Named Entity Recognition on Indonesian Microblog Messages

Natanael Taufik, Alfan F. Wicaksono, Mirna Adriani
Information Retrieval Lab.
Faculty of Computer Science, Universitas Indonesia

Outline

- Named Entity Recognition for Entity Names
- Background
- Dataset
- Feature Representation
- Features
- Result
- Conclusion

Named Entity Recognition for Entity Names

Task to identify named entity from text.

Scope:

- Person
- Organization
- Location

Chinchor et al. - 1999 Named Entity Recognition Task Definition

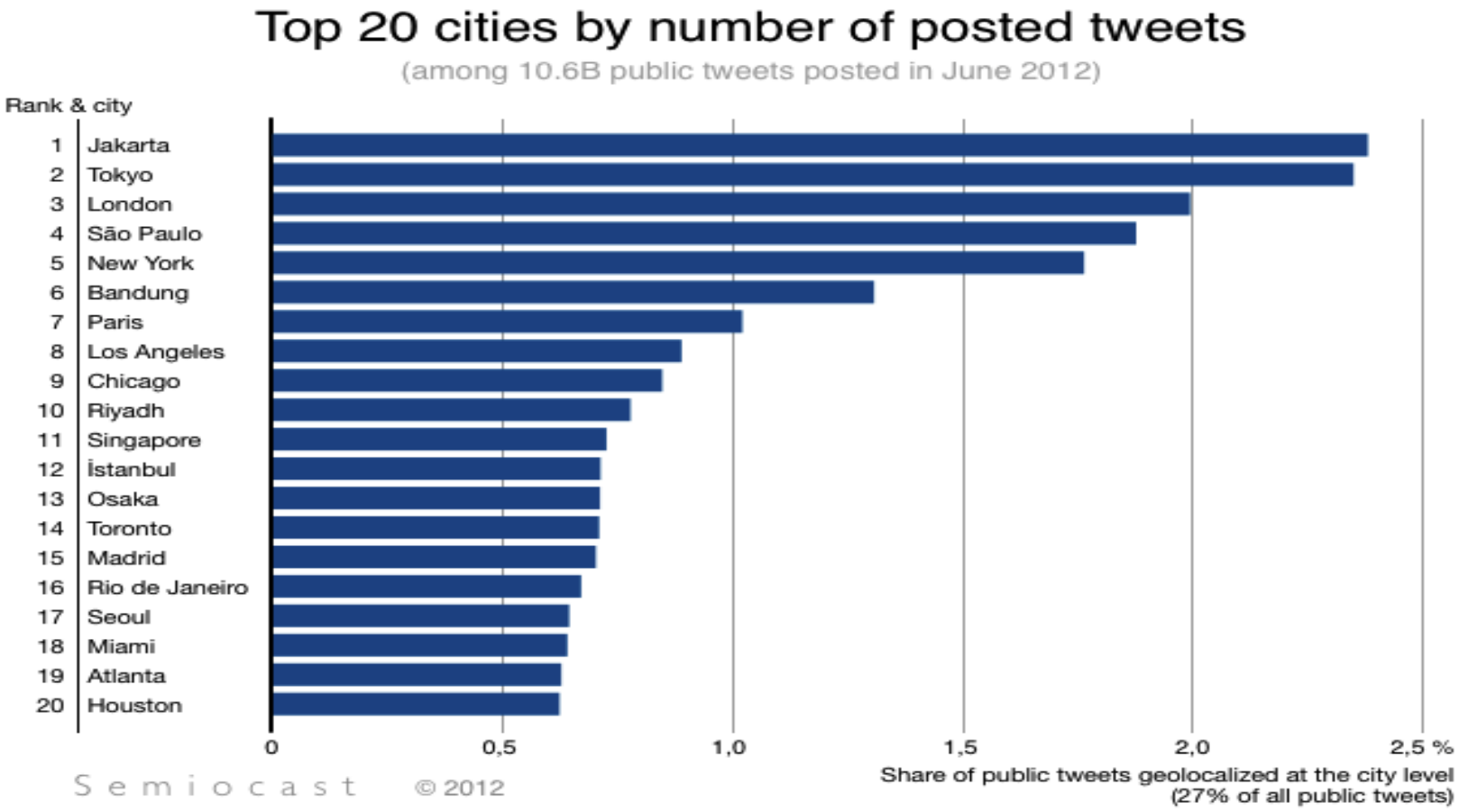
Named Entity Recognition for Entity Names

11/20/2016 - http://www.bbc.com/news/world-us-canada-37925014
11/20/2016 - http://superstore.wnd.com/TIME-TO-PLAY-THE-TRUMP-CARD-Magnetic-or-Adhesive-Bumper-Sticker

Background

11/20/2016 - http://iconcept-seo.com/article/social-media-and-its-effects-to-different-aspects-of-life-part-1-2/
11/20/2016 - http://www.mid-day.com/articles/what-not-to-do-on-the-social-network/15462181

