JOURNAL OF INFORMATION SCIENCE AND ENGINEERING 39, 655-670 (2023)
DOI: 10.6688/JISE.202305_39(3).0013

EventGo! Mining Events Through Semi-Supervised Event Title Recognition and Pattern-based Venue/Date Coupling*

YUAN-HAO LIN[1], CHIA-HUI CHANG[1,+] AND HSIU-MIN CHUANG[2]
[1]*Department of Computer Science and Information Engineering*
National Central University
Taoyuan, 320 Taiwan
[2]*Department of Information and Computer Engineering*
Chung Yuan Christian University
Taoyuan, 320 Taiwan
E-mail: luff543@gmail.com; chia@csie.ncu.edu.tw; showmin1205@gmail.com

Looking for local activities and events is a common task for most users during travel or daily life. Events are usually announced on the event organizers' website or spread by posting on social networks such as Facebook Event or Facebook Fanpages. Integrating all these activities/events allows us to explore the city and understand its dynamics. In this article, we study the problem of event extraction, including event title recognition, venue extraction, and relationship coupling. Although distant supervision is a common technique for generating annotated training data, how to choose proper seed entities depends on the nature of the entities to be extracted, and the automatic labeling strategy adopted. To improve the performance, we proposed model-based distant supervision for event title recognition and Point Of Interest (POI) extraction, which reached 0.565 and 0.536 F1, respectively. Meanwhile, we conduct sequential pattern mining from Facebook event posts to determine the event venue and start/end date when multiple addresses/POIs or temporal expressions are recognized in a message. Overall, the average F1 of the proposed model in event extraction is 0.620.

Keywords: event title extraction, venue recognition, social event search, relation coupling, semi-supervised learning

1. INTRODUCTION

With improvements in the convenience of transportation, traveling is no longer about sightseeing or taking professional photographs but more about experiencing local cultures. Thus, finding events or activities interested to users has become a necessity desired by users. Facebook's Events Calendar is one of the platform's oldest features, which gradually became a tool for finding local parties after the "*Event For You*" feature was added in 2014. Facebook launched Events, a standalone event discovery and calendar application and relaunched as *Facebook Local* in 2017. The new application enables users to search for nearby events and restaurants from Facebook's event and place database.

Received January 23, 2022; revised March 28, 2022; accepted May 20, 2022.
Communicated by Chao-Lin Liu.
+ Corresponding author.
* This work was partially supported by the Ministry of Science and Technology, Taiwan under grant MOST109-2221-E-008-060-MY3.

To challenge Facebook's dominance in the local event space, Google has collaborated with numerous event-related websites, such as Eventbrite, Allevents.in, Ticketmaster, and StubHub to ensure that their content is displayed in Google searches since 2017. As associating events with locations are crucial tasks for event platforms, Google rolled out new public event features in Google Maps in 2019, enabling businesses and event managers to create events associated with specific locations through the "*Contribution Tab*" in its Android application.[1]

The motivation of this study is to provide event search service to users interested in attending local events. An event search application can be regarded as a type of location-based service (LBS) application. Similar to geographical information retrieval systems and geosocial search systems, event search or recommendation systems aim to fulfill users' information needs. The event database is not only a key component behind the event search service, but also can be regarded as a record of people's events in a city, reflecting the community culture to a certain extent.

Most businesses and government organizations announce event information on their websites. However, discovering such event sources is still challenge. Therefore, Wang *et al*. [1] applied an event detection model to filter webpages that do not contain event announcement information. However, the cost of crawling the entire web and filtering pages that contain event or activity information is still too high. Although event-related websites are good event sources, many local activities, such as high school band concerts, cultural activities, and handicraft sales, are community-based and are unlikely to be advertised on commercial paid ticketing websites. Instead, posting activity information on their websites and social networks such as Facebook Fanpages is a common way for activity promotions.

In this paper, our objective was to extract gathering-type events that users may wish to add to their calendars and attend. We examined the problem of extracting information on upcoming events from the web for event database construction. We focused on event extraction from Facebook Fanpages in Taiwan, where a post may contain an announcement containing details on the title, venue, date and host of the event. An example post and the corresponding target output are displayed in Fig. 1.

Because not all posts contain event announcements, Wang. *et al*. [1] uses an event detection model to determine whether a given page contains events to be extracted. Instead, we consider event title extraction as an alternative method of event detection and focus on three major problems in event extraction: (1) event title; (2) event venue; and (3) event start/end date. Therefore, we only perform event venue recognition for posts containing event titles.

For event title extraction, although distant supervision is a common technique used for preparing training data, the long event titles greatly influence the quality of auto-labeling. Besides, the Facebook posts from Fanpages do not synchronize well with the Facebook events because only 13% of Facebook events are posted on their Facebook Fanpage.[2] How to prepare annotated training data with high quality is the main challenge.

[1] https://www.searchenginejournal.com/google-maps-rolls-out-new-public-events-feature/300151/

[2] This statistic is estimated by querying the event title from our Fanpage post retrieval system. If the event title has a Levenshtein similarity greater than 0.8 with any of the top 10 Fanpage posts, we consider that the Facebook event was indeed published on the fan page post. Of the 336K events tested, only 44K events were considered reposts on the Fanpage.

