

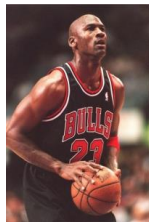
AI6122 Text Data Management & Analysis

Topic: NER on Social Media




NER and EL

- Named-entity recognition (NER)
 - The task to locate and classify named entities in text into pre-defined categories
 - **names** of persons, organizations, **locations**,
 - expressions of **times**, quantities, monetary values, percentages, **etc.**
 - Example: [Jim]_{Person} bought 300 shares of [Acme Corp.]_{Organization} in [2006]_{Time}.
- Entity linking (EL)
 - The task of determining the **identity** of entities mentioned in text, with reference to a knowledge base.
 - Example: Michael Jordan will give a talk at the conference



NER from Social Media Text

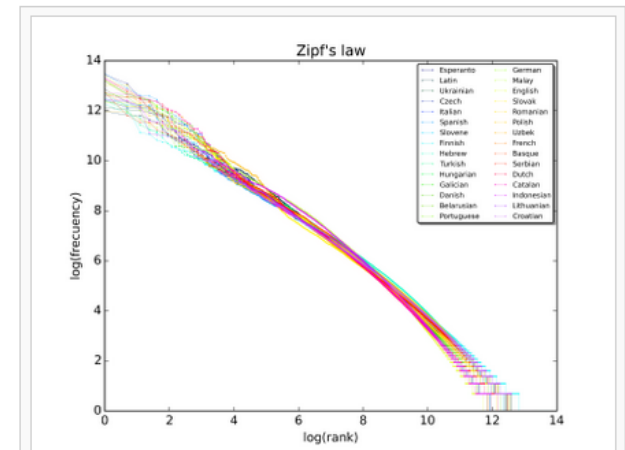
- Why is the task difficult?
 - Informal language
 - Misspellings
 - Grammatical errors
 - Self-defined abbreviations
 - And many others....
 - The next “why”
 - Why does most social media text has these issues?
 - Facebook posts, Tweets, Weibo, WeChat moments, comments, forum posts...
 - The same group of users write research papers, theses, formal documents, homepages
- 




3/31/2020

Informal text serves the purpose

- Principle of least effort
 - “...neither speakers nor hearers using a given language want to work **any harder than necessary** to reach understanding...”
 - https://en.wikipedia.org/wiki/Zipf's_law
- A piece of text for “communicating information”
 - Assuming the right (domain-specific) knowledge
 - Assuming the right context
 - Trending/hot event, Group gathering
 - Parents of students from the same school
 - Students from same research group



A plot of the rank versus frequency  for the first 10 million words in 30 Wikipedias (dumps from October 2015) in a **log-log** scale.

Assuming the right context

Content Without Context is Meaningless

Ramesh Jain
Dept of Computer Science
University of California, Irvine
jain@ics.uci.edu

Pinaki Sinha
Dept of Computer Science
University of California, Irvine
psinha@ics.uci.edu

ABSTRACT

We revisit one of the most fundamental problems in multimedia that is receiving enormous attention from researchers without making much progress in solving it: the problem of bridging the semantic gap. Research in this area has focused on developing increasingly rigorous techniques using the content. Researchers consider that *Content is King* and ignore everything else. In this paper, first we will discuss how this infatuation with content continues to be the biggest hurdle in the success of, ironically, content based approaches for multimedia search. Lately, many commercial systems have ignored content in favor of context and demonstrated better success. Given that the mobile phones are the major platform for the next generation of computing, context becomes easily available and more relevant. We show that it is not Content Versus Context; rather it is Content and Context that is required to bridge the semantic gap. In this paper, first we will discuss reasons for our

community and suggest that we approach important problems from a different perspective. We need to liberate ourselves from our current *koopmanduk* (Frog-in-the-Well) mentality, otherwise all our research will only result in making our approaches irrelevant to the mainstream computing community. In this paper, we propose to address one of the most fundamental problems: bridging the semantic gap. Based on research and emerging technology from multiple related areas, we adopt a new out-of-the-box perspective to bring revolutionary changes in the current research paradigm. At the first sight, it may appear to be something that is known, but we will show that despite a lot of lip-service to the use of context, it is mostly ignored. The current situation is exactly like that in the famous story: *The Emperor's New Clothes* [4].

