MINI-HACKATHON NLP-NameEntityRecognition

SuperAI2-513
Airin Intaratat - Gaem

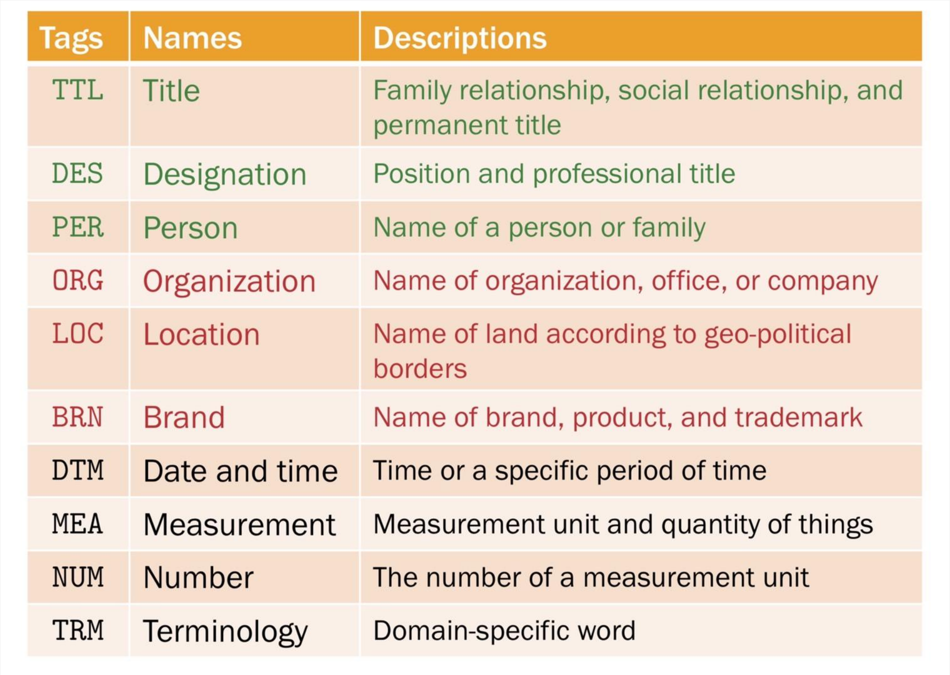
CONTENTS

- **Task**
- **Data Exploration**
- **Literature Review**
- **Experiment**
  1. *BiLSTM - Thai2fit Embedding*
  2. *BiLSTM - Bert Embedding*
  3. *BiLSTM - CRF - Thai2fit Embedding*
- **Result and Evaluation**
- **Future Improvement**

CONTENTS

- **Task**
- **Data Exploration**
- **Literature Review**
- **Experiment**
  1. ***BiLSTM - Thai2fit Embedding***
  2. ***BiLSTM - Bert Embedding***
  3. ***BiLSTM - CRF - Thai2fit Embedding***
- **Result and Evaluation**
- **Future Improvement**

Our Task



Tags	Names	Descriptions
TTL	Title	Family relationship, social relationship, and permanent title
DES	Designation	Position and professional title
PER	Person	Name of a person or family
ORG	Organization	Name of organization, office, or company
LOC	Location	Name of land according to geo-political borders
BRN	Brand	Name of brand, product, and trademark
DTM	Date and time	Time or a specific period of time
MEA	Measurement	Measurement unit and quantity of things
NUM	Number	The number of a measurement unit
TRM	Terminology	Domain-specific word

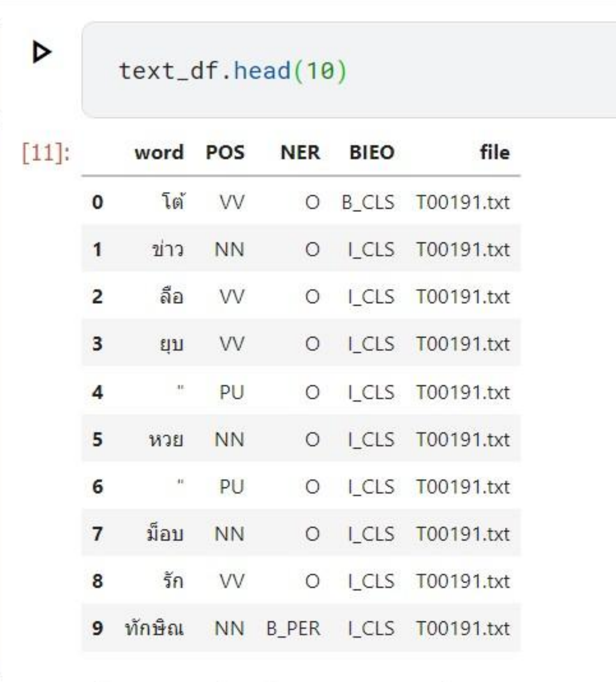
Ex.1 นายกรัฐมนตรี/B_DES|ดร./B_TTL |
มหาธีร์/B_PER | บิน/I_PER |
โมฮัมหมัด/E_PER |