Event Title	Start	End	Venue
「稅樂融融幸福你我」音樂會 "Happy Tax Day, Happiness with You and Me" Concert	109/9/8	None	臺中中山堂 Taichung Zhongshan Hall

Fig. 1. (top) Post on a Facebook fan page that contains an event announcement and (bottom) the target output.

As for event venue extraction, the main challenge is that many venue names can be abbreviated in informal writing. In addition, event venues can be specific to an address or extended across streets and neighborhoods.

Another challenge of event extraction is the relation between event title and POI. This is because event announcements may mention the location of the past events, so they may contain multiple locations. Finally, we adopt a fine-grained geocoding strategy via consistency checking to determine the geographic scope. In general, the contribution of this paper can be summarized as follows.

- We adopt seed-based distant supervision to prepare two training corpus (Google search snippets and Facebook Fanpage posts) using Facebook Event and CityTalk events. Since only a small percentage of event posts contain event titles (11.03% for Facebook Event and 13.15% for CityTalk), we introduce longest common subsequence (LCS) matching and the core-word filtering technique from [2] to overcome the imprecise annotation due to approximate matching. The experimental results show that these two mechanisms can effectively improve the F1 performance of the Facebook corpus (from 0.298 to 0.331) and the Google corpus (from 0.402 to 0.526).

- We further proposed model-based distant supervision, which used the best model trained from Google corpus to automatically label the Facebook corpus. The experimental results show that the model-based distant supervision can effectively generate more accurate labeled training data, therefore improving the F1 performance on the test set from 0.526 to 0.565.

- For venue extraction, we consider both address extraction and POI recognition for candidate venue extraction, and construct a POI recognition model by filtering POIs in Facebook placeDB as seeds for training data preparation.

- Finally we determine event attribute coupling through sequential pattern mining, and apply the filtered patterns to determine the event venue and start/end date to achieve 0.620 F1 event extraction performance.

The remainder of this paper is organized as follows. Section 2 describes studies related to event or event extraction and venue recognition. Sections 3 and 4 introduce the proposed methods for event title recognition and venue recognition, respectively. Conclusions are drawn in Section 5.

2. RELATED WORK

As stated in the book *Bowling Alone*, face-to-face social interaction is essential to the active civic participation that a democratic society requires. Therefore, there are many event-based social networks, such as Meetup.com and Facebook Events, which provide us with a platform to attract various event organizers and participants with common interests. However, these networks represent only a subset of event organizers. Most companies, organizations and authorities announce events on their official websites and use social networks or event-related websites to promote their activities. For example, most organizations also use Facebook Fanpages to interact with their customers.

Two research teams from Google, Foley *et al*. [3] and Wang *et al*. [1] revealed the challenge involved in constructing an event database. Foley *et al*. [3] used 217,000 unique events from **Freebase** to conduct distant supervision to extract local events from the ClueWeb12 data set. By dividing a document into small text strings separated by HTML tags, the system constructs a LIBLINEAR classifier for each event field: "When," "Where" and "What". They proposed 13 features, including textual (*e.g*., unigrams and bigrams), natural language (*e.g*., capitalization, address overlap, and date overlap), and structural (*e.g*., parent and sibling HTML tags). Each text string is assigned the event field with the highest score. The extracted fields are then grouped into regions of complete events through the resolution of a general subset selection problem. Each region is then ranked by a scoring function based on field scoring, region scoring, and document scoring.

Foley *et al*. evaluated the precision of three event fields at three recall points (very high, high and moderately high). Notably, the precision for the "What" field (event title) classification decreased significantly from 0.76 (very high) to less than 0.50 (high) and then 0.30 (moderately high). The average precision of the "What," "When" and "Where" fields was 0.36, 0.32, and 0.66, respectively. However, whether general subset selection outputs a correct coupling between event title and venues or dates has not been evaluated in the paper.

To improve the performance of the work of Foley *et al*., Wang *et al*. proposed an event extraction pipeline, which consists of six modules, namely, event page classifier, single/multiple event classifier, single event extractor, multiple events extractor with repeated patterns, event consolidation module, and a wrapper induction module. The process are divided into two stages. The first stage uses the event page classifier and the single/multiple event classifier to filter web pages, and then applies the single event extractor to extract event data (namely, title, date, and location).

The second phase improves the quality of raw event extraction in the first stage by using repeated patterns from multiple-event pages, the event consolidation module, and the wrapper induction module. A precision of 0.88 was achieved for raw event title extraction from the top-ranked text string (*i.e*., Precision@1) by using a feed-forward neural network with sparse, Boolean, and bucketized features. With regard to event date and location ex-

traction, pattern-based approaches were employed to obtain all dates and locations on a page as event date and location candidates. Subsequently, a joint model was trained to predict the probability that each date and location pair would appear on the page. The precision of the predictions for date and location was 0.96 and 0.93, respectively.