In the last two decades, multimedia computing has evolved to become the dominant main stream, first in computing and now in mobile computing. The Web has clearly become more multimedia

WITHOUT CONTEXT
WORDS AND
ACTIONS HAVE NO
MEANING AT ALL

GREGORY BATESON
PICTUREQUOTES.COM

PICTUREQUOTES

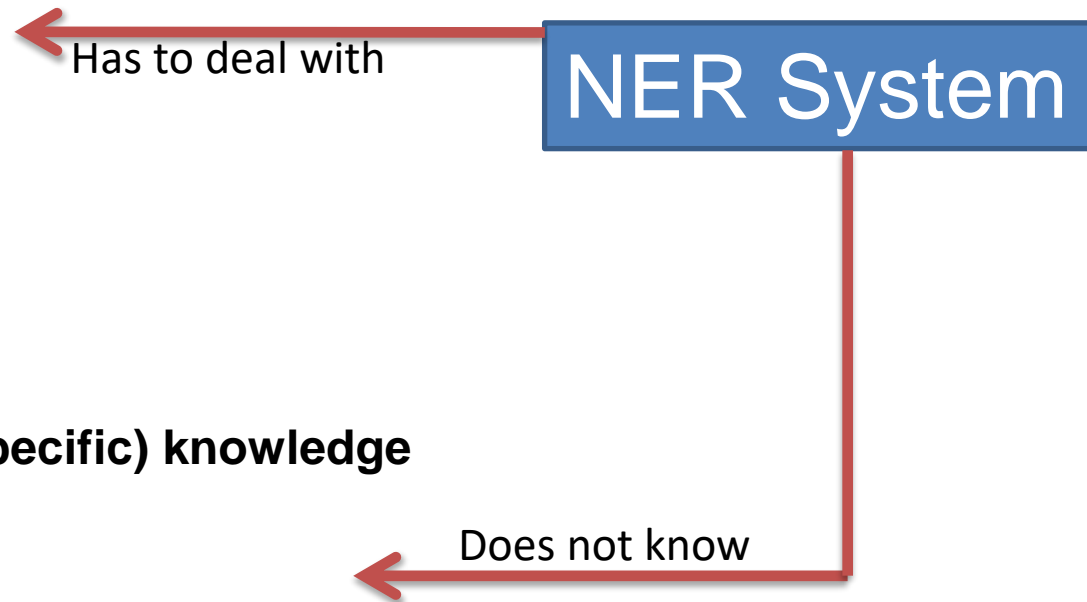
3/31/2020



NANYANG TECHNOLOGICAL UNIVERSITY | SINGAPORE

NER from Social Media Text

- Why is the task difficult?
 - Informal language
 - Misspellings
 - Grammatical errors
 - Self-defined abbreviations
 - And many others....
- Principle of least effort
 - Assuming the right **(domain-specific) knowledge**
 - Assuming the right **context**



Utilize External Resources

- Domain-specific knowledge in User Language
 - Collection of terms used by users to name entities in a specific domain
 - Domain → defines term meanings
- Why not general (open-domain) knowledge bases?
 - Wikipedia, Freebase, ProBase ...
 - What does this term mean: “TCU 2/52”
- Case study:
 - Extract mobile phone names from user forum
 - Extract fine-grained locations from tweets

Mobile phone name extraction and normalization

- True, **Desire** [HTC Desire] might be better if compared to **X10** [Sony Ericsson Xperia X10] but since I am using **HD2** [HTC HD2] , it will be a little boring to use back HTC ...
- I just wanna know what problems do users face on the **OneX** [HTC One X] ... of course I know that knowing the problems on **one x** [HTC One X] doesn't mean knowing the problems on **s3** [Samsung Galaxy SIII]
- oh, the mono rich recording at **920** [Nokia Lumia 920] no better than stereo rich recording at **808** [Nokia 808 PureView]

3/31/2020

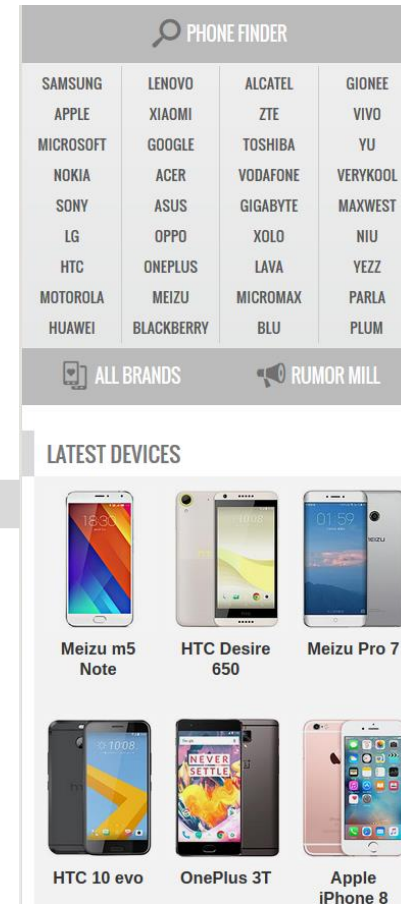
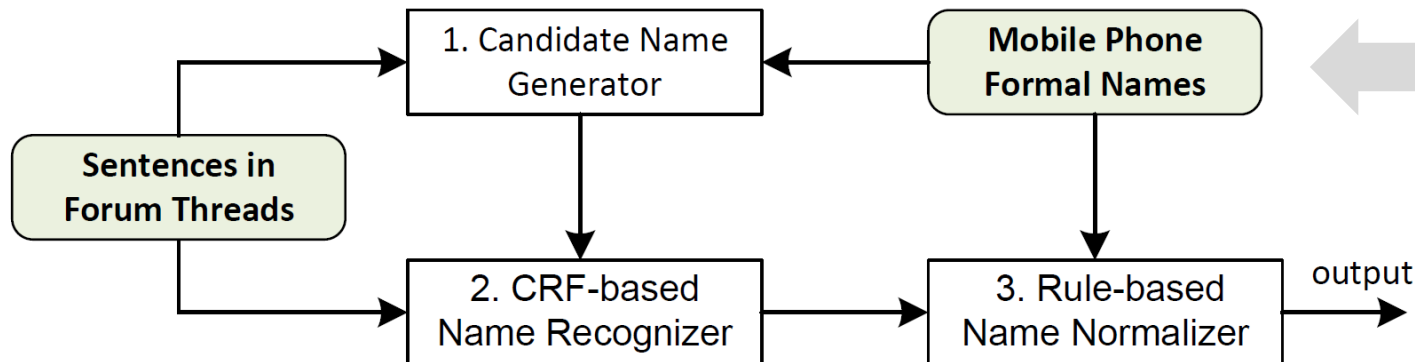


