R	F1
Transliteration	.463	.463	.463
Google Translate	.562	**.562**	.562
Bing Translator	.425	.425	.425
Taiwanese-English Gazetteer	**.960**	.485	**.645**

Table 5: P, R, and F1 of baselines

Method	P	R	F1	MRR
FB	.215	.215	.215	.439
FB + PN	.220	.220	.220	.454
SB	.640	.640	.640	.730
SB + PN	**.680**	**. 670**	**.675**	**.752**

(a) Without administrative hierarchy

Method	P	R	F1	MRR
FB	.515	.515	.515	.641
FB + PN	.624	.615	.620	.730
SB	.655	.655	.655	.733
SB + PN	.706	.695	.700	.763
FB + SB + PN	**.751**	**.740**	**.746**	**.806**

(b) With given hierarchy

Table 6: Effect of FB, SB, PN, and the hierarchy

Measures: We measured *precision* (P), *recall* (R), *F1-Score* (F1), and *mean reciprocal rank* (MRR) where MRR $= \frac{1}{|P|}\sum_{(c,e_0)\in P}\frac{1}{rank(c,e_0)}$, for which P is a set of gold standard pairs (c, e_0) of a Chinese SE c and its correct translation e_0, and $rank(c, e_0)$ indicates the rank of $w(c, e_0)$ among all $w(c, e)$ s.t. $e \in E_c$.

4.2 Experimental Results

Comparison to Baselines: The proposed approach (**SB + PN**) with or without the administrative hierarchy provided higher R and F1 than did the baseline methods (Table 5, 6).

The baseline methods showed generally low P, R, and F1. Especially, the gazetteer produced high precision, but poor recall because it could not translate lesser-known SEs such as '兔子餐廳 (To House)' and '典華旗艦館 (Denwell Restaurant)' (Table 4).

Effect of SB and PN: We experimented on the effect of the combinations of the features (Table 6). Using all the features **FB+SB+PN** with hierarchy, which translated the upper level of the hierarchy with **FB** and the lower level with **SB**, showed the best effectiveness. Simple **FB** gave both low precision and very low recall regardless of whether we used the hierarchy. Replacing **FB** with **SB** yielded both higher F1 and higher MRR.

PN increased F1, especially greatly when it was used with **SB** or the hierarchy because **PN** filtered out different types of noises, non-SEs. Applying **PN**, we classified 361 non-SEs and 6 SEs as noises in total. Despite some misclassifications, it improved the overall accuracy by ignoring highly ranked non-SEs such as 'dog' and 'food'.

5 Conclusion

We propose a scalable map translator that uses a geo-tagged corpus from social media to mine translation evidence to translate between English and maps in local languages. Our approach leverages both co-occurrence of the SE tags in Chinese and English and their scarcity and spatial property. Our approach can translate small or emerging spatial entities such as restaurants, which major map services cannot support currently. We empirically validated that our approach provided higher P, R, F1, and MRR than the existing methods including popular off-the-shelf translation services.

Acknowledgments

This research was supported by the MSIP (The Ministry of Science, ICT and Future Planning), Korea and Microsoft Research, under IT/SW Creative research program supervised by the NIPA(National IT Industry Promotion Agency). (NIPA-2013-H0503-13-1009).

References

Jinhan Kim, Seung-won Hwang, Long Jiang, Y Song, and Ming Zhou. 2013. Entity translation mining from comparable corpora: Combining graph mapping with corpus latent features. *IEEE Transactions on Knowledge and Data Engineering*, 25(8):1787–1800.