Fine-Grained Location Extraction from Tweets with Temporal Awareness

Chenliang Li[1], Aixin Sun[2]

[1]Wuhan University, China

[2]Nanyang Technological University, Singapore

Agenda

- Introduction on Twitter
- Inferring Locations from Tweets
  - Vision for temporal awareness resolution for locations
  - Existing related works review
- ***Petar*** for POI Extraction with Temporal Awareness
  - POI Inventory
  - Data analysis and observations
  - Time-aware POI tagger
  - Efficiency issue
- Experimental Results

武漢大學 Wuhan University

NANYANG TECHNOLOGICAL UNIVERSITY

Introduction on Twitter

- **Large** volume of **timely** data
  - 200+ million active users worldwide every month
  - Users share about their mood, activities, and opinions through a short message, called a tweet.
- Social and business **value**
  - Event detection and summarization
  - Users’ opinion, crisis detection and response
  - Business marketing and advertising
- **Challenges**
  - Grammar errors, misspellings, informal abbreviations…
  - Tweets are (very) **short**
  - **Effectiveness** and **Efficiency** are both important

武漢大學 Wuhan University
NANYANG TECHNOLOGICAL UNIVERSITY

Inferring Locations from Tweets

- Through tweets, users casually or implicitly reveal their **locations** and **short term visiting plans**

- Extracting fine-grained locations (point-of-interest) from tweets with **temporal awareness**
  - The user has visited, is currently at or will soon visit the POI.
  - Support precise location-based services/marketing and personalization

just back from **L'Artusi**, wonderful dinner :> like to try **the smile** tmr for lunch.

L'Artusi — The user **has just visited** this restaurant

The Smile — The user **will soon** visit this restaurant

Inferring Locations from Tweets: Existing Studies

- Build **spatial language model** from the spatial usage of words
  - [Cheng et al. CIKM10; Chang et al. ASONAM12; Kinsella et al. SMUC11; Li et al. CIKM11]
- **Gazetteers** and **external knowledge** like Geonames, DBPedia Spotlight are used to derive the locations
  - [Mahmud et al. ICWSM12; Schulz et al. ICWSM13]
- **Latent variable models** are used to analyze the interplay between geographic locations, topics and users' interests
  - [Eisenstein et al. EMNLP10; Hong et al. WWW12; Yuan et al. KDD13]

sit at mac, enjoying a big mac

Ambiguity of a specific POI ???
Temporal awareness of a specific POI ???

Inferring Locations from Tweets: Challenges

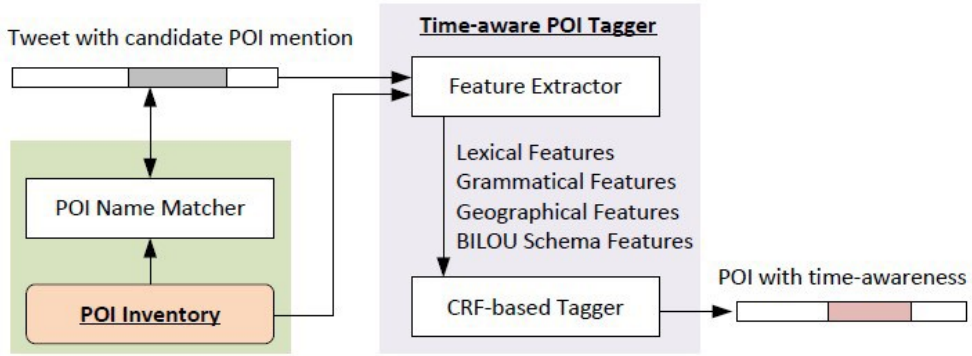
- **POI** extraction with **temporal awareness** from tweets is difficult
  - POI a focused geographic entity or a specific point location [Lingad et al. WWW13; Rae et al. SIGIR12]
- Predominate usage of **short names** or **informal abbreviations**
  - Existing NER techniques for location detection experience a significant performance degradation.
  - Capturing temporal awareness based on **existing** temporal expression extraction tools become **less practical**
- Many POI names are **ambiguous**
  - Pre-built gazetteer leads to an ineffective solution

mac => McDonald's chain restaurant
Apple's product
McDonald's product

武漢大學 Wuhan University
NANYANG TECHNOLOGICAL UNIVERSITY

Petar for POI Extraction with Temporal Awareness

- Overview of *Petar*:

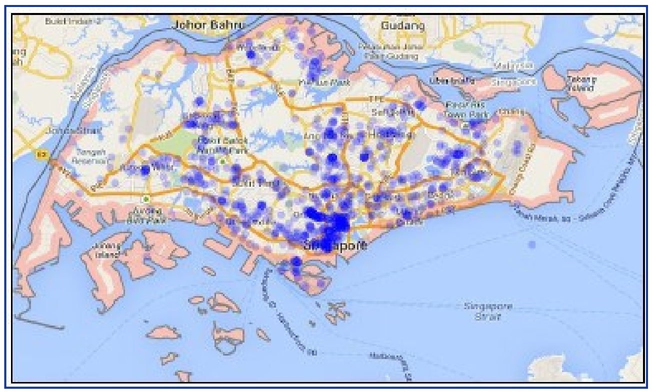


- Focus on a predefined geographical region (e.g., a city)
  - Tweets from Singaporean users
- POI Inventory
  - A collection of candidate POI names, each of which may refer to a POI.
- Time-aware POI Tagger
  - Simultaneously disambiguate the candidate POI name and resolve its temporal awareness.

Wuhan University NANYANG TECHNOLOGICAL UNIVERSITY

POI Inventory: exploiting *Foursquare* community

- 239,499 Foursquare check-in tweets made by Singaporean users
- The POI coverage is **broad** or even **exhaustive** in a fine-grained scale



POIs covered by 1K sample Foursquare check-ins

  - Each check-in tweet is well formatted and associated with a latitude/longitude coordinate

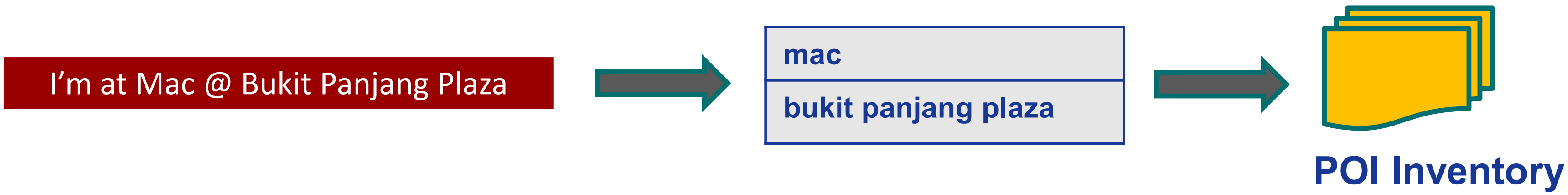
t_1	I'm at Mac @ Bukit Panjang Plaza
t_2	I'm at ITE College Central MacPherson Campus Main (201 Circuit Road)
t_3	Birthday dinner (@ Ambush @ JP w/ 2 others)
t_4	Watching "Hello Stranger" (@ Golden Village Cinema 9 @ Plaza Singapura)

