Master's Thesis

# Map Translation Using Geo-tagged Social Media

Sunyou Lee (이 선 유)

Division of Electrical and Computer Engineering

(Computer Science and Engineering)

Pohang University of Science and Technology

2014

# 소셜미디어 지명 태그를 이용한 지도 번역

# Map Translation Using Geo-tagged Social Media

# Map Translation Using Geo-tagged Social Media

by

Sunyou Lee
Division of Electrical and Computer Engineering
(Computer Science and Engineering)
Pohang University of Science and Technology

A thesis submitted to the faculty of the Pohang University of Science and Technology in partial fulfillment of the requirements for the degree of Master of Science in the Division of Electrical and Computer Engineering (Computer Science and Engineering)

Pohang, Korea
6. 19. 2014
Approved by
Seung-won Hwang

Academic Advisor



# Map Translation Using Geo-tagged Social Media

Sunyou Lee

The undersigned have examined this thesis and hereby certify that it is worthy of acceptance for a master's degree from POSTECH.

6. 19. 2014

Committee Chair Seung-won Hwang
Member Hwanjo Yu
Member Wook-shin Han

MECE 이선유, Sunyou Lee
20130999

Map Translation Using Geo-tagged Social Media

소셜미디어 지명 태그를 이용한 지도 번역

Division of Electrical and Computer Engineering (Computer Science and Engineering), 2014, 025p, Advisor: Seung-won Hwang, Text in English

**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.

# Contents

# List of Figures

# List of Tables

Chapter 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 1.1 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.

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

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


Figure 1.1: A map of Taipei in English. Google Maps, as of Oct 14, 2013

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 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

[2] As of Dec 26, 2013.

off-the-shelf translators.

The contributions of our approach can be summarized as follows.

- We study the problem of spatial entity translation, which is an important task for supporting multilingual map services.
- We propose a novel framework that leverages geo-tagged social resources as corpora and using spatial signals from such resources as an important translation evidence.
- Our evaluation results comparing with the off-the-shelf translators suggests that our proposed framework significantly improves the accuracy of translation (0.562 to 0.746).
- Our approach does not require any language-dependent resources and thus can translate any language pair.

Chapter 2

# Related Work

## 2.1 Corpus-based Approaches

Corpus-based approaches build on bilingual resources mined from parallel and comparable bilingual corpora. For example, bilingual Wikipedia entries on the same person [1, 2] can be used for translation with near-perfect precision. However, such information is scarce for lesser-known entities, which was addressed by exploring richer resources of comparable corpora by [3] and [4] These approaches would not be effective for translating SE s due to limited parallel or comparable corpora support, our approach of extracting multilingual tags from Flickr can be viewed as one way of mining parallel/comparable resources.

## 2.2 Holistic Approaches

Holistic approaches have overcome the low coverage of corpus-based approaches by aggregating many features, including phonetic features (PH). Specifically, [5] used transliteration to generate candidates and then web corpora to identify translations. This approach was later enhanced to use transliteration to guide web mining [6]. More recently, by combin-

ing phonetic feature, context feature (CX), relation (R), and spatiotemporal feature (ST), the name translation is solved as a graph matching problem [7]. In detail, after matching entities with PH and CX, the matching scores are reinforced using graphical relation of entities and parallel spatiotemporal features. Our proposed framework uses the GPS coordinates of a tag as a ST feature of translation and validates their effectiveness.

Chapter 3

# Overall Framework

We provide the framework of our proposed method using predefined symbols (Table 3.2). 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[1].

**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

[1]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.