Background


Top 20 cities by number of posted tweets
(among 10.6B public tweets posted in June 2012)
Rank & city
1 Jakarta
2 Tokyo
3 London
4 São Paulo
5 New York
6 Bandung
7 Paris
8 Los Angeles
9 Chicago
10 Riyadh
11 Singapore
12 İstanbul
13 Osaka
14 Toronto
15 Madrid
16 Rio de Janeiro
17 Seoul
18 Miami
19 Atlanta
20 Houston
0
0,5
1,0
1,5
2,0
2,5 %
Share of public tweets geolocalized at the city level
(27% of all public tweets)
Semiocast ©2012

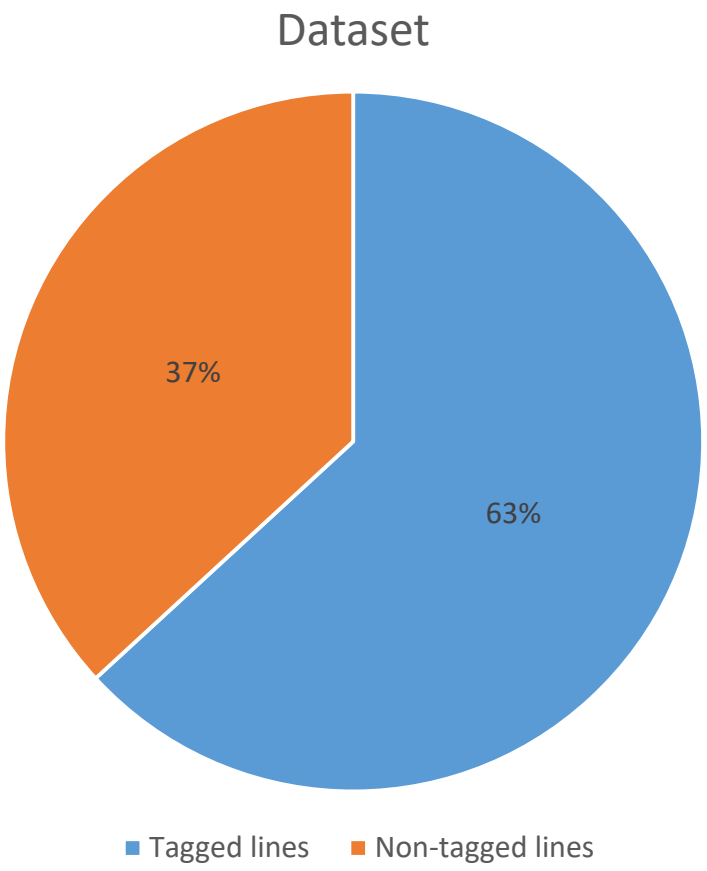
NER for Indonesian Microblog

404 - Not Found

Dataset

	# count
Tagged lines	379
Non-Tagged lines	221
Total	**600**

Type	# count
Person	225
Location	257
Organization	204



Feature Representation

Feature 1	Feature 2	Feature 3	Label	
Feature 1	Feature 4	Label		
Feature 1	Feature 5	Label		
Feature 1	Label			
Empty line				
Feature 1	Feature 3	Label		
Feature 1	Feature 2	Label		
…	…	…		

Feature Representation(example)

Jokowi Diminta Perhatikan Angkutan Umum

means

Jokowi Asked To Pay Attention To Public Transportation

Feature Representation(example)

Source	Features	Label
Jokowi	capitalized	Person
Asked	capitalized	Other
To	capitalized	Other
Pay	capitalized	Other
Attention	capitalized	Other
To	capitalized	Other
Public	capitalized	Other
Transportation	capitalized	Other

Feature Representation(example)

<B_ENAMEX TYPE="PERSON">Jokowi<E_ENAMEX> Asked To Pay Attention To Public Transportation

Features

- Word

Jokowi, written as is: “**jokowi**”

- Last 3 letters

Example of family names: *Setia**wan***, *Hendra**wan***, *Hima**wan***

- Word length

Written as follow: “**wordLength:6**”

Features

- Pattern function[1]:

[A-Z] => “A”

[a-z] => “a”

[0-9] => “0”

other character => “-”

Jokowi => Aaaaaa

=> Aa

Result: [“Aaaaaa”, “Aa”]

[1] M. Collins - Ranking algorithms for named-entity extraction: Boosting and the voted perceptron

Features

- Inside bracket

Example:

You're like a moon, can be seen, but can't be owned(Ari, Dewi)

Feature: "**insideBracket**"

Features

- Part-of-speech

Train Stanford Log-linear Part-Of-Speech Tagger using tagged tweets made by Canggadibrata and Bressan

Example:

"Susilo Bambang Yudhoyono"

f("Yudhoyono") = ["1stLeftPOS-NNP", "2ndLeftPOS-NNP"]

- Surrounding words

f("Yudhoyono") = ["1stLeftWord-Bambang", "2ndLeftWord-Susilo"]

Features

- Lookup list – common location and stopwords[1]

f(“Jawa”) = [“isRegion”]

f(“dan”) = [“isStopword”]

- Non-standard word list[1]

f(“nggak”) = [“tidak”]

[1] Vania, C., Ibrahim, M., dan Adriani, M. - Sentiment lexicon generation for under-resourced language