User's current location

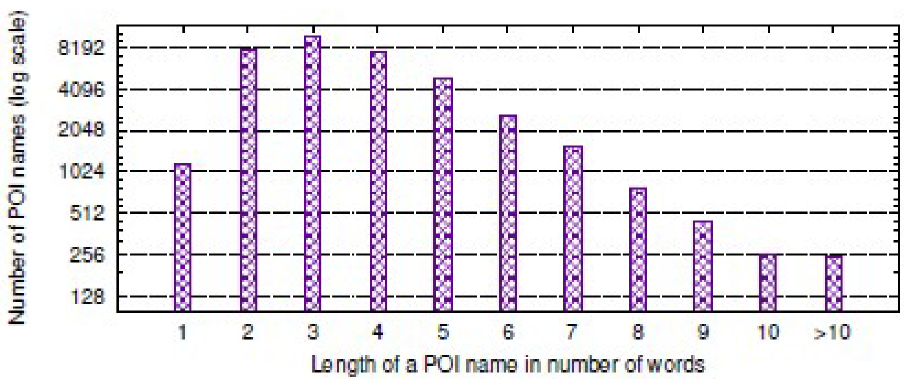
User's activities at the location

POI Inventory

- POI names are extracted by applying handcrafted rules



- 37,160 POI names are extracted, the average length is **3.9 words**

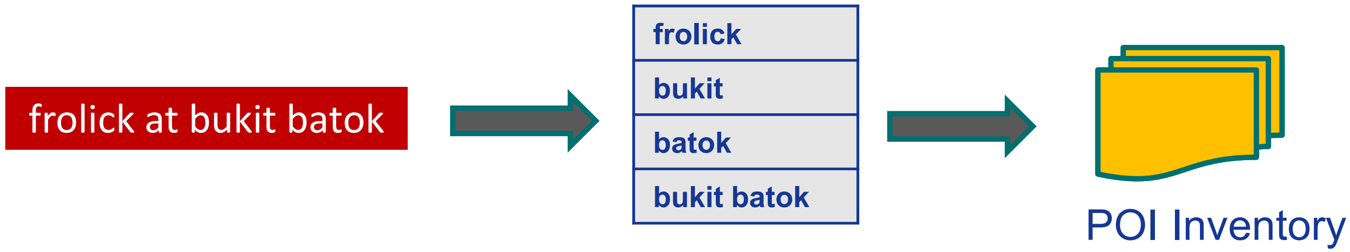


Most POI names are in the range of 2 to 5 words

- A tweet is very short, people often mention POIs with **abbreviations / partial names**, assuming the audience's context-awareness [Lieberman et al. ICDE10]

POI Inventory

- Partial POI names are extracted by taking all the sub-sequences of the names (up to 5 words)



  - Stopwords are ignored and used as separators
  - Filtering is conducted to remove infrequent candidate POI names

- Not all candidate POI names are valid
  - noisy data is included as well: “my room”, “my work place”, “my bed”
  - the candidate POI mention **may not** be a **true POI** in a tweet (i.e., **ambiguity**)

Data Analysis and Observations

- Data Sets
  - 4.33M tweets from 19,256 unique Singapore-based users during June 2010
  - 222,201 tweets mentions at least one candidate POI name (5.1%)

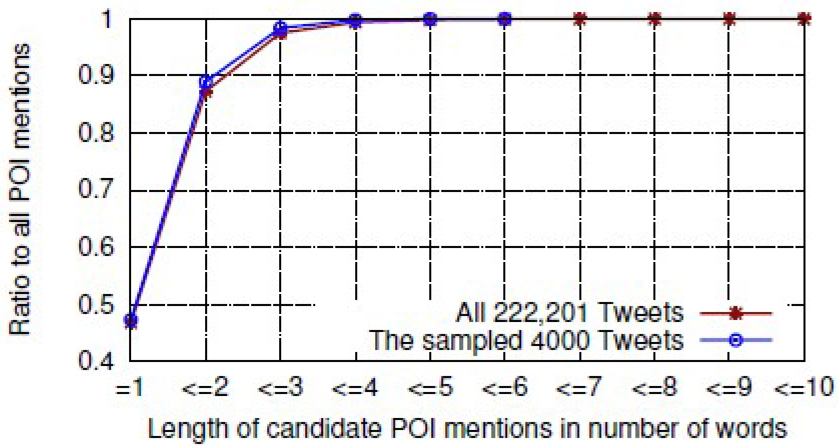
- Observation 1:
  Many users reveal their fine-grained locations in their tweets.
  - 71.4% of all users in the dataset
  - **91.3%** of the users who had published at least 20 tweets

Casually or implicitly reveal their locations in the form of **fine-grained POIs** like restaurant or shopping mall names

Data Analysis and Observations

- Observation 2:
  The candidate POI mentions are mostly very short with one or two words. Many of the mentions are partial location names.

  - 46.7% of the candidate POI names are **unigrams** (likely to be ambiguous)
  - 41.6%+ of the candidate POI names are **partial POI names**
  - POI names with **3 or more words** com are about **2.5%** only.