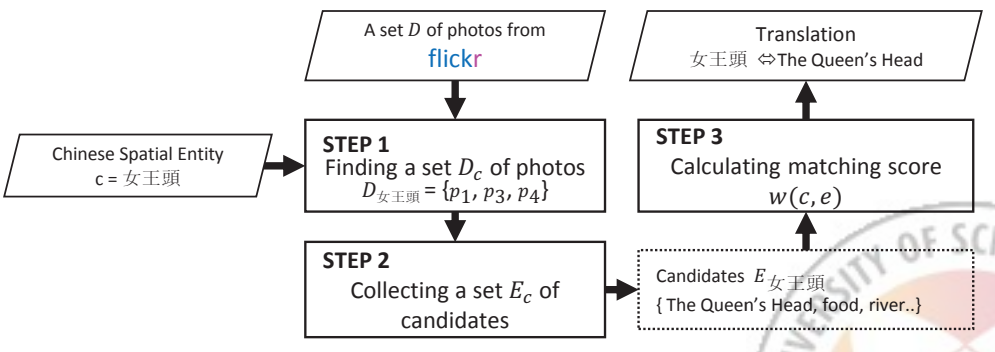


Figure 3.1: Framework

| 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 3.1: Structure of crawled photos $D = \{p_1, p_2, p_3, p_4, p_5\}$

in $D_c$, we obtain a set of tags in multiple languages and GPS coordinates of the photo as translation evidence (Table 3.1).

**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 4.

*Example.* We illustrate how we translate $c$ = '女王頭' with the example (Table 3.1, Fig. 3.1). We crawl a photo set $D = \{p_1, p_2, p_3, p_4, p_5\}$ from Flickr. We first find the set $D_{女王頭}$ among $D$, as shown shaded in Table, or $\{p_1, p_3, p_4\}$. We then extract set $E_{女王頭}$ of photos composed of English tags co-occurred with '女王頭' such as 'The Queen's Head', 'food', 'river', and 'love river', which we denote as $E_{女王頭}$. Lastly, we compute the matching score $w(女王頭, e)$ for all $e \in E_{女王頭}$ and output 'The Queen's Head' as a translation, as $w(女王頭, \text{The Queen's Head})$ scores the highest.

| 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 3.2: Overview of symbols

Chapter 4

# Method

## 4.1 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,

$$CF(c, e) = |D_c \cap D_e|, \tag{4.1}$$

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 4.2)
- C2 : Non-SEs such as 'dog', 'food' (Section 4.3)

## 4.2 Overcoming C1: Scarcity-biased Translation (SB)

We overcome the first challenge due to large regions by two approaches: applying an additional feature to the score function and blacklisting large regions using administrative hierarchy.

### 4.2.1 Applying Tag-scarcity Feature

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 3.1); therefore, $CF(\text{女王頭}, \text{Taipei})$ larger than $CF(\text{女王頭}, \text{The Queen's Head})$ (Table 4.1).

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

$$TS(c, e) = \log |D - D_c|/|D_e|. \tag{4.2}$$

| 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 4.1: **SB** vs. **FB**

**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

$$w(c, e) = CF(c, e) \times TS(c, e). \tag{4.3}$$

To illustrate the effect of **SB** with our running example (Table 3.1), we compare 'The Queen's Head' to 'Taipei' for translating '女王頭' (Table 4.1). **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.

### 4.2.2 Blacklisting Large Regions using Hierarchy

Apart from **SB**, to overcome the challenge of large regions, we could use geographical hierarchy to give priority of translation of some Chinese entities over the others. A Chinese entity with lower priority has a blacklist composed of large regions' translation. For instance, suppose an Chinese entity '台北' takes upper level than '女王頭' of the hierarchy. Then, we give a priority for '台北' to be translated and its translation results in 'Taipei'. Because we remove 'Taipei' from $E_{女王頭}$, the translation of '女王頭' cannot be 'Taipei'.