The event extraction problem of the above two papers [1, 3] and our paper is different from the traditional event traditional addressed in Automatic Content Extraction (ACE),[3] which includes the extraction tasks of 33 event types (8 categories), 28 semantic roles, and 9 entity types. The event extraction in ACE is mostly carried out by supervised learning methods. For example, Yang and Mitchell [4] modeled a joint inference framework as an integer linear program to determine the event type, semantic role, and entity type. On open-domain, Ritter *et al*. collected approximately 100 million Twitter posts, from which they constructed a Tweet calendar system [5] to extract events in the open field, where each event can be described by four attributes: person, event phrase, date, and event category. For the most frequently discussed events, the aforementioned system has a high accuracy of 0.9. However, its accuracy decreases to 0.66 when the top 500 extracted events (by frequency) are considered.

3. EVENTGO SYSTEM ARCHITECTURE AND EVENT TITLE RECOGNITION

This research is an extension of our previous work [6]. In this paper, we give the whole picture of the system architecture, including the corpus preparation and model training for event extraction, as well as downstream application such as event search and event dynamic analysis. As shown in Fig. 2, EventGo system includes three parts: preparation of training corpus, extraction of event title/venue/date, and event field coupling. The design and performance of seed event crawling as well as event dynamic analysis can be found in [7]. This article mainly focuses on event title extraction, event venue recognition (POI and address), and the coupling between event attribute and venue.

3.1 Seed Event and Training Corpus Preparation

To prepare the training corpus for event extraction in Chinese, we explore two event sources, CityTalk and Facebook Events. CityTalk is a ticketing website that provides upcoming events in Taiwan. Facebook events were considered because the test data originated from Facebook posts. We do not use Freebase because the number of Chinese events contained in Freebase is limited.

We crawled the CityTalk (http://www.citytalk.tw/) website and built a wrapper to collect events as our seed events. For Facebook events, we search for the events from Facebook Event explore URL (https://www.facebook.com/events/) to find EVENT IDs from the search results. We then visit each event post to collect event information and parse the returned HTML page to extract the title, date, location, description, and Global Positioning System (GPS) information. A total of 117K and 273K events from CityTalk and Facebook Event are collected as our seed events as shown in Table 1.

[3] ACE: https://www.ldc.upenn.edu/collaborations/past-projects/ace

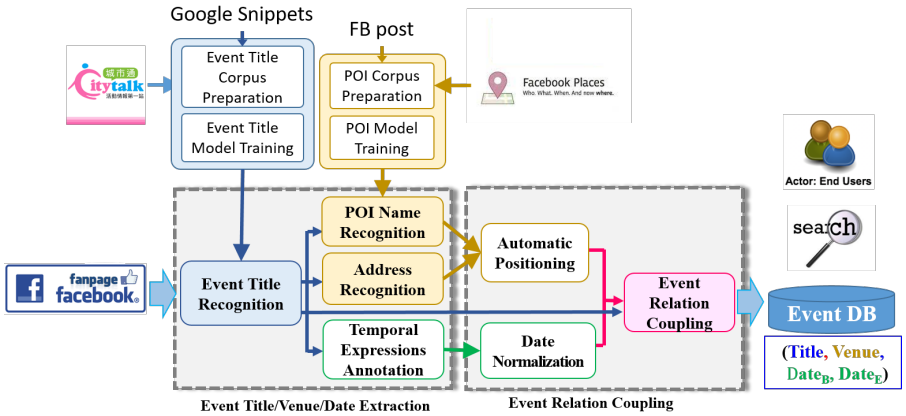

Fig. 2. An overview of the event mining architecture.

Table 1. Seed event titles statistics.

Seed Set	# of Titles	Avg Len	Corpus	# of Snippets
FB events	273,430	16.36	FB posts	3,730 K
CityTalk events	117,736	19.51	Google	1,339 K

Distant supervision is a learning strategy in which a classifier is learned from a weakly labeled training set, which is usually obtained by auto-labeling through seed entities. In this paper, we selected both CityTalk events and Facebook events as seed events, and collected Google snippets and Fanpage posts as our corpus.

To be more specific, We use the event titles from the CityTalk website as queries to collect the top 100 search results by Google search engine for seed-based automatic labeling. For Facebook Fanpage posts, we adopt Apache Solr to index 42.6 million posts from 230K Facebook Fanpages in Taiwan to simulate the Google search engine. Similarly, we use the top 30 posts ranked by the BM25 model for each Facebook event as the training corpus. As shown in Table 1, we collect 1,339K snippets and 3,730K posts from Google and Facebook, respectively, for automatic labeling. This seed-based preparation of automatically labeled training data could greatly save manual labeling costs.

3.2 Seed-based Auto-Labeling

Compared with typical named entities in ACE, event titles are considerably longer and can include various elements, such as the names of persons, locations, organizations, dates and other keywords such as exhibitions and topics. With exact matching (EM), we could only label a limited amount of training data from the crawled Facebook posts through distant supervision, missing a lot of false negative.

Therefore, we employed approximate matching based on the longest common subsequence (LCS) with a threshold of 0.7 to increase the possibility of long titles being matched. However, approximate matching can result in false positives. For example, seed event title “台灣工藝之家府城聯展” (Taiwan Craft House Fucheng Exhibition) can incorrectly label the false positive event “Taiwan Craft House” in the post “國寶級木雕大

師吳榮賜曾榮獲「台灣工藝之家」之尊榮" (National Treasure Wood Carving Master Wu Rongci has won the honor of "Taiwan Craft House").