Recognize names based on a dictionary in user language

- Generate **candidate names** based on naming convention

Dictionary defined in
user language

- Recognize** true product names from candidate names
- Normalize** names based on naming convention



3/31/2020



Samsung Galaxy SIII – real data from Singaporean users

- Many variants
- Many users do use formal names
- Brand, series, model
- The usage context shall be similar

– Brown Clustering

Name variation	#users	Name variation	#users
1. galaxy s3	553	14. lte s3	46
2. s3 lte	343	15. galaxy s3 lte	45
3. samsung galaxy s3	284	16. s3 non lte	32
4. s iii	242	17. samsung galaxy siiii	32
5. galaxy s iii	225	18. sgs 3	27
6. samsung s3	219	19. samsung galaxy s3 lte	22
7. sgs3	187	20. sg3	21
8. siiii	149	21. gsiii	16
9. samsung galaxy s iii	145	22. samsung galaxy s3 i9300	15
10. i9300	120	23. samsung i9300 galaxy s iii	13
11. gs3	82	24. s3 4g	11
12. galaxy siiii	61	25. 3g s3	11
13. i9305	52	–	

3/31/2020



Words grouping by Brown Clustering

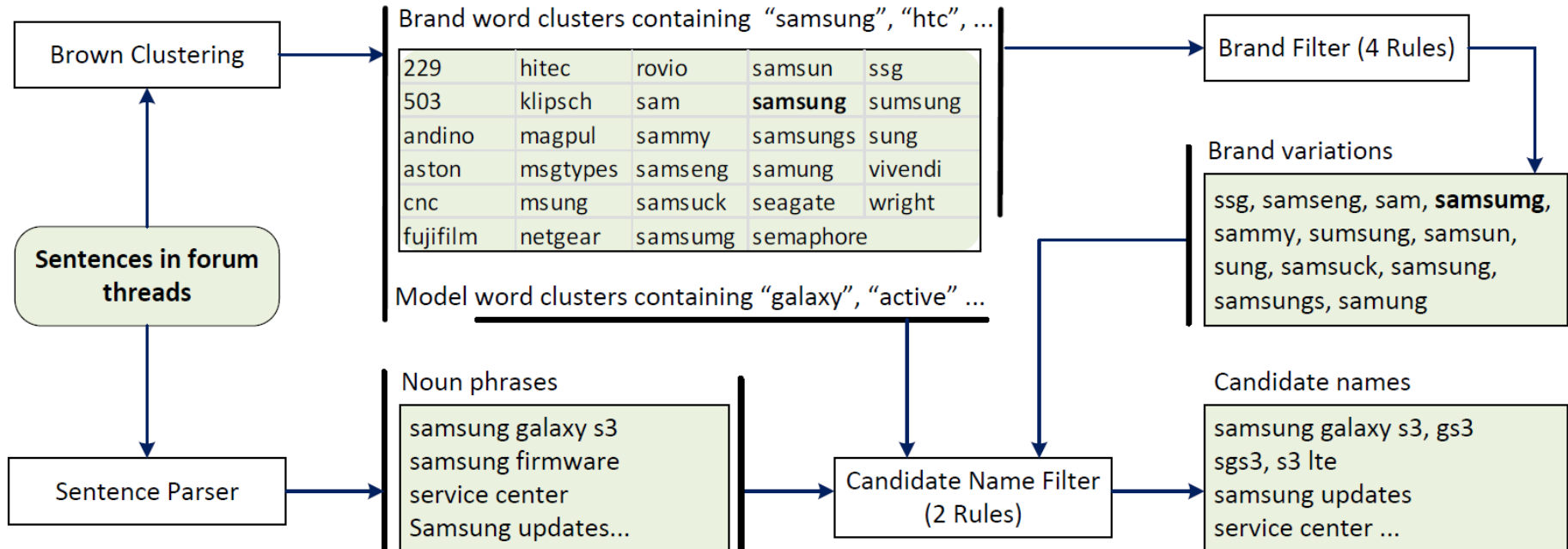
^01110111001 (43)	everything everythin everthing evrything everythang everythingq everythng eveything everythinggg everyting evrythng errthang errything #everything errthing erything everythinggg errythang evrythin erthang 1thing erthing everythingggg jony everythig everytin everything everithing everythign everythn everyhting everythingqq everythink erythang everything- everythingiing everythingggggg errrthang everythg everyword evreything everysong er'thing
^01110111010 (85)	nothing nothin nun nuthin nuttin nuffin 10x noting nthn nowt nuthn nothing 100x nothingg nothn nothingq nutin notin nuffn nutn whatever #nodisrespect nuttn nothinggg #dontmeantobrag 1000x zilch nothinn #nothing nothng nutting nufin nuin nout nothingq nthng nthing nothingggg nufn nofin nothen nthin nottin ntn nought ntg nothinnn nothign n0thing nothig
^011101110110 (108)	something somethin sumthin sumthing sumn somthing sth sumthn sumtin smth somthin suttin sumin somethingq summin treaters somethingg someting sumfin smthn somethn something summat smthng smthing sum'n sumthng smthg somn sumtn smething sometin somethign sum10 soemthing somethinn somethinggg something sumpin somin sumting something/someone somtin somehting somthn someshit sumptin something- smtg thereabouts
^011101110111 (36)	anything anythin nething anythng anythingg anythingq anyting anythang anyth anytin anthing anythinggg anyfin vaart anythn anythign nethin nything anyhting #anything anything- anythg anythingggg nothing- anythink anythig anythingqq somethingggg nethng anyhing woodsen endsmeat anything/anyone aything anythiing enything