Ex.2 ที่/O | โรงแรม/B_LOC |อินโดจีน
/E_LOC | ␣ |อำเภอ/B_LOC |
อรัญประเทศ/E_LOC | ␣ |
จังหวัด/B_LOC |สระแก้ว/E_LOC |

CONTENTS

- **Task**
- **Data Exploration**
- **Literature Review**
- **Experiment**
  1. *BiLSTM - Thai2fit Embedding*
  2. *BiLSTM - Bert Embedding*
  3. *BiLSTM - CRF - Thai2fit Embedding*
- **Result and Evaluation**
- **Future Improvement**

Data Exploration

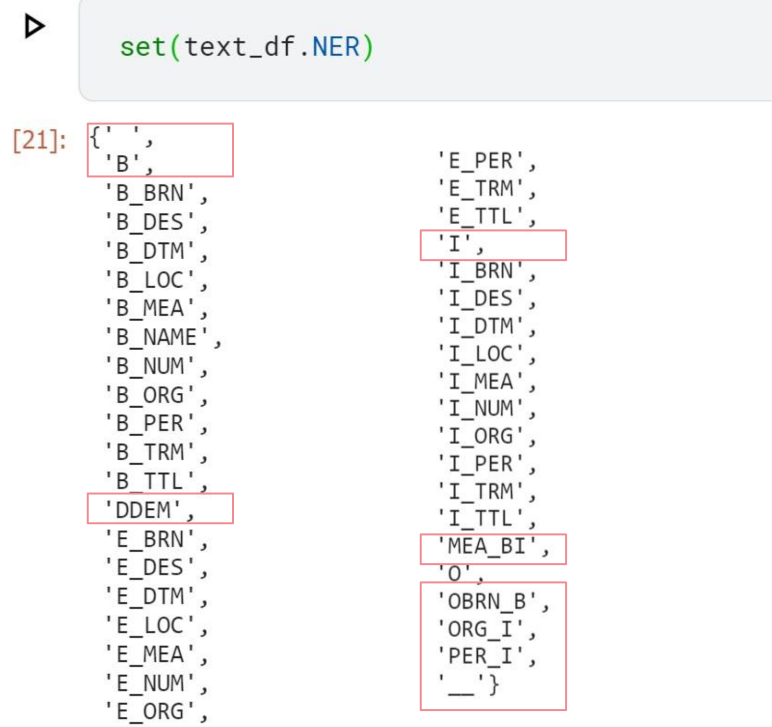


```
text_df.head(10)
```

[11]:

	word	POS	NER	BIEO	file
0	โต้	VV	O	B_CLS	T00191.txt
1	ข่าว	NN	O	I_CLS	T00191.txt
2	ลือ	VV	O	I_CLS	T00191.txt
3	ยุบ	VV	O	I_CLS	T00191.txt
4	"	PU	O	I_CLS	T00191.txt
5	หวย	NN	O	I_CLS	T00191.txt
6	"	PU	O	I_CLS	T00191.txt
7	ม็อบ	NN	O	I_CLS	T00191.txt
8	รัก	VV	O	I_CLS	T00191.txt
9	ทักษิณ	NN	B_PER	I_CLS	T00191.txt