To reduce false-positive examples, we keep only matched string with the core words in the event title. Core names are special words that differentiate one event from other events. Thus, we referred to Facebook's work on place name deduplication [2] and adopted a core word filtering mechanism to eliminate false positive examples. The idea is to identify core words and background words for each seed event title, and eliminate mentions that did not contain core words during automatic labeling to avoid false-positive mentions caused by the partial match with the event title.

To define formally the method of core word filtering, assume that each event title n consists of at least one *core* word and other *background* words and $|n|$ denotes the number of words in the title n. Let B and C be two probability distributions over dictionary W, where $C(w)$ represents the probability of word w under the core-word distribution and $B(w)$ represents the probability of word w under the background distribution. Let $z(w,n)$ denote a binary event that word w is the core word of an event n, *i.e.*, $z(w,n) = 1$ if $C(n) = w$ and 0 otherwise. The optimization problem is to learn the distribution B, C, and z so that the likelihood of $P(N|B,C,z)$ is maximized via Expectation Maximization algorithm, where the E-step estimate the probability of each $w \in n$ to be the core-word of each known seed n as shown in Eq. (1). The M-step recalculated the probability distribution B, C and z by maximizing the likelihood objective function as shown in Eqs. (2) and (3).

$$z^{(t)}(w,n) = \frac{C^{(t)}(w)/B^{(t)}(w)}{\sum_{w' \in n} C^{(t)}(w')/B^{(t)}(w')} \tag{1}$$

$$C^{(t+1)}(w) = \frac{\sum_{n \in N | w \in n} z^{(t)}(w,n)}{\sum_{n \in N} \sum_{w' \in n} z^{(t)}(w',n)} = \frac{\sum_{n \in N | w \in n} z^{(t)}(w,n)}{|N|} \tag{2}$$

$$B^{(t+1)}(w) = \frac{\sum_{n \in N | w \in n} (1 - z^{(t)}(w,n))}{\sum_{n \in N} \sum_{w' \in n} (1 - z^{(t)}(w',n))} = \frac{\sum_{n \in N | w \in n} (1 - z^{(t)}(w,n))}{(\sum_{n \in N} |n|) - |N|} \tag{3}$$

Table 2 presents the top-ranked core words for some event titles. To establish a suitable balance regarding the occurrence of core words in a mention, we required that one of the top-three-ranking core words be present for a positive training example.

Table 2. Core-background analysis example.

Event title	Top 3 core word ranked by $z(w,n)$		
2009 台中夏季旅展	旅展	2009	台中
2009 Taichung Summer Travel Exhibition	Travel Exhibition	2009	Taichung
施孝榮民歌30音樂會	施孝榮	民歌	音樂會
Shi Xiao Rong Folk Song 30 Concert	Shi Xiao Rong	Folk Song	Concert
曾文40・老照片心故事	曾文	老照片	故事
Zeng Wen 40・Old photo heart story	Zeng Wen	Old photo	story

In addition to approximate matching based LCS, we also adopted an encoding mechanism for numbers and punctuation to increase the matching chance of dates and symbols. We defined six regular expressions for dates and time encodings and 20 rules for

symbolic encodings according to Wikipedia's definition of punctuation terms.[4] For example, "2009台中夏季旅展" (2009 Taichung Summer Travel Exhibition) and "施孝榮民歌30音樂會" (Shi Xiao Rong Folk Song 30 Concert) were converted into "_YEAR台中夏季旅展" and "施孝榮民歌_TITLENUM音樂會".

3.3 Feature Extraction for CRF++

Given labeled training data, the next step in machine learning is to extract features for the sequence labeling models. In this paper, we use conditional random fields (CRF) CRF++ for event title recognition. We refer to DS4NER tool (see [8]) and employ pattern mining to obtain entity prefix/suffix and common before/after n-grams patterns of the event titles as our features. Two additional features are whether the current token is alphanumeric or contains a special symbol. A total of 15 original features and 215 templates are employed for the CRF++ model. The feature template used for the CRF++ model is given at our GitHub.[5]

3.4 Testing Dataset and Event Title Evaluation

To evaluate the performance of event title recognition, we randomly selected 1,300 posts that contained event announcements from Facebook Fanpages and manually annotated 2,199 event titles.

Table 3. Testing data statistics – 1300 FB fanpage posts.