Source: http://www.cs.cmu.edu/~ark/TweetNLP/cluster_viewer.html

Dictionary (knowledge) in user language

Brand	User spellings
Apple, HTC, LG	–No brand variations–
Nokia	nokia, nokie, nk
BlackBerry	blackberry, bbry, blackbery, bb, bberry
Samsung	ssg, samseng, sam, samsumg, sammy, sumsung, samsun, sung, samsuck, samsung, samsungs, samung
Sony Ericsson	sony erricson, sony ericsson, sony ericson, sony ericcson, sonyericsson, sony ericssion, sn, sony, sonyeric
Motorola	motorola, moto, motorolla, mot

Dictionary: candidate mobile phone names



Rules derived on naming conventions and usage patterns

- A word cluster W_b contains brand b , then a word w in W_b is a variation of brand b if:
 - The **phonemic edit distance** between w and b is 0, or
 - The first and the last characters in w and b are the same
 - ... rules derived based on naming conventions ...
- A noun phrase is a candidate mobile phone name if it satisfies:
 - The phrase contains a brand variation, or the phrase appears after a brand variation at least once in the whole dataset; and
 - At least one word in the phrase appears in a model word cluster and all the remaining words appear in either model word clusters or brand word clusters.

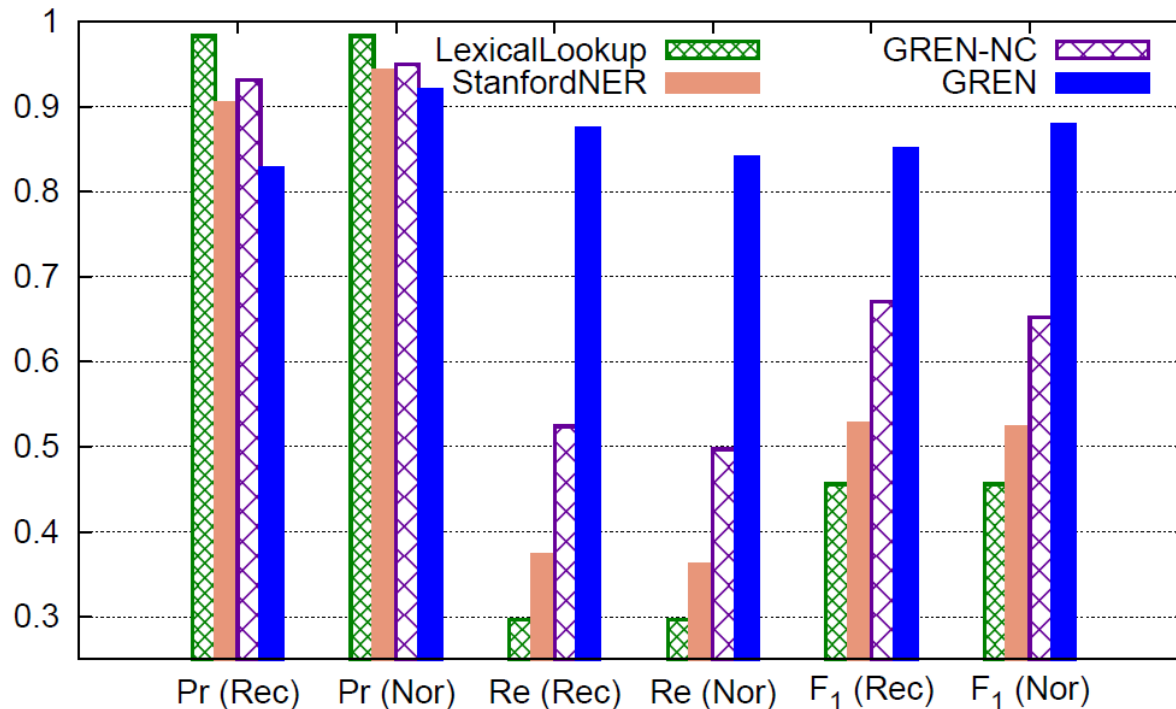
NER from mobile phone forums

- Positive names P :
 - All formal/official names from mobile phone comparison website(s)
 - Formal names by replacing Roman number with Arabic number: SIII \rightarrow S3
 - Model names if containing more than one word e.g., “galaxy note”
- Negative names N :
 - Manual annotation from the set of candidate names, e.g., “service center”, “firmware”, “update”.
- 33,072 sentences selected **automatically**:
 - The sentence contains at least one entity in either set P or set N ;
 - The sentence does not contain any entry appears in unlabeled candidate names
- CRF to classify a candidate name

3/31/2020



Experimental results



Rec: name recognition Nor: name normalization



Dataset

- 1,026,190 posts in 25,251 threads from March 2002 to May 2013



Evaluation

- 4,121 sentences with 946 phone name mentions

A short summary

- What we have learnt from this work
 - Brown clustering is effective in “grouping” product name variants
 - Rule-based approach is useful in product name recognition if there exist naming convention
 - With rule-based approach, training examples can be obtained in semi-automatic manner
- Limitations