- DataFrame from LST20

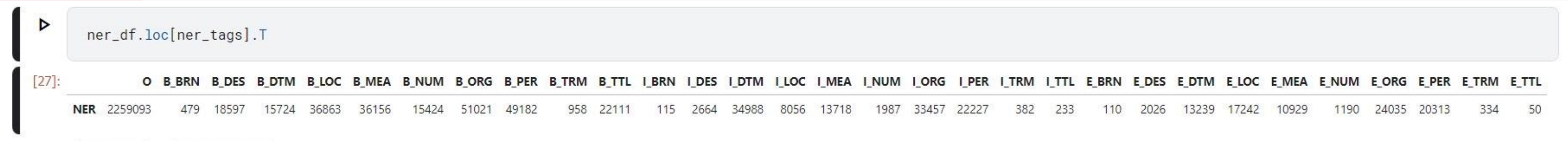


```
set(text_df.NER)
```

```
[21]: {' ',
        'B',                'E_PER',
        'B_BRN',            'E_TRM',
        'B_DES',            'E_TTL',
        'B_DTM',            'I',
        'B_LOC',            'I_BRN',
        'B_MEA',            'I_DES',
        'B_NAME',           'I_DTM',
        'B_NUM',            'I_LOC',
        'B_ORG',            'I_MEA',
        'B_PER',            'I_NUM',
        'B_TRM',            'I_ORG',
        'B_TTL',            'I_PER',
        'DDEM',             'I_TRM',
        'E_BRN',            'I_TTL',
        'E_DES',            'MEA_BI',
        'E_DTM',            'O',
        'E_LOC',            'OBRN_B',
        'E_MEA',            'ORG_I',
        'E_NUM',            'PER_I',
        'E_ORG',            '__'}
```

- NE column - found abnormal data
Solution: Drop all sentences that have unusual NE

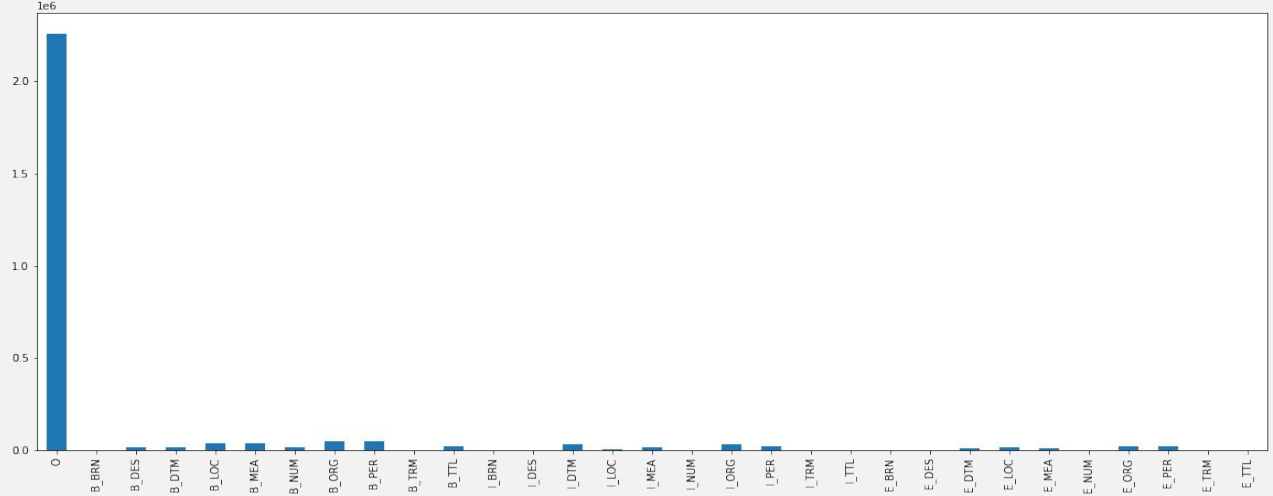
Data Exploration



```
ner_df.loc[ner_tags].T
```