We can obtain such hierarchy by either mining given resources or generating one using GPS data of photos. Although the former one is language-dependent, these resources are open-sourced from Wikipedia or Google Maps. For example, Google Maps provides *"What's*

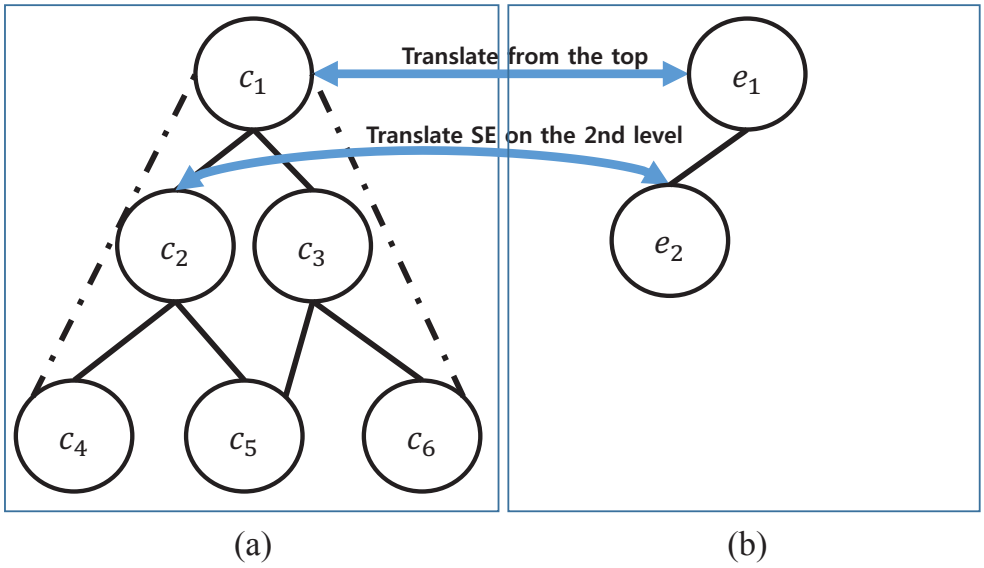


Figure 4.1: Level-by-level construction of corresponding English hierarchy

*here?"* which gives an entity's local address. We transform the address into hierarchy.[1]

When such resources are not available, we could exploit a simplified application of *Place Hierarchy*[2][8] to automatically generate hierarchy. In detail, we used Minimum Bounding Rectangle (MBR) of the GPS coordinates of photos to construct hierarchy of Chinese entities. Specifically, given two Chinese entities $c_1$ and $c_2$, suppose the MBR of the tag $c_1$ contains that of $c_2$. We then consider $c_1$ as a larger region spatial entity than $c_2$. Because we identify all possible parents of an entity, a spatial entity could have multiple larger regions than itself, i.e. a graph structure (Figure 4.1a), not a tree structure. By translating from the top of the Chinese hierarchy, we can make corresponding English hierarchy (Figure 4.1b).

[1]We can mine two sets of hierarchy in two languages from the addresses and their translation, because the translation of large regions' addresses are often provided together. Thus, we can solve spatial entity translation of large regions on the upper level as a graph matching problem. However, small spatial entities still require our proposed framework in order to be translated.

[2]It makes a polygon which represents how a tag is geographically spread. *Place Hierarchy* is constructed using inclusion and exclusion of the polygons.

## 4.3 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 $G_{\text{food}}$ 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.

$$PN(e) = \sum_{(c_1,c_2)\in M} |G_{c_1} \cap G_e| \times |G_{c_2} \cap G_e|. \tag{4.4}$$

For example of translating '大稻埕' (Table 4.2b), because $C_{\text{building}}$ intersects with a distant pair '大同區' and '中正區', we remove 'building' from the candidates. In the same way, we also remove 'wharf' and 'temple'. Otherwise, $C_{\text{Dadaocheng}}$ has an intersection only with $C_{\text{大稻埕}}$; we determine the English tag 'Dadaocheng' as a SE.

| **Rank** | **e** $\in E_c$ | $w$(大稻埕, $e$) |
|---|---|---|
| 1 | wharf | 9.82 |
| 2 | temple | 4.31 |
| 4 | building | 2.99 |
| 6 | **Dadaocheng** | 2.25 |
| 7 | city | 1.80 |
| 21 | ROC | 1.20 |
| 50 | Datong | 0.50 |
| 67 | Taipei | -8.85 |