# Testing Posts	# Sent.	Sent Len	# Entity Titles	# POIs	# Start/End Date
1,300	28,856	34	2,199	2,030	1,192 / 805

Because an event title is considerably longer than a typical named entity, we employed a partial scoring method for performance evaluation. For each extracted title e that overlaps with the true answer title a, we defined the $P\text{-}core(e,a)$, $R\text{-}score(e,a)$ scores, as presented in Eq. (4). To calculate the intersection and unification of e and a, we use Chinese characters (*i.e.*, uni-gram) of the golden answer and extracted string as basic units. The precision and recall were averaged over all the extracted titles E and true answer titles A, respectively as presented in Eq. (5).

$$\text{P-score}(e,a) = \frac{P(e \cap a)}{|e|}, \quad \text{R-score}(e,a) = \frac{P(e \cap a)}{|a|} \tag{4}$$

$$\text{Precision} = \frac{\sum_{e \in E} P - score(e,a)}{|E|}, \text{Recall} = \frac{\sum_{a \in A} R - score(e,a)}{|A|} \tag{5}$$

$$\text{F1-score} = \frac{2(\text{Precision} \times \text{Recall})}{\text{Precision+Recall}} \tag{6}$$

Table 4 presents a comparison of the models trained with various labeling strategies on

[4] https://en.wikipedia.org/wiki/Punctuation

[5] https://github.com/formatchou/DS4NER/blob/master/crfpp/template_dictionary.txt

the Facebook Fanpage corpus and Google search snippets using Facebook or CityTalk as seed events for auto-labeling.

The baseline approach based on exact match yielded F1 values of approximately 0.004∼0.09 and 0.07∼0.146 with FB events and CityTalk, respectively (as presented in the rightmost column of Table 4). These results indicate that dictionary-based exact matching has very limited performance.

Table 4. Performance of CRF++ models using various training data preparation.

Corpus+Labeling	Seed event	# Sentences	P	R	F1	Baseline Dict-EM
FB-EM (CRF++)	FB Event	243,035	0.336	0.267	0.298	0.070
FB-LCS (CRF++)	FB Event	537,019	0.266	0.306	0.285	0.140
FB-LCS-Core (CRF++)	FB Event	287,653	0.344	0.319	0.331	0.146
G-EM (CRF++)	CityTalk	95,532	0.608	0.300	0.402	0.004
G-LCS (CRF++)	CityTalk	305,797	0.581	0.477	0.524	0.090
G-LCS-Core (CRF++)	CityTalk	263,955	0.595	0.471	0.526	0.070
FB-CRF (CRF++)	G-LCS-Core	603,925	0.602	0.532	**0.565**	N/A

We consider three labeling strategies: exact match (EM), longest common subsequence (LCS), and core word filtering, and use CRF++ sequence labeling tool for event title extraction from both Google (G) and Facebook (FB) corpus. The interesting thing is that although the test set is collected from Facebook posts and more seed events are matched in the FB training corpus, the models trained on the Google corpus are far superior to the models trained on the Facebook corpus. As we all know, the size of the training corpus and the correctness of the labeled training data play an important role in the model performance. Compared with the Google corpus labeled by exact matching, more training data is added through approximate matching (95K to 305K), and the F1 performance of the former model is improved from 0.402 to 0.524. However, the same mechanism on the Facebook corpus will reduce F1 performance (exact match is 0.298 F1, approximate match is 0.285 F1), indicating too many noisy data are introduced in the Facebook corpus through approximate matching (243K vs. 573K).

With the core word filtering mechanism, the model trained from Facebook corpus has an 16.1% F1 improvement over the model trained from approximate matching (0.331 vs. 0.285). However, core word filtering only slightly improves the performance of the model trained on the Google Corpus (from 0.523 to 0.526 F1). One possible reason is that core word filtering removes too many good sentences from Google search results (305K to 263K). Overall, we can say the retrieval of sentences by Google are much more accurate than the Solr retrieval system we build for Facebook Fanpage posts.

3.5 Model-based Auto-Labeling

Since seed-based distant supervision does not work well on the test data, we consider the possibility of model-based distant supervision to prepare the training corpus. We selected Facebook Fanpage posts that contain keywords such as "展覽" (exhibition), "演講" (lecture), "比賽" (competition), and "活動" (activity) as the corpus. Applying the G-LCS-Core model to the 13 million Fanpage posts, we obtain 770K posts with at least one

event title recognized. We sample 80K posts (from 770K posts) and only keep paragraphs that contain annotated event titles in these 80K posts to obtain a total of 603,925 sentences (including 88,855 event titles) and train a new model with the CRF++ tool kit. The idea is similar to the retraining strategy in *data programming* [9], which is a paradigm for the programmatic creation and modeling of training datasets.

The performance of the retrained model is presented in the last row of Table 4. With the new annotated data, we obtained an F1 score of 0.565, substantially higher than the F1 values (0.298 to 0.331) of the seed-based auto-labeling on FB corpus. We conducted t-test of the new model with the best model from Google corpus, and obtained a p-value of 0.0011, showing the improvement in F1 is significant.

3.6 Performance of BERT-Based Models

Finally, to see how state-of-the-art deep learning models perform on this task, we also trained BERT-based models (with CRF added to the last layer) using the G-LCS and G-LCS-Core training corpus. The results show that the BERT-based models achieve only 0.195 to 0.188 F1, which is much lower than the CRF++ models (0.524 to 0.526 F1). We speculate that seed-based auto-labeling contains too many false positives and false negative examples, resulting in poor model performance.

On the other hand, in terms of model-based distant supervision, the BERT-based model achieves 0.573 F1, outperforming the trained CRF++ model (0.565 F1). This suggests that we can achieve better performance with state-of-the-art deep learning models when the annotated training data is of a certain quality.

