


Geotagging Location Information Extracted from Unstructured Data

Kyunghyun Min

Department of Civil and Environmental Engineering, Seoul National University 35-209,
Gwanak-gu, Seoul, Republic of Korea
minkh@snu.ac.kr

 <https://orcid.org/0000-0002-4835-7215>

Jungseok Lee

Department of Civil and Environmental Engineering, Seoul National University 35-209,
Gwanak-gu, Seoul, Republic of Korea
rightstone@snu.ac.kr

Kiyun Yu

Department of Civil and Environmental Engineering, Seoul National University 35-209,
Gwanak-gu, Seoul, Republic of Korea
kiyun@snu.ac.kr

Jiyoung Kim

Institute of Construction and Environmental Engineering, Seoul National University 35-215,
Gwanak-gu, Seoul, Republic of Korea
soodaq@snu.ac.kr

Abstract

Location information is an essential element of location-based services and is used in various ways. Unstructured data contain different types of location information, but coordinate values are required to determine the exact location. In Twitter, a typical social network service (SNS) platform of unstructured data, the number of geotagged tweets is low. If we can estimate the location of text by geotagging a large number of unstructured data, we can estimate the location of the event in real-time. This study is a base study on extracting the location information by using the named entity recognizer provided by the Exobrain API and applying geotagging to unstructured data in Hangul (Korean). We used Chosun news articles, which are grammatically correct and well organized, instead of tweets to extract three location-related categories, namely “location,” “organization,” and “artifact”. We used the named entity recognizer and geotagged each sentence in combination of the fields in each category. The results of the study showed that 61% of the 800 test sentences did not have the location-related information, thus hindering geotagging. In 11.75% of the test sentences, geotagging was possible with only the given location information extracted using the named entity recognizer. The remaining 27.25% of the sentences contained information on more than two locations from the same subcategories and hence required location estimation from candidate locations. In future research, we plan to apply the results of this study to develop location estimation algorithm that makes use of the extracted location-related entities from purely unstructured data such as that on SNSs.

2012 ACM Subject Classification Information systems → Content analysis and feature selection

Keywords and phrases Location Estimation, Information Extraction, Geo-Tagging, Location Information, Unstructured Data

Digital Object Identifier 10.4230/LIPIcs.GIScience.2018.49

Category Short Paper



© Kyunghyun Min, Jungseok Lee, Kiyun Yu, and Jiyoung Kim;
licensed under Creative Commons License CC-BY

10th International Conference on Geographic Information Science (GIScience 2018).

Editors: Stephan Winter, Amy Griffin, and Monika Sester; Article No. 49; pp. 49:1–49:6

Leibniz International Proceedings in Informatics



LIPICs Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

■ **Table 1** Percentages of Tweets with Location Information.

	Max	Min	Average
% of Geotagged Tweets Per Day	0.22%	0%	0.11%

Funding This study was supported by the research funding of the project on the development of big data management, analysis, and service platform technology for the national land spatial information research project of the Ministry of Land, Infrastructure, and Transport (18NSIP-B081023-05).

1 Introduction

Recently, location-based services are growing rapidly owing to the large amount of data generated in people's lives. A person's behavior or the occurrence of an event is often accompanied by location information. Recently, the use of social network services (SNSs) has increased as a method for human expression. However, less than 0.42% of tweets were geotagged even though Twitter is providing a function to determine the location information [9]. In fact, we collected 611,687 tweets for the entire month of March 2018 and confirmed that they are geotagged only on an average of 0.11% tweets a day, as shown in Table 1. If the tweet is geotagged, a location where a specific article was written or a location that it describes is known. Hence, an incident or an accident mentioned in the SNS or the news article can be checked in real time. Therefore, by extracting the location information from these unstructured data and adding the location information, the occurrence of a specific event and its location can be monitored.

As mentioned earlier, the number of geotagged SNSs is small. As a result, many studies have been carried out only on geotagged posts [5, 7]. Therefore, other factors such as user profiles, text content, and location labeling are used to aid in an ongoing location estimation [4]. One study detected earthquake in real time and inferred the location from the registered location and GPS data created when users sign up the unstructured data platform, Twitter [8]. Further, a study on the extraction of location-related entities from each tweet on twitter using named entity recognition and concept-vocabulary-based extraction has been performed [1]. Recently, research has been performed to detect the location information in text using the conditional random fields (CRF) model [3]. However, a case in which the location information is extracted by using the named entity recognizer for Hangul (Korean) does not exist. In this study, we aim to geotag the unstructured Hangul data with the location information extracted with the entity recognizer.

2 Detection of Location Information by Named Entity Recognition

2.1 Named Entity Recognition

Named entities are the names of persons, organizations, locations, dates, and times. Named entity recognition refers to recognizing and tagging the corresponding entity name among proper names or noun phrases. Named entity recognition is one of the language analysis techniques that is essential in natural language processing tasks used in information retrieval or information extraction. In the English language, high-level recognition and classification performance were shown by using language characteristics such as capital letters [6]. However, in the Korean language, it is difficult to recognize an entity name in the absence of certain features such as capital letters in English. As an alternative,

■ **Table 2** Lists of items in each category (in Parts).