Data Analysis and Observations

- Observation 3:
  *About **half** of the candidate POI mentions indeed refer to locations and their associated temporal awareness can be determined.*

- 4,000 tweets are sampled from these 222,201 tweets for manual annotation
  - The **previous** and **following two tweets** (and their timestamps) from the same user are used as context for annotation.
  - Search engine is allowed for context understanding

14.9%	58.5%	26.6%			
$\#POI_p$	$\#POI_z$	$\#POI_f$	$\#NPOI$	$\#Unknown$	Total
307	1,202	547	1,801	120	3,977
Total $\#POI$s: 2,056			-	-	

906 distinct candidate POI names

51.7% are truly locations

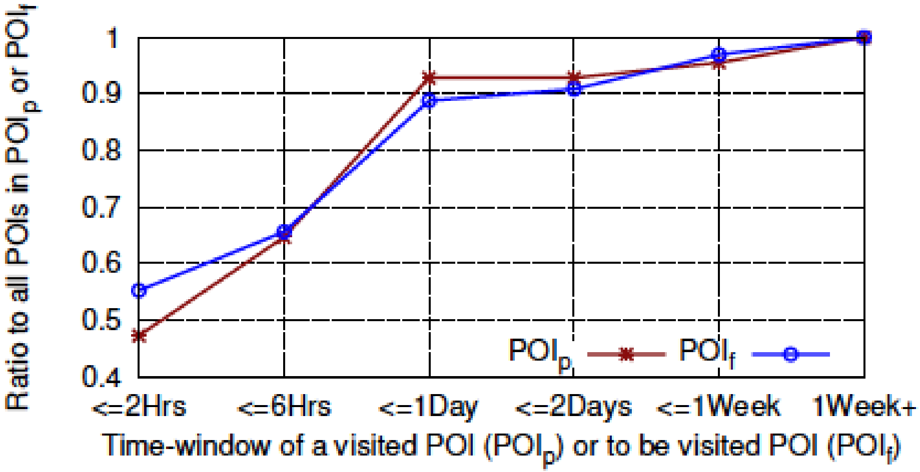
武漢大學 Wuhan University

NANYANG TECHNOLOGICAL UNIVERSITY

Data Analysis and Observations

- Observation 4:
  *Among all POIs that were visited, or to be visited, about **90%** of the visits to these POIs happen **within a day**.*

- Temporal awareness of POIs in ***POI_p*** and ***POI_f*** (854 POIs)



Efficiency is an important factor for fine-grained location-based services/marketing

Time-Aware POI Tagger

- Simultaneously **disambiguate** the candidate POI mentions and resolve the **temporal awareness**
  - Fast learning and inference
- Contextual knowledge is very important

Off to jp now! Hope it DOESN'T rain

"*jp*" is POI_f

- Linear-Chain Conditional Random Field (CRF)
  - Four types of features are investigated
  - lexical, grammatical, geographical and BILOU schema features.

Wuhan University

NANYANG TECHNOLOGICAL UNIVERSITY

Time-Aware POI Tagger: Lexical Features

- Basic lexical features
  - Word itself, lowercase form, prefixes & suffixes;
  - Word shape (all-cap., is-cap., all-numeric, alphanumeric);
  - Prior probability of being in cap. or all-cap. (discretized as binary features)
- Contextual features
  - BOW of context window up to 5 words;

    Off to jp now ! Hope it DOESN't rain
  - BOW of the preceding 2 words (left-hand side window);

    Off to jp now ! Hope it DOESN't rain
  - BOW of the following 2 words (right-hand side window);

    Off to jp now ! Hope it DOESN't rain

武漢大學 Wuhan University

NANYANG TECHNOLOGICAL UNIVERSITY

Time-Aware POI Tagger: Grammatical Features

- Part-of-speech (POS) (TwitterNLP [Ritter et al. EMNLP11])
- Word group by Brown clustering
- Time-trend score of tweet
  - A dictionary of 36 commonly used English words with their time-trend scores is manually compiled (time-trend dictionary);
- The closest verb.
  - Verb., the tense of the verb., distance, left/right-hand indicator
- The closest time-trend word
  - Word, time-trend score, distance, left/right-hand indicator