Table 5. Performance of BERT-based models for various event title recognition.

Corpus+Labeling	Seed event	# Sentences	P	R	F1	Baseline Dict-EM
G-LCS (BERT)	CityTalk	305,797	0.113	0.728	0.195	0.090
G-LCS-Core (BERT)	CityTalk	263,955	0.107	0.756	0.188	0.070
FB-CRF (CRF++)	G-LCS-Core	603,925	0.602	0.532	**0.565**	N/A
FB-CRF (BERT)	G-LCS-Core	603,925	0.641	0.518	**0.573**	N/A

Finally, since multiple event titles may be recognized by the trained event title recognition model, we compare each event title with all other recognized event titles and select the event title with the highest overlap ratio as the main event title for a post. Thus, the next task is to extract the corresponding event venue for the selected event title.

4. EVENT VENUE RECOGNITION AND EVENT ATTRIBUTE COUPLING

Event venues can be as specific as addresses or as large as the Geographical Political Entity (GPE). To make it easier to remember, event venues are usually associated with POIs such as landmarks, organization or business names. In this paper, we referred to [10] for address extraction and focused on POI extraction using Facebook PlaceDB as our seed POIs.

4.1 POI Recognition Model Training and Evaluation

Similar to event title recognition, we use an automatic labeling strategy to annotate the retrieved posts (from the information retrieval system) or returned snippets (from the Google search engine) for each given POI name in the *Facebook PlaceDB*. However, not all Facebook check-in POIs are equally good for automatic labeling, because people may abbreviate or simplify location names. For example, "中埔分局" (Zhongpu Precinct), appears to be a POI; however, "嘉義縣警察局中埔分局" (Jhongpu Precinct of Chiayi County Police Department) is considerably more precise. Moreover, the location database contained many long POI names that people arbitrarily used to specify exact locations, such as "文化部所屬國立中正紀念堂管理處志清廳" (Zhiqing Hall of the National Chiang Kai-shek Memorial Hall Management Office of the Ministry of Culture).

To overcome this challenge and obtain a high-quality training corpus, we used two POI selection mechanisms: reliability and popularity. The former refers to POIs with more than 200 check-ins and likes, while the later refers to POIs with more than 1,000 check-ins. See Table 6 for the number of POIs filtered by the two mechanisms.

Table 6. Seed POI filtering mechanisms and seed length distribution.

Seed Set	Criteria	# of seeds	Len<=5	Len>5
All POIs	None	865K	32%	68%
Reliable POIs	#checkin>200 & #likes>200	497K	30%	70%
Popular POIs	#checkin>1000	103K	31%	69%

We tested the POI selection mechanisms on the 1,300 event posts mentioned in Table 3 for event venue recognition. For comparison, we use the Stanford Named Entity Recognizer (NER) as the baseline. As shown in Table 7, dictionary-based matching with the Facebook PlaceDB achieves best F1 with 0.367 when popular POIs are used.

Table 7. Comparison of POI recognition models.

Model		# Training Sentences	# Extract	P	R	F1
Baseline - Stanford NLP		–	134	0.137	0.017	0.031
FB-Dict	All	–	6,205	0.222	0.558	0.317
	Reliable	–	4,640	0.271	0.511	0.354
	Popular	–	2,828	0.345	0.393	0.367
FB-POI-CRF	All	136.9M	7,360	0.222	0.695	0.337
	Reliable	79.9M	6,065	0.258	0.671	0.372
	Popular	3.3M	3,268	0.451	0.658	**0.536**
G-POI-CRF	Popular	3.9M	1,256	0.468	0.265	0.339

Next, we compared the performance of the models trained from Facebook posts (called FB-POI) and Google snippets (called G-POI). For FB-POI corpus, we collect the top 30 posts for each POI query. For G-POI corpus, we only use popular POIs as queries to collect the top 150 Google search snippets to generate 170k sentences and annotate

62K location mentions. We can see from Table 7 that the performance of the trained model based on the Facebook corpus with popular POI as seeds has the best performance of 0.536 F1. The common problem of these models is the low precision due to too many extractions (from 3,268 to 7,360). Instead, the trained model based on the G-POI corpus extracts only 1,256 POIs and achieves the best precision 0.468 but lowest recall 0.265.

4.2 Event Title and Venue Coupling

Since multiple addresses, GPEs, or POIs may be mentioned in the post, the next question arises is how to determine which POI name is the event venue. To solve this issue, we conducted sequential pattern mining on sentences that contain mentions of event venues to discover patterns that can identify correct venues. The procedure is outlined below:

- First, we use the trained event title recognition model to annotate 13M posts containing event-related keywords, *e.g.* "exhibition," "concert," and "competition," from Facebook Fanpages. A total of 770K posts are found to contain at least one event title.
- Next, we use the FB-POI-CRF model (popular) and the address extraction model in [10] to annotate the 770K FB posts, and select sentences containing POI names for sequential pattern mining. A similar approach is used for start and end dates. A total of 350K sequences were found to contain either POIs or addresses.
- Finally, We used the Jieba word segmentation tool to segment 340K sentences into word tokens and replace POIs and addresses with a special "VENUE" token. We adopt BIDE algorithm [11] which is implemented in the open source data mining library SPMF [12] for sequential pattern mining.