LC	OG	AF
LCP_COUNTRY	OGG_EDUCATION	AF_BUILDING
LCP_CAPITALCITY	OGG_SPORTS	AF_ROAD
LCP_COUNTY	OGG_FOOD	AF_TRANSPORT
LCP_CITY	OGG_HOTEL	AF_CULTURAL_ASSET
LC_TOUR	OGG_POLITICS	:
LCG_MOUNTAIN	OGG_RELIGION	:
LCP_PROVINCE	OGG_ECONOMY	:
:	:	:

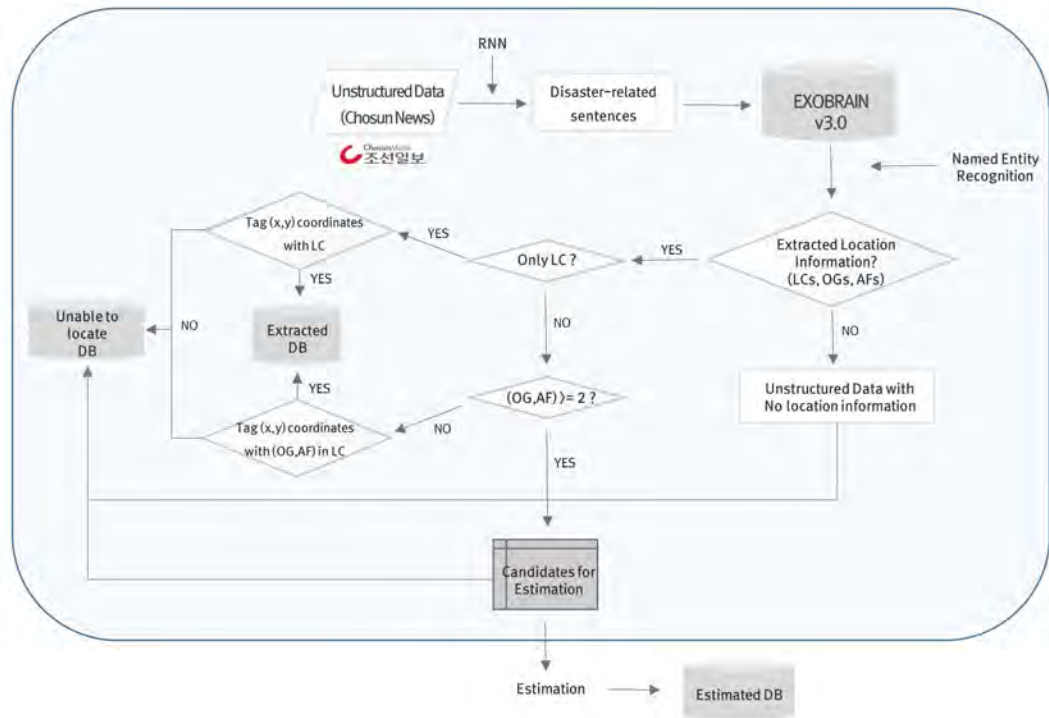
there is a study using word embedding features in recognition and classification of Korean entity names [2]. The entity name recognizer used in this study is the Exobrain language analysis open API provided by Korea Electronics and Telecommunications Research Institute (ETRI). The entity recognition corpus for Exobrain comprises of 10,000 sentences from news articles. It uses the Telecommunications Technology Association's (TTA) standard object name tag set consists of 15 main categories and 146 subcategories for object types in various fields. Location (LC), organization (OG), and artifact (AF) were selected as the necessary main categories for this study. Subcategories that can be used to extract location-related information are partly introduced in Table 2. There are fourteen subcategories for LC, fifteen for OG, and thirteen for AF. LC contains the geographical name, the administrative district name, and the like. OG contains the names of educational institutions, medical institutions, accommodations, and the like. AF indicates the names of cultural properties, buildings, and roads.

2.2 Extracting Location Information

The workflow of this study is presented in Figure 1. We extracted the LC, OG, and AF information from sentences related to fire accidents by using the entity recognizer. If no location-related information that belongs to the three major categories is obtained in the sentence, such a sentence is stored in a database (DB) that cannot be geographically located by geotagging. If extracted location information are geographically hierarchical, the coordinates corresponding to the area are tagged and stored in the extracted DB. If only one OG or one AF information exists in addition to the LC, only one coordinate value can be assigned. However, if more than two OG or AF information are to be assigned, the allocation of the location cannot be determined. In other words, if the text, in this case a sentence, is mentioning more than two locations that are not geographically hierarchical, then location estimation is needed. In our future study, several OGs and AFs will be temporarily stored as estimated candidates so that location estimation can proceed.