Time-Aware POI Tagger

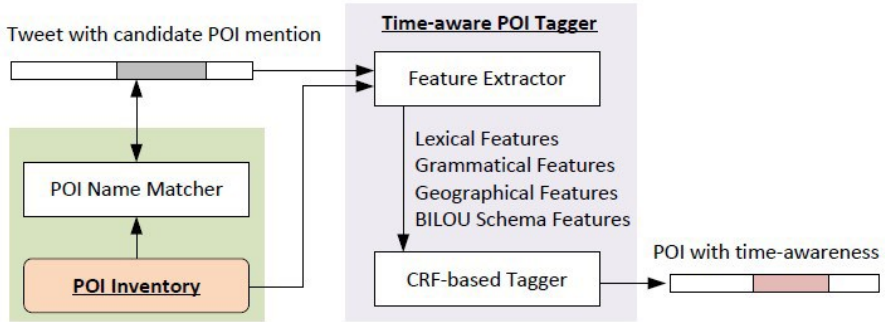
- Geographical Features
  - Spatial randomness of each candidate POI name
  - Location name confidence
  - Multiple candidate POI mention
- BILOU Schema Features
  - Identify **B**eginning, **I**nside and **L**ast word of a multi-word POI name, and **U**nit-length POI name, and the words **O**utside of any POI names
  - Each candidate POI name is pre-labeled with BILOU schema

We're all for Asian delights! Thai express today, suki sushi tomorrow

We're\O all\O for\O Asian\O delights\O ! \O Thai\B express\L today\O ,\O suki\B sushi\L tomorrow\O

Time-Aware POI Tagger

- Efficiency
  - Scan each tweet against POI inventory
    - Prefix-tree algorithm with linear complexity [Li et al. JASIST13]
  - Most features are simple to derive
    - Brown clustering, geographical features are pre-computed
    - Linear-chain CRF inference for POS tagging
  - Overall linear-chain CRF inference x 2



- Quantitative results
  - 1.86GHz Xeon quad-core and 12GB RAM
  - 400 raw tweets per second (i.e., 1.44M/Hr) by using a single CPU core

武漢大學 Wuhan University
NANYANG TECHNOLOGICAL UNIVERSITY

Experiments: settings

Experiment Setup

- NPOIs and true POIs with their temporal awareness
- 5-fold cross validation is applied
- Evaluation metrics: Precision, Recall and F1

#POI$_p$	#POI$_z$	#POI$_f$	#NPOI	Total
307	1,202	547	1801	3,857
Total #POIs: 2,056				

Comparative Methods

- Random Annotation (RA)
  - A candidate POI mention is randomly assigned as POI$_f$, POI$_z$, POI$_p$, NPOI
- K-Nearest Neighbor (KNN)
  - A candidate POI mention is represented by its surrounding 4 words

$$sim(\ell_a, \ell_b) = \frac{|\mathcal{W}_a \cap \mathcal{W}_b|}{|\mathcal{W}_a \cup \mathcal{W}_b|}$$

- StanfordNER (CRF-Classifier)
  - Mainly **lexical features** are used in this system.
  - **POI inventory** is provided as an **external gazetteer**

Wuhan University 武漢大學

NANYANG TECHNOLOGICAL UNIVERSITY

Experiments: results

- POI extraction with temporal awareness
  - Petar > StanfordNER >> KNN >> RA

Method	POI_f			POI_z			POI_p		
	Pr	Re	F_1	Pr	Re	F_1	Pr	Re	F_1
RA	0.1438	0.2464	0.1816	0.3040	0.2582	0.2792	0.0795	0.2426	0.1197
KNN	0.4622	0.2792	0.3481	0.5685	0.4593	0.5081	0.1333	0.0066	0.0125
StanfordNER	0.5701	0.4526	0.5046	0.5886	0.5264	0.5558	0.3147	0.1475	0.2009
PETAR	**0.6895**	**0.5511**	**0.6126**	**0.6752**	**0.7108**	**0.6925**	**0.5266**	**0.3574**	**0.4258**

  - POI_f, POI_z >> POI_p

- Disambiguating POIs (ignoring temporal awareness)
  - Petar > StanfordNER

Method	POI			$NPOI$		
	Pr	Re	F_1	Pr	Re	F_1
RA	0.5254	0.7419	0.6152	0.4509	0.2428	0.3156
KNN	0.7761	0.4980	0.6067	0.5948	0.8385	0.6959
StanfordNER	**0.9397**	0.6931	0.7977	0.7308	**0.9493**	0.8259
PETAR	0.9094	**0.8436**	**0.8753**	**0.8354**	0.9042	**0.8684**