Table 8 presents the common words before or after an event venue. We manually examined the top 800 patterns with the highest support to select 43 patterns. The patterns that preceded event venues included "位於" (located in), "報到" (register at), "表演場地" (performance venue), and "報到地點" (registration location), while common-after patterns were mostly verbs like "舉行" (hold), "舉辦" (host), "推出" (launch), and "舉辦活動" (hold an event).

Table 8. Common words before and after VENUE tokens.

Rule	Common Before	Common After
1-Pattern	位於(located in), 報到(register at), 即將(coming soon)	舉行(hold), 舉辦(host), 推出(launch)
2-pattern	表演場地(performance venue), 報到地點(registration location)	舉辦活動(hold an event)

Table 9 shows the experimental result of event venue coupling through pattern filtering. Note that if all recognized POIs match with no pattern, we directly regard the POI with the highest marginal probability as the event venue. As we can see, filtering addresses with the 43 patterns is effective to select the event venue with 0.890 F1. This is

because explicitly mentioned addresses are usually the event venue. As for POI, applying patterns can successfully filter 965 event venues from 3,268 POIs. However, the pattern filtering for POI only decreases the F1 score from 0.536 to 0.216.

Table 9. Performance of event venue and date coupling.

	Golden Ans	Pattern Filtering	P-Score	R-Score	P	R	F1
Venue - POI	910	3,268 → 965	320.1	146.1	0.332	0.161	0.216
Venue - Address	440	524→ 442	398.3	387.1	0.901	0.880	0.890
Venue - All	974	3,792→1030	601.9	533.2	0.584	0.548	0.565
Start Date	1,192	6,294→1,214	1,012	1,012	0.834	0.849	0.841
End Date	805	6,294→510	469	469	0.920	0.583	0.713

We apply the same sequential pattern mining approaches to start/end date as well to select the proper event start/end date. The results are shown in the last two rows of Table 9, where the F1 score achieves 0.841 and 0.713 for start/end date, respectively.

Finally, we show the performance of the complete event extraction in Table 10. An event is composed of four arguments: title, venue, start date, and end date. Among the 1,300 test posts, most (584+574) events mentioned four or three arguments (called 4- or 3-tuple, respectively). A few events (124+45) mentioned only two or one argument (2-tuple and 1-tuple). The precision/recall of a k-tuple event is the average P-score and R-score of the k fields, respectively. The F1 of a k-tuple is the harmonic mean of the precision and recall for the event. The performance of 1-tuple event reaches 0.767 F1, while the performance of 2-tuple or 3-tuple events achieve 0.633. Extracting all arguments of 4-tuple events is more challenging than the other three event types. The extraction performance of 4-tuple events is 0.593 F1. On average, the performance of event extraction with correct coupling is 0.620 F1.

Table 10. Tuples of event performance.

k-Event	# of Events	P	R	F1
4-tuples	584	0.545	0.650	0.593
3-tuples	547	0.589	0.686	0.634
2-tuples	124	0.624	0.642	0.633
1-tuples	45	0.776	0.758	0.767
Total	1,300	0.579	0.668	0.620

4.3 Event Dynamics

Finally, we demonstrate the collection of social events from Accupass, Facebook Event and Fanpage from 2016 to 2019. As displayed in Fig. 3, the number of events submitted to Facebook Event gradually increases from monthly 4.6K (2016) to 23K (2019). In addition, the number of extracted events from Facebook Fanpages per month increases slightly from 6K (in 2016) to 10K (2019). In practice, the event overlap percentage between Facebook Event and Fanpages is less than 10%. In other words, event organizers

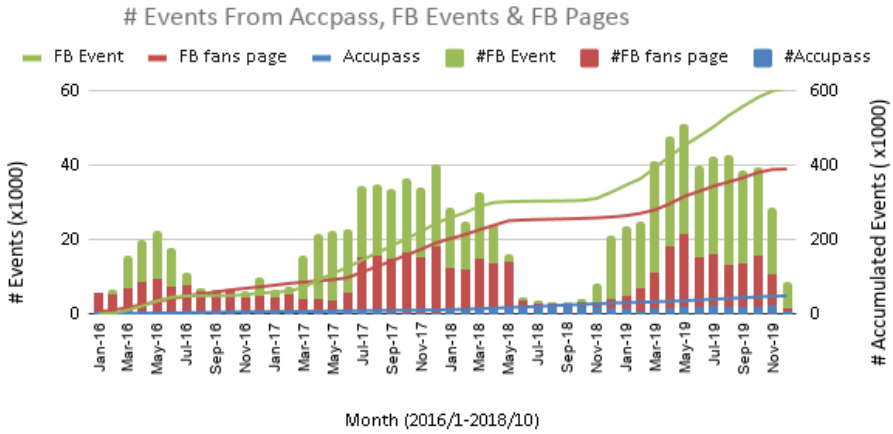

Fig. 3. The number of events collected during 2016 and 2019. Due to the the Facebook Cambridge Analytica data scandal, Facebook ended some of the Facebook Graph API functions after May 2018. Therefore, we have missed more than a half year data in 2018.

tend to choose either Fanpage or Facebook Event to publish an event, but seldom do both. Therefore, it is necessary to identify events from Facebook Fanpage posts as well. Exploring the dynamics of events through social-media posts has become a key issue in our research. The interface of our social event search service *EventGo!* can be found at the following URL (`https://eventgo.widm.csie.ncu.edu.tw/`).