 - Candidate name set needs to be updated from time to time
 - Code name cannot be normalized to phone name, e.g., “Nozomi” to “Sony Xperia S”
 - Mobile phone names do follow certain patterns: Brand, Series, Model ...

Utilize External Resources

- Domain-specific Knowledge in **User Language**
 - Collection of terms used by users to name entities in a specific domain
 - Domain defines term meanings
- Why not general (open-domain) knowledge bases?
 - Wikipedia, Freebase, ProBase ...
 - What does this term mean: “TCU 2/52”
 - Hint: a phrase widely used in (Singapore) medical records
- Case study:
 - Extract mobile phone names from user forum
 - **Extract fine-grained locations from tweets**

NER in Social Media Text

- Mobile phones generally follow some naming conventions
 - We derive candidate names or user dictionary from the data itself
- How about named entities in other domains?
 - Your new **mac** is beautiful
 - Enjoying world cup at **mac**
- How about “popular”



Fine-grained locations or POI

- “see you later at **popular** [The Popular Book Store] @ **jp** [Jurong Point]”



Concise English Dictionary

popular ['pɒpjələ(r) / 'pɒpjul-]
adj.

1. regarded with great favor, approval, or affection especially by the general public
2. carried on by or for the people (or citizens) at large
3. representing or appealing to or adapted for the benefit of the people at large
4. (of music or art) new and of general appeal (especially among young people)

➤ Locations may not follow naming conventions

- Hard to derive candidate names from the data itself
- Exploit external resources



Location profile and check-in tweets

"Combat" Top Quality Durian
Dessert Shop, Miscellaneous Shop, and Diner
Novena, Singapore

7.8 / 18 27 ratings

6 Tips and reviews

Filter: durian

Sort: POPULAR RECENT

Awesome durians. I ate one that was creamy, sweet, bitter, and left a slight tingling sensation on the tongue which some folks think it's because it was freshly picked. Free water available.
P.C. - January 16

Save Like

Golden Village
Multiplex, Movie Theater, and General Entertainment
VivoCity (#02-30 & #03-04), 098585, Singapore
At: VivoCity

7.5 / 10 Based on 424 votes
People like this place

Hours: Likely open (See when people check in)
Credit Cards: Yes

Total Visitors: 19316
Total Visits: 40213

Save

http://4sq.com/8HTd5G Share

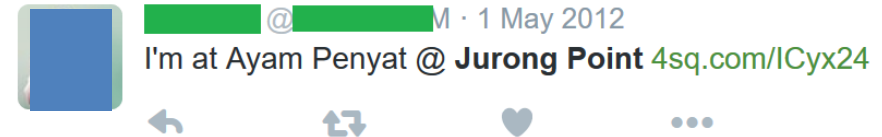
I'm at Ayam Penyat @ Jurong Point 4sq.com/lCyx24