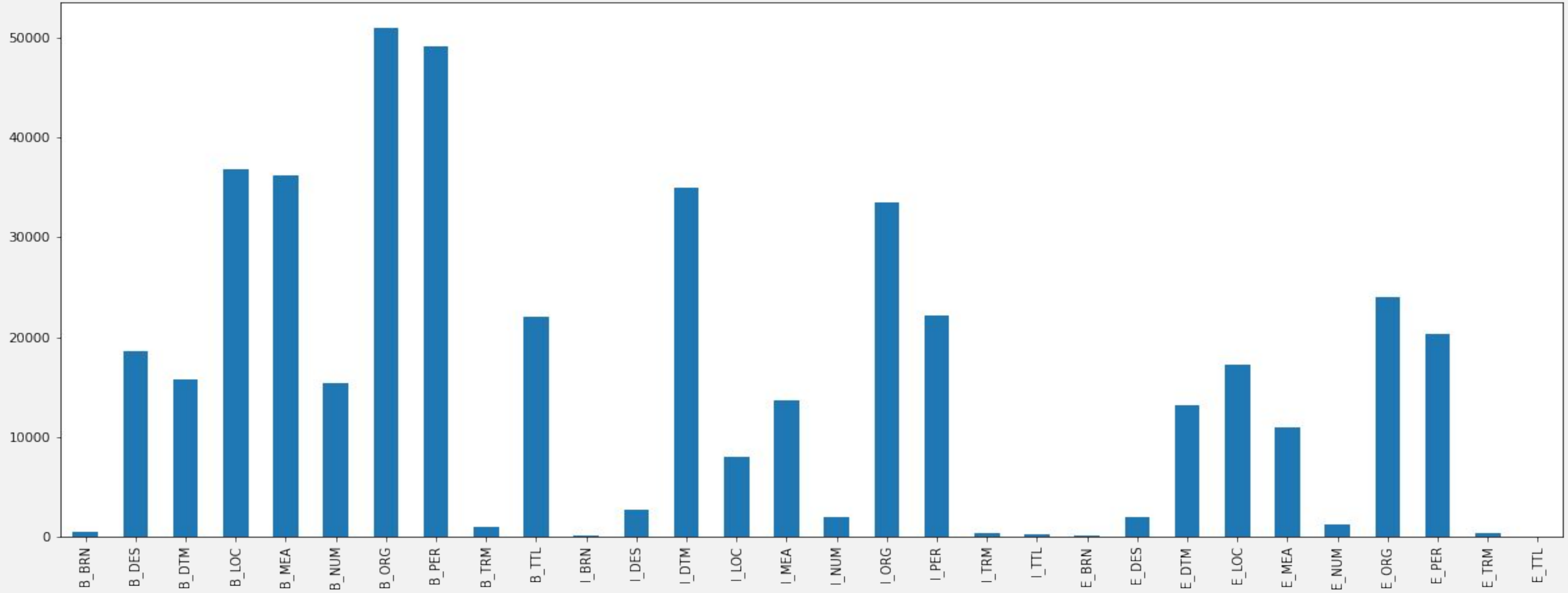
[27]:

	O	B_BRN	B_DES	B_DTM	B_LOC	B_MEA	B_NUM	B_ORG	B_PER	B_TRM	B_TTL	I_BRN	I_DES	I_DTM	I_LOC	I_MEA	I_NUM	I_ORG	I_PER	I_TRM	I_TTL	E_BRN	E_DES	E_DTM	E_LOC	E_MEA	E_NUM	E_ORG	E_PER	E_TRM	E_TTL
NER	2259093	479	18597	15724	36863	36156	15424	51021	49182	958	22111	115	2664	34988	8056	13718	1987	33457	22227	382	233	110	2026	13239	17242	10929	1190	24035	20313	334	50



- Plot bar chart - O class much more than others

Data Exploration



- Bar chart - without O class

Literature review

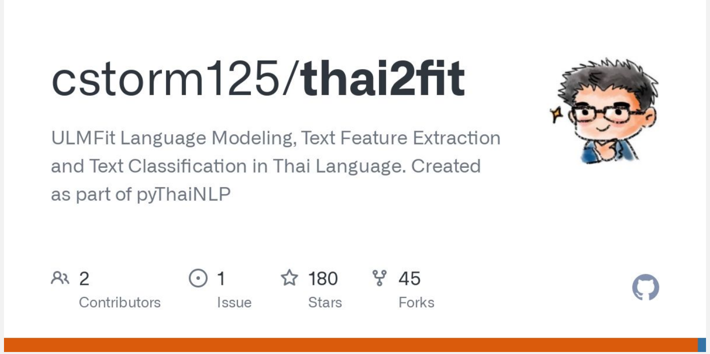
- **How to represent text?**
- **What model should we use?**

Literature review

- **How to represent text?**
    a. **Word2vec - skip gram, CBOW**
    b. **GLOVE**
    c. **ULMFiT**
    d. **ElMo**
    e. **BERT**

Literature review

- **How to represent text?**
  a. **Word2vec - skip gram, CBOW**
  b. **GLOVE**
  c. **ULMFiT - Thai2fit (pythainlp)**
  d. **ElMo**
  e. **BERT - Geotrend/bert-base-th-cased (huggingface transformers)**



cstorm125/**thai2fit**

ULMFit Language Modeling, Text Feature Extraction and Text Classification in Thai Language. Created as part of pyThaiNLP