(a) SB

| **Rank** | **e** $\in E_c$ | $w$(大稻埕, $e$) |
|---|---|---|
| 1 | **Dadaocheng** | 2.25 |
| 3 | Coolpix | 1.41 |
| 7 | ROC | 1.20 |
| 8 | church | 1.06 |
| 20 | wave | 0.57 |
| 23 | Datong | 0.50 |
| 24 | boat | 0.48 |
| 28 | Taipei | -8.85 |

(b) SB+PN

Table 4.2: Effect of **PN**

Chapter 5

# Evaluation

## 5.1 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* [1].

**Baselines:** We choose four baselines; Phonetic similarity, Google Translate, Bing Translator, and Gazetteer. Our framework matches a Chinese SE with an English SE without language specific tools such as phonetic algorithm, dictionary, and parser, to translate a map. Thus, we choose baselines that are available for many languages, and exclude methods using specific textual corpora that have been shown that are not suitable for translating location or

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

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

Table 5.1: P, R, and F1 of baselines

organization.

- Phonetic Similarity (PH) [7]
- Off-the-shelf Translator: Google Translate[2], Bing Translator[3]
- Taiwanese-English Gazetteer (official SE translation[4])

**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. $e \in E_c$.

## 5.2 Experimental Results

### 5.2.1 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 5.1, 5.3).

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 5.2).

[2] http://translate.google.co.kr/
[3] http://www.bing.com/translator

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

Table 5.2: Example translation from our method and the baselines (Correct translations are boldfaced.)

### 5.2.2 Effect of SB

We experimented on the effect of the combinations of the features (Table 5.3). 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.

To verify **SB** is more effective when applied to lower (small and emerging regions), we translated the upper level of the hierarchy with **FB** and the lower level with **SB**. It was represented as **FB+SB+PN** with hierarchy, showed the best effectiveness. In detail, we translated upper and lower level separately (Table 5.4. It definitely showed that SEs on the lower level were much better translated by **SB** than by **FB**. In the meantime, **SB** is a little affected by using the hierarchy since both **SB** and the hierarchy are devised for the same problem, sparseness due to large regions.

### 5.2.3 Effect of PN

**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’.

| 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 administrative hierarchy

| Method | P | R | F1 | MRR |
|---|---|---|---|---|
| **FB** | .295 | .295 | .295 | .472 |
| **FB + PN** | .442 | .435 | .438 | .591 |
| **SB** | .650 | .650 | .650 | .798 |
| **SB + PN** | .706 | .695 | .700 | .826 |
| **FB + SB + PN** | **.711** | **.700** | **.705** | **.833** |

(c) With automatic hierarchy

Table 5.3: Effect of FB, SB, PN, and the hierarchy

| Method | Upper Level | Lower Level | Overall |
|---|---|---|---|
| **FB** | .811 | .589 | .620 |
| **SB** | .649 | .718 | .700 |

Table 5.4: **FB** vs. **SB** for SEs on upper and lower level

| Gold Pair $(c, e_0)$ | **FB** | **FB + PN** | **SB + PN** |
|---|---|---|---|
| (大稻埕碼頭, Dadaocheng Wharf) | 25 | 18 | 11 |
| (台北花園大酒店, Taipei Garden Hotel) | 10 | 9 | 3 |
| (永安漁港,Yongan Harbour) | 15 | 10 | 4 |
| (福爾摩沙高速公路, Formosa Freeway) | 5 | 5 | 2 |
| (自然科學博物館, National Museum of Natural Science) | 7 | 7 | 4 |
| (巴黎波波,Lepain) | 8 | 6 | 3 |
| (南京東路,Nanjing East Road) | 11 | 7 | 2 |

Table 5.5: Rank of gold pair $(c, e_0)$.