武漢大學 Wuhan University

NANYANG TECHNOLOGICAL UNIVERSITY

Experiments: Feature Analysis

- ***Lex*** features are better for POI mention **disambiguation**

Feature	POI			NPOI		
	Pr	Re	F_1	Pr	Re	F_1
Lexical	0.9161	0.8109	0.8603	0.8095	0.9154	0.8592
Grammatical	0.8688	0.8152	0.8411	0.8033	0.8597	0.8306
Geographical	0.7787	0.5762	0.6624	0.6276	0.8135	0.7085

- ***Gra*** features are better for resolving **temporal awareness**

Feature	POI_f			POI_z			POI_p		
	Pr	Re	F_1	Pr	Re	F_1	Pr	Re	F_1
Lexical	0.4727	0.2682	0.3423	0.5701	0.6915	0.6250	0.2264	0.0393	0.0670
Grammatical	0.6525	0.5310	0.5855	0.6425	0.6764	0.6590	0.4727	0.3410	0.3962
Geographical	0.1667	0.0055	0.0106	0.4519	0.5666	0.5028	0	0	0

- ***Lex + Gra*** features are better in most cases

Feature	POI_f			POI_z			POI_p			POI		
	Pr	Re	F_1	Pr	Re	F_1	Pr	Re	F_1	Pr	Re	F_1
Gra+Geo	0.6453	0.5191	0.5753	0.6480	0.6858	0.6663	0.5026	**0.3725**	0.4279	0.8741	0.8241	0.8484
Lex+Gra	**0.6895**	**0.5511**	**0.6126**	**0.6752**	**0.7108**	**0.6925**	**0.5266**	0.3574	0.4258	0.9094	0.8436	**0.8753**
Lex+Geo	0.4748	0.2755	0.3487	0.5811	0.6873	0.6298	0.2373	0.0459	0.0769	**0.9206**	0.8045	0.8586
Lex+Gra+Geo	0.6788	0.5438	0.6039	0.6712	0.7083	0.6892	0.5211	0.3639	**0.4286**	0.8702	**0.8709**	0.8706

武漢大學 Wuhan University

NANYANG TECHNOLOGICAL UNIVERSITY

Experiments: Feature Analysis

➢ Effectiveness of individual features in *Lex + Gra*

- 5 features is considered by removing it from *Lex + Gra* combination

Feature	Description
ContextWindow	The BOW of 5-word context window, the preceding two words, and following two words
LRContextWindow	The preceding 2 words, and following 2 words
TimeTrend	The overall time-trend score of the whole tweet
ClosestVerb	The closest verb, its time-trend score, distance, left/right-hand side indicator
ClosestTrend	The closest time-trend word, its time-trend score, distance, left/right-hand side indicator

- ***ClosestTrend > ClosestVerb > ContextWindow > LRContextWindow > TimeTrend***

Features	Pr	Re	F_1
Lex+Gra	**0.6895**	**0.5511**	**0.6126**
Lex+Gra - ContextWindow	0.6360	0.5420	0.5852
Lex+Gra - LRContextWindow	0.6520	0.5401	0.5908
Lex+Gra - TimeTrend	0.6736	0.5310	0.5939
Lex+Gra - ClosestVerb	0.6628	0.5237	0.5851
Lex+Gra - ClosestTrend	0.6590	0.5255	0.5848

Impact for POI_f

➢ Effectiveness of BILOU schema features

Features	Pr	Re	F1
Lex+Gra - BILOU	0.6522	0.5201	0.5787

武漢大學 Wuhan University
NANYANG TECHNOLOGICAL UNIVERSITY

Conclusion

- Facilitate the fine-grained location-based services/marketing and personalization
- Crowd wisdom of Foursquare community is exploited
  - Exhaustive coverage for fine-grained locations
- A effective and efficient time-aware POI tagger
  - Enable real-time applications
- Four types of features are extensively investigated
  - Lexical, grammatical and BILOU schema features are all useful