3/31/2020



Location names in user language

- Location profile in Foursquare
 - Formal location names (relatively)
- Location mentions in check-in tweets
 - Location names in user language



3/31/2020

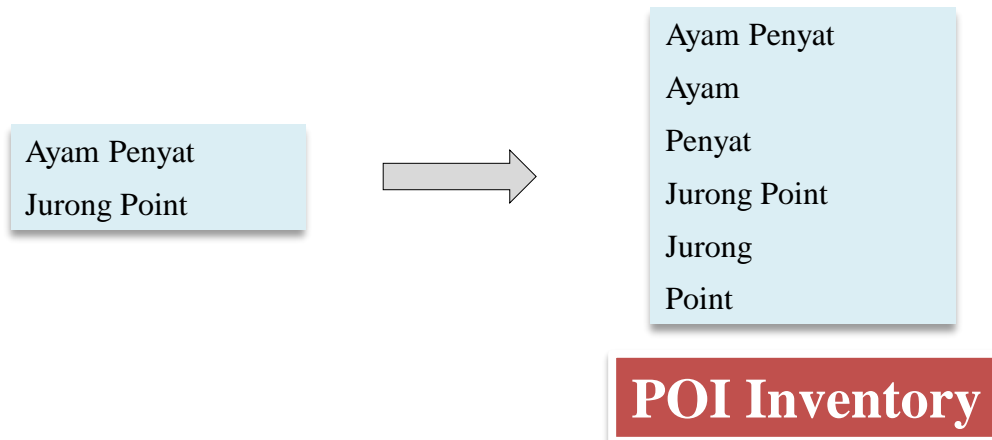


POI inventory: a location dictionary in user language

- Construct POI inventory from check-in tweets and location profiles



- Extend POI inventory with Partial POI names, by taking all the sub-sequences of the names



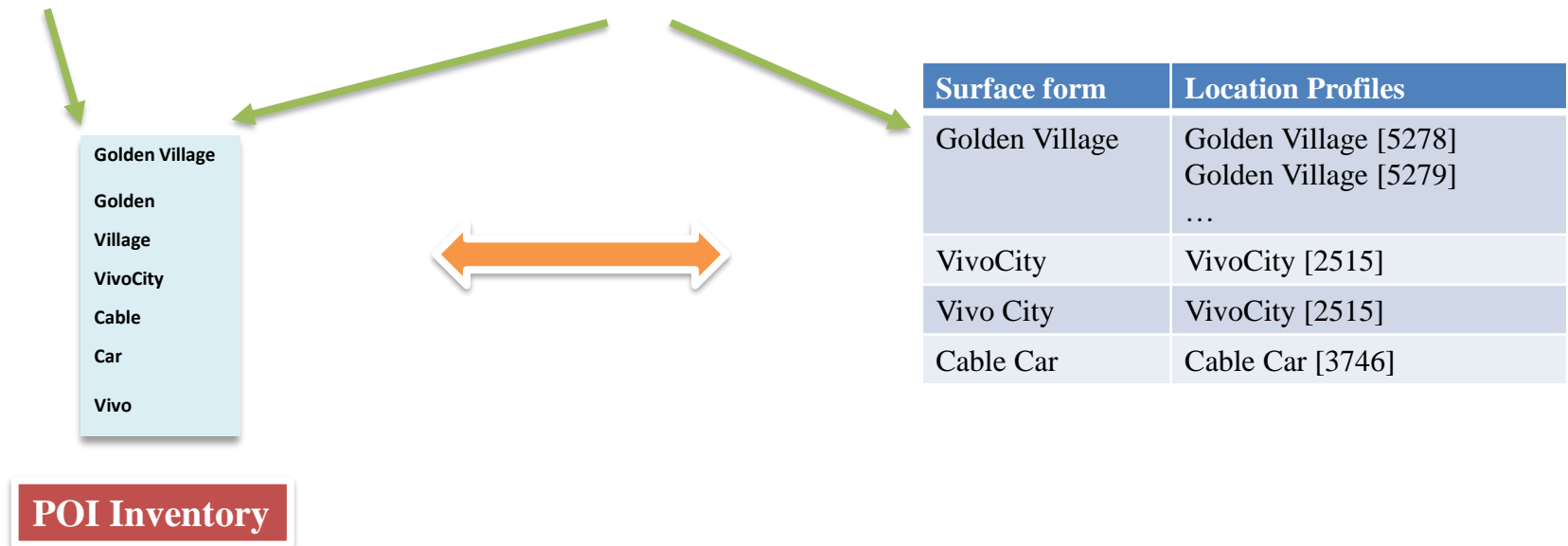
3/31/2020



Candidate linking

Check-in Tweets	Location Profiles
I'm at Golden Village (Singapore) [link]	Golden Village [5278]
I'm at Vivo City (Singapore) [link]	VivoCity [2515]
Ice Age 4 (@ Golden Village) [link]	Golden Village [5279]
Quiet day at work (@ Cable Car) [link]	Cable Car [3746]