### 5.2.4 Error Analysis

We show examples of wrong translations using our approach in Table 5.5. Although the correct translation was ranked better by applying **SB** and **PN**, it cannot be top-1 answer. We found that the wrong translations were caused by two reasons; a) non-filtered or wrongly filtered non-SEs by **PN**, and b) incorrect spelling. The first problem can be solved by more sophisticatedly constructing the set $M$ of distant pair, but the spelling problem is out of our scope.

Chapter 6

# 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.

# Bibliography


[1] Donghui AAFeng, Yajuan L¨u, and Ming Zhou. A new approach for English-Chinese named entity alignment. In *Proceedings of the 2004 Conference on Empirical Methods on Natural Language Processing*, volume 2004, pages 372–379, 2004.

[2] Julian Kupiec. An algorithm for finding noun phrase correspondences in bilingual corpora. In *Proceedings of the 31st Annual Meeting on Association for Computational Linguistics*, pages 17–22. Association for Computational Linguistics, 1993.

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

[4] Li Shao and Hwee Tou Ng. Mining new word translations from comparable corpora. In *Proceedings of the 20th International Conference on Computational Linguistics*, page 618. Association for Computational Linguistics, 2004.

[5] Yaser Al-Onaizan and Kevin Knight. Translating named entities using monolingual and bilingual resources. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 400–408. Association for Computational Linguistics, 2002.

[6] Long Jiang, Ming Zhou, Lee-Feng Chien, and Cheng Niu. Named entity translation with web mining and transliteration. In *Proceedings of the 20th International Joint Conferences on Artificial Intelligence*, volume 7, pages 1629–1634, 2007.

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

[8] Donggi Jung, Hyunwoo Park, Ryong Maeng, and Sangki Han. A geometric patternbased method to build hierarchies of geo-referenced tags. In *Social Computing (Social-Com), 2010 IEEE Second International Conference on*, pages 546–551. IEEE, 2010.

[9] Lee, Sunyou, Taesung Lee, and Seung-won Hwang. Map Translation Using Geo-tagged Social Media. *EACL 2014* (2014): 59.

# 요 약 문

세계화가 가속되고 관광 산업이 발전함에 따라 한 지역사회에 콘텐츠를 다언어로 공급할 필요가 증가하고 있다. 지도 역시 다언어로 제공되어야 하는 필수 콘텐츠 중 하나이다. 현재 많은 온라인 지도 서비스에서도 백가지 이상의 언어로 지도를 제공하는 것을 목표로 하고 있다. 이 논문에선 사용자가 원하는 언어로 지도를 제공할 수 있도록, 지명을 번역하는 방법을 제안한다. 특히 실제 지역주민이 사용하는 지명으로 번역하기 위해서는 발음과 의미를 적절하게 조합하여 번역해야하는 특성을 갖는다. 예를 들어 '十分車站 (Shifen Station)' 은 '十分' 은 발음을 이용해 'Shifen' 으로, '車站' 는 의미를 이용해 'Station'으로 번역해야 한다. 이 논문에선 소셜미디어인 Flickr의 사진 데이터에 태그된 사용자가 실제 사용하는 지명을 이용하여, 지도를 번역하는 방법을 제안했다. 기존의 객체 번역방법들이 '사람' 타입의 객체번역에만 치우친 성능을 보이고, 관련 연구들이 웹문서를 주로 이용하여 그러한 자료가 없을 때는 번역에 큰 한계를 보인다. 하지만 제안한 방법에선 지명 태그와 사진으로 연계된 위치정보를 이용함으로써 지명에만 특화된 정보를 뽑아낼 수 있었다. 결과적으로 실제 대만의 지명 번역에 적용하였을 때, 대조군으로 사용한 온라인 번역기 Google Translate와 공식적인 지명 번역 사전보다 높은 정확도 및 재현률을 보였다.

# 감사의 글

먼저 모든 순간순간을 함께 해주신 하나님께 감사드립니다.