5. CONCLUSION AND FUTURE WORK

In this study, we considered event title recognition a critical task for event identification and extraction. Therefore, we only conducted address extraction and POI recognition for posts containing an event title; and coupling the extracted address and POIs with event title through manually selected rules from sequential pattern mining. For event title recognition, we proposed the idea of core word filtering to diminish the effect of approximate matching for long event titles. Furthermore, we proposed model-based distant supervision with the best model trained from Google corpus to produce new annotated Facebook posts of better quality and improve the event title recognition rate from 0.331 to 0.565.

For event venue extraction, we adopted both address extraction and POI recognition model trained from 103K popular Facebook place names. Finally, we applied a sequential pattern mining approach to select the extracted address or POIs. Overall, event extraction achieve an average of 0.620 F1. For future work, we plan to exploit the trained model to other event resources on the Web. For example, city governments and organization websites usually announce speech/activity on the recent news. If we could discover such event resources and monitor such websites for new event announcements, we can resolve the dependency on Facebooks APIs.

REFERENCES

1. Q. Wang, B. Kanagal, V. Garg, and D. Sivakumar, "Constructing a comprehensive events database from the web," in *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, 2019, pp. 229-238.
2. N. Dalvi, M. Olteanu, M. Raghavan, and P. Bohannon, "Deduplicating a places database," in *Proceedings of the 23rd International Conference on World Wide Web*, 2014, pp. 409-418.
3. J. Foley, M. Bendersky, and V. Josifovski, "Learning to extract local events from the web," in *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2015, pp. 423-432.
4. B. Yang and T. M. Mitchell, "Joint extraction of events and entities within a document context," in *Proceedings of Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2016, pp. 289-299.
5. A. Ritter, Mausam, O. Etzioni, and S. Clark, "Open domain event extraction from twitter," in *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2012, pp. 1104-1112.
6. Y.-H. Lin, C.-H. Chang, and H.-M. Chuang, "Mining events through activity title extraction and venue coupling," in *Proceedings of International Conference on Technologies and Applications of Artificial Intelligence*, 2020, pp. 136-141.
7. C.-H. Chang, Y.-H. Lin, and H.-M. Chuang, "Eventgo! exploring event dynamics from social-media posts," in *Proceedings of International Computer Symposium*, 2020, pp. 548-552.
8. C.-L. Chou, C.-H. Chang, Y.-H. Lin, and K.-C. Chien, "On the construction of web NER model training tool based on distant supervision," *ACM Transactions on Asian and Low-Resource Language Information Processing*, Vol. 19, 2020, Article 87.
9. A. J. Ratner, C. M. De Sa, S. Wu, D. Selsam, and C. R´e, "Data programming: Creating large training sets, quickly," in *Advances in Neural Information Processing Systems*, D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, (eds.), Vol. 29, Curran Associates, Inc., 2016.
10. C.-H. Chang and S.-Y. Li, "Mapmarker: extraction of postal addresses and associated information for general web pages," in *Proceedings of IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, 2010, pp. 105-111.
11. J. Wang and J. Han, "Bide: efficient mining of frequent closed sequences," in *Proceedings of the 20th International Conference on Data Engineering*, 2004, pp. 79-90.
12. P. Fournier-Viger, J. C.-W. Lin, A. Gomariz, T. Gueniche, A. Soltani, Z. Deng, and H. T. Lam, "The SPMF open-source data mining library version 2," in *Machine Learning and Knowledge Discovery in Databases*, B. Berendt, B. Bringmann, É. Fromont, G. Garriga, P. Miettinen, N. Tatti, and V. Tresp, (eds.), Springer, Cham, 2016, pp. 36-40.

Yuan-Hao Lin is a Ph.D. student in Department of Computer Science and Information Engineering from National Central University, Taiwan since 2019. He received an MS degree from the same department in National Central University in 2016.

Chia-Hui Chang is a Full Professor at National Central University, Taiwan. She obtained her Ph.D. in Computer Science and Information Engineering from National Taiwan University, Taiwan in 1999. Her research interests focus on information extraction, web intelligence, data mining, machine learning and system integration. Dr. Chang has published more than 80 papers at refereed conferences and journals including WWW, PAKDD, TKDE, IEEE Intelligent Systems, *etc*. She was the President of Taiwan Association for Artificial Intelligence (TAAI) and the Association for Computational Linguistics and Chinese Language Processing (ACLCLP) from 2020 to 2021.

Hsiu-Min Chuang is an Assistant Professor at the Chung Yuan Christian University, Taiwan. She received an MS degree in Information and Computer Education from the National Tainan University in 2006 and Ph.D. degree in Computer Science and Information Engineering from the National Central University in 2016. Her research interests include web mining, information extraction and natural language processing.