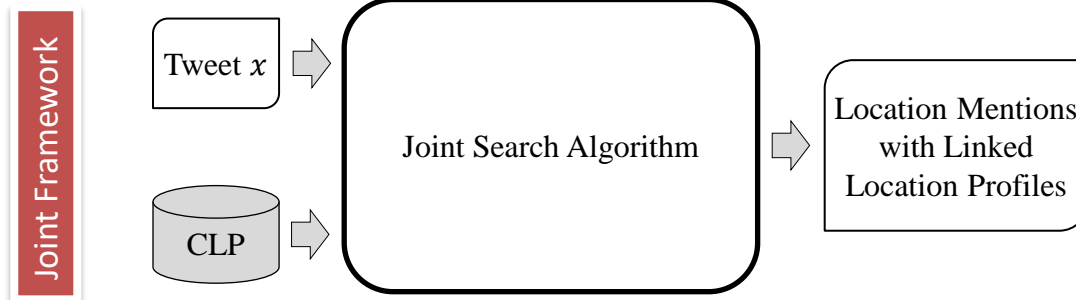
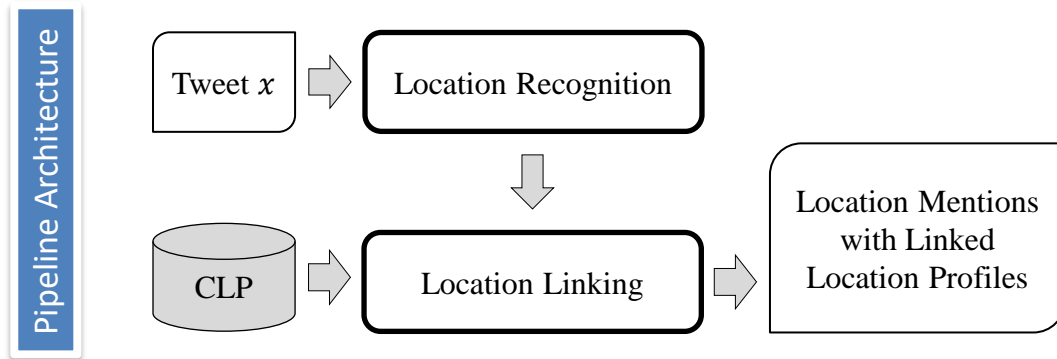
Foursquare as the knowledge base



3/31/2020



Recognition and linking



- CLP
 - Collection of location profiles from Foursquare
- Location Recognition
 - CRF on **pre-labeled** tweets
 -
- Location Linking
 - Local features + geographical coherence
- Joint Search Algorithm
 - Structure prediction

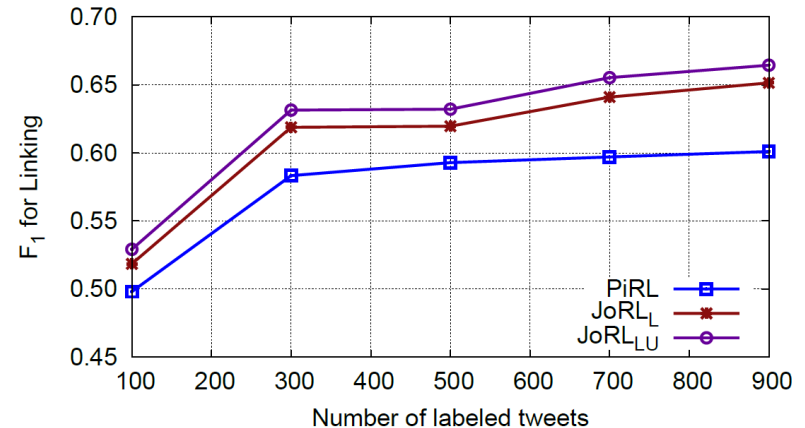
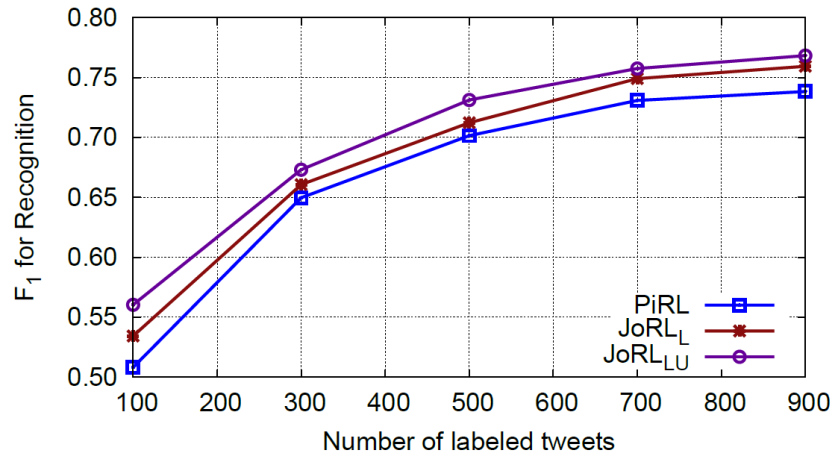
Experiment data