3 Test and Results

The sentences used in this study are 800 fire accident-related sentences from the Chosun news articles published in 2017. Since tweets are written by the users in colloquial style that is hard for computers to understand, we chose news articles as an alternative as they are grammatically correct and well structured. To geotag the sentence, not estimate, at least one LC is required. For example, if "Starbucks" is the only retrieved location information



■ **Figure 1** Work Flow Chart.

for OG, the specific Starbucks branch cannot be determined because there are more than thousand Starbucks stores in Korea. As many as 488 sentences through the named entity recognizer did not contain location information, comprising 61% of the total number of sentences. In contrast, sentences with location information including LC, OG, or AF, were 312 in number. Among them, only 94 sentences, i.e., only 11.75% out of the total, could be geotagged; for the remaining 27.25% of sentences, location estimation is required. The results are summarized in Table 3. Figure 2 shows the example visualization of named entity recognition and morphological analysis performed using the Exobrain API. The sentence at the top is written in Hangul, and the one below is the corresponding translated sentence.

4 Conclusion

Recently, the use of SNS has increased, but the location information extracted from unstructured data is lacking. We confirmed the lack of geotagging through the twitter data collected for a month and aimed to solve it through the location estimation from the named entity recognition. In this study, geotagging was performed by extracting the location-related information on LC, OG, and AF from fire accident-related sentences using the Exobrain named entity recognizer as a base study for location estimation. Our experimental results showed that 61% of 800 sentences had no extracted location information, 11.75% of sentences were geotagged, and 27.25% of sentences required location estimation. As the number of sentences has a large number of candidates that can be used for estimation, future studies will focus on improving the accuracy using named entity recognition and CRF model, and the location information can be provided to more unstructured data by developing a location estimation algorithm that uses the extracted location information.

Table 3 Application Result.

	LC, OG, AF		No Location Information	Total
	Extracted Location	Location Estimation needed	-	
Named Entity Recognition	312		488	800
Percentage [%]	11.75%	27.25%	61%	100%

밀양 세종병원 화재는 2018년 1월 26일 경상남도 밀양시 중앙로 114(가곡동)에 있는 세종병원에서 발생한 화재 사고이다.
LCP_CITY OGG_MEDICINE LCP_PROVINCE LCP_CITY AF_ROAD LCP_COUNTY OGG_MEDICINE

The fire in the Miryang Sejong Hospital is a fire accident at Sejong Hospital on
LCP_CITY OGG_MEDICINE OGG_MEDICINE
 January 26, 2018 in Gangok-dong, 114, Middle Road, Miryang-si, Gyeongsangnam-do.
LCP_COUNTY AF_ROAD LCP_CITY LCP_PROVINCE



Figure 2 Example of named entity recognition result for fire-related sentence.

References

- 1 Puneet Agarwal, Rajgopal Vaithiyathan, Saurabh Sharma, and Gautam Shroff. Catching the long-tail: Extracting local news events from twitter. In *Proceedings of the Sixth International AAAI Conference on Weblogs and Social Media*, pages 379–382, 2012.
- 2 Yunsu Choi and Jeongwon Cha. Korean named entity recognition and classification using word embedding features. In *Journal of Korean Institute of Information Scientists and Engineers*, pages 678–685, 2016.
- 3 Diana Inkpen, Ji Liu, Atefeh Farzindar, Farzaneh Kazemi, and Diman Ghazi. Location detection and disambiguation from twitter messages. *Journal of Intelligent Information Systems*, 49(2):237–253, 2017.
- 4 Farhad Laylavi, Abbas Rajabifard, and Mohsen Kalantari. A multi-element approach to location inference of twitter: A case for emergency response. *ISPRS International Journal of Geo-Information*, 5(5):56, 2016.
- 5 Ryong Lee, Shoko Wakamiya, and Kazutoshi Sumiya. Discovery of unusual regional social activities using geo-tagged microblogs. *World Wide Web*, 14(4):321–349, 2011.
- 6 Xiaohua Liu, Ming Zhou, Furu Wei, Zhongyang Fu, and Xiangyang Zhou. Joint inference of named entity recognition and normalization for tweets. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*, pages 526–535. Association for Computational Linguistics, 2012.

49:6 Geotagging Location Information Extracted from Unstructured Data

- 7 Kenta Oku, Koki Ueno, and Fumio Hattori. Mapping geotagged tweets to tourist spots for recommender systems. In *Advanced Applied Informatics (IIAIAAI), 2014 IIAI 3rd International Conference on*, pages 789–794. IEEE, 2014.
- 8 Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. Earthquake shakes twitter users: real-time event detection by social sensors. In *Proceedings of the 19th international conference on World wide web*, pages 851–860. ACM, 2010.
- 9 Cheng Zhiyuan, Caverlee James, and Lee Kyumin. You are where you tweet: a content-based approach to geo-locating twitter users. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, pages 759–768, 2010.