Result

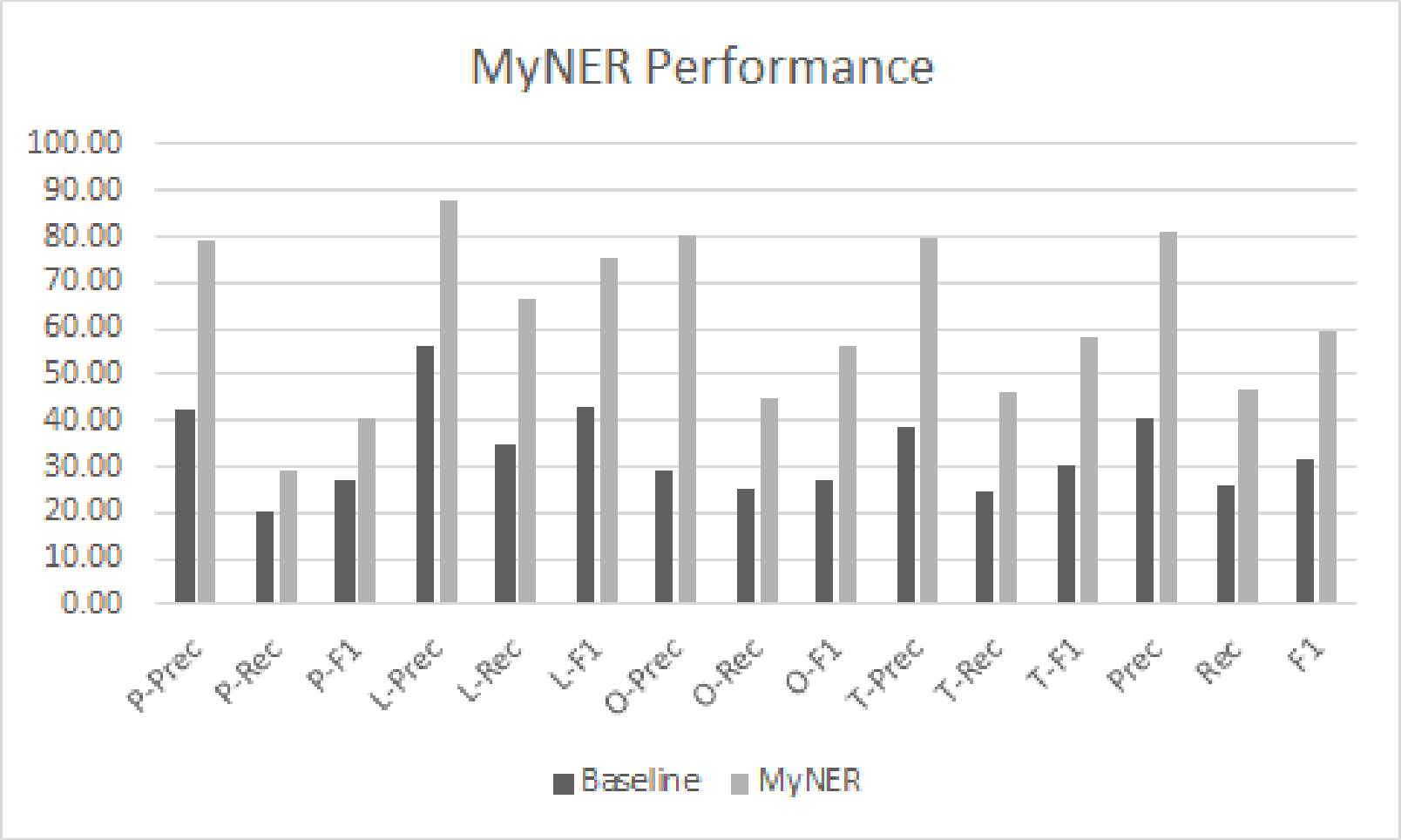
Baseline

Research by G. Wahyudi, using rule based approach based on contextual and morphological information as well as part-of-speech of the words.

Result

	Type	Precision(%)	Recall(%)	F1(%)
Baseline	PER	42.45	20.00	27.19
	LOC	56.33	34.63	42.89
	ORG	29.38	25.49	27.30
MyNER (10-fold cross validation)	PER	79.02	29.16	40.75
	LOC	88.04	66.16	75.13
	ORG	80.27	44.63	56.35

Result


MyNER Performance
100.00
90.00
80.00
70.00
60.00
50.00
40.00
30.00
20.00
10.00
0.00
P-Prec
P-Rec
P-F1
L-Prec
L-Rec
L-F1
O-Prec
O-Rec
O-F1
T-Prec
T-Rec
T-F1
Prec
Rec
F1
Baseline
MyNER

Conclusion

- NER on Indonesian microblog is challenging because it’s sort and written in a non-standard way
- Specific model built for Indonesian microblog messages outperforms model built for formal Indonesian language
- Now we have a NER for Indonesian microblog messages
- Further improvements are needed since the best results do not seem sufficient for higher level applications
- Our model is based on sequence labeling task that employs Conditional Random Fields as our machine learning algorithm

Terima kasih ☺
Thank you ☺
謝謝 ☺
ありがとうございました ☺

Any question?