- Data preparation
 - Location profiles: 326,853 check-in tweets → 28,134 location profiles
- Evaluation data
 - Randomly sampled 100 Singapore Twitter users, collect at most 2000 tweets per sampled user
 - 100,058 tweets after cleaning
 - 24,129 tweets contain at least one match in POI inventory (e.g., “popular”, “jurong”, “point”, “jp”)
 - Manually label 4,012 tweets
 - 169 filtered out for containing mostly non-English words
 - 232 tweets containing at least one location Unknown
 - 3,611 labeled tweets for evaluation
- 900 for training, 211 for development, 2500 for testing

3/31/2020



Experimental results

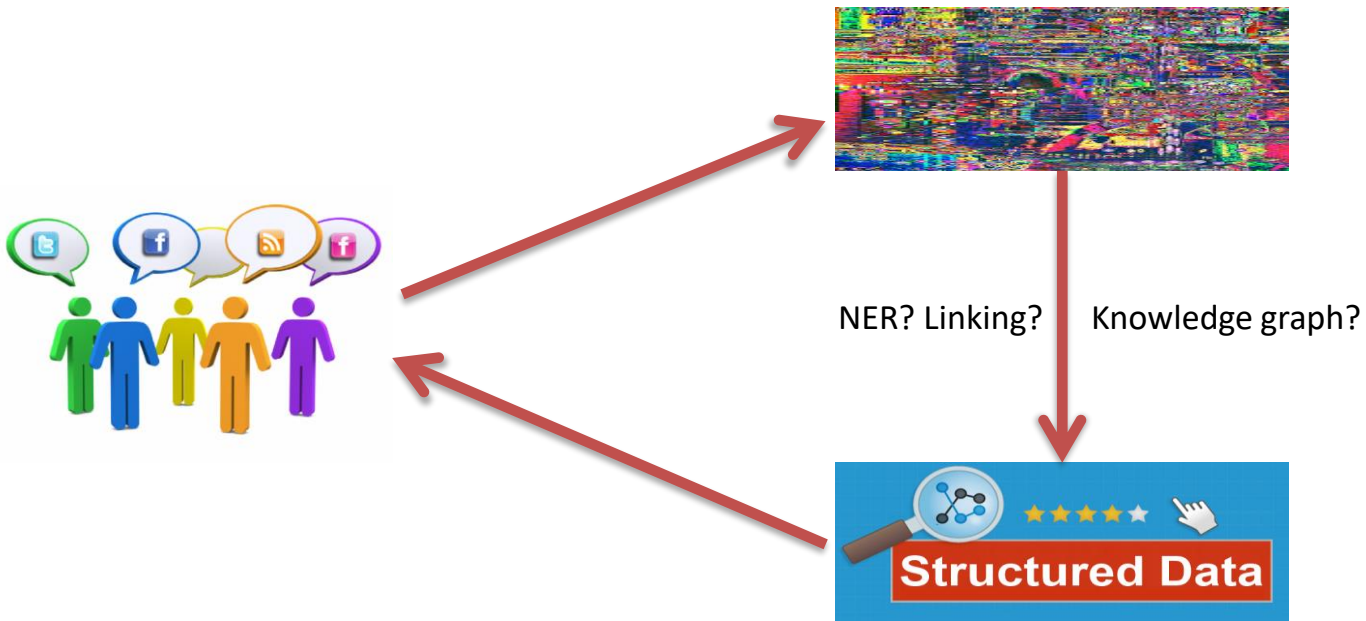


- PiPL: the pipeline architecture
- JoRL_L: the joint framework learning from labeled data
- JoRL_{LU}: JoRL learning from labeled and unlabeled data
 - Minimize the error on the labeled data, and
 - Maximize the agreement on the unlabeled data from multiple models

Another short summary

- Better results are achieved through
 - Make good use of dictionary
 - POI inventory ← The (very noisy) dictionary in user language
 - Examples: point, popular, mac, home ...
 - Joint considering POI recognition and linking
 - Relatively more context from the tweets
- Remains an extremely challenging problem
 - Informal writing style is not well addressed
 - Lack of context, which may not be able to derive from the tweet itself
 - Limited to locations

Users, Noisy Content, Structured data



3/31/2020