열정적으로 지도해주시고 적극적으로 동기 부여해주신 황승원 교수님께 감사드립니다. 아직 부족하지만, 주도적으로 일을 진행해 나가는 자세를 배울 수 있었습니다. 또한 바쁘신 와중에 석사 디펜스 심사를 해주신 유환조, 한욱신 교수님께도 감사드립니다.

언제나 지지해주시는 엄마, 아빠, 오빠, 우리 아들 헌댕 헌형이, 내 엔돌핀 예쁜 한지 지영이, 지혜찡, 정미송 자매님과 은화언니를 비롯한 네비게이토 형제자매님들, 자주 찾아뵙지 못해 죄송한 양태성 선생님 감사드립니다. 여러분의 사랑과 기도와 훈계와 동행이 없었다면 포스텍에서 이 5년을 헤쳐 나올 수 없었을 겁니다. 진심으로 감사드립니다.

고맙고 존경하는 포스텍 합창단 코러스 선후배동기님들 감사합니다. 선유졸업중창 멤버 혜민언니, 남규오빠, 준영오빠, 승현선배, 동훈이, 서영이, 철훈이, 영리, 항상 관심과 사랑 주시는 두엽오빠. 모두 제게 잊지 못할 추억을 주셨어요.

그리고 우리 IDS 연구실 식구들! 오빠들 덕분에 이렇게 잘 있다 갑니다. 은근히 챙겨주시고 연구실의 중심을 잡아준 진한오빠, 언제나 다정하고 연구 외적으로도 배려해주신 진우오빠, 제 마음의 두 번째 지도교수님 태성오빠, 제 가치관확립에 도움을 주신 토론왕 진영오빠, 점점 말이 많아지는 잘생긴 영록빠, 글로벌 박사님 우리 연구실의 미래 성환오빠. 모두 고맙습니다. 특히 제 짝꿍 불평불만 뭐든지 귀 기울여 들어주시고 연구 조언도 해주시고 놀자고 꼬드기면 같이 놀아주시고 밥도 많이 사주신 연구실 지킴이 섬세한 상훈이 오빠와 우리 파티션의 플러스맨 개그맨 겸손왕 기어올라도 많이 봐주시고 땡깡부려도 이해해주신 벌써 두 아이의 아빠가 된 현석

오빠! 제 비빌 언덕이 돼주셔서 감사드립니다. 현석오빠와 함께 두 아이의 어머니가 되신 정은언니와 밀당하는 캄보 하난이와도 지금의 기쁨을 나누고 싶습니다.

마지막으로 절 발견하고 키워준 자랑스러운 우리 학교 포스텍에게도 고맙다는 말 전하고 싶습니다.

지금 이렇게 감사의 글을 쓰기까지 저 혼자 해낸 것은 아무것도 없었습니다. 제가 이 마음을 잊을 때마다 경책해 주세요. 저도 모두에게 미래와 희망이 있길 기도하겠습니다. 감사합니다.

# CURRICULUM VITAE

*Name*. Sunyou Lee

## *Education*

**Bachelor's Degrees**

Electrical Engineering

Pohang University of Science and Technology (POSTECH), Korea, Republic of

March 2009 – February 2013

**Master's Degrees**

Computer Science & Engineering

Pohang University of Science and Technology (POSTECH), Korea, Republic of

March 2013 – October 2014

## *Publications*

1. Sunyou Lee, Taesung Lee, and Seung-won Hwang, "Map Translation Using Geo-tagged Social Media" *14th Conference of the European Chapter of the Association for Computational Linguistics (EACL)* Gothenburg, Sweden, 2014.

2. Jongwuk Lee, Hyunsouk Cho, Sunyou Lee, and Seung-won Hwang, “Toward Scalable Indexing for Top-k Queries” *IEEE Transactions on Knowledge and Data Engineering*, 1, 2013.

## *Award*

1. **Magna Cum Laude**, Pohang University of Science and Technology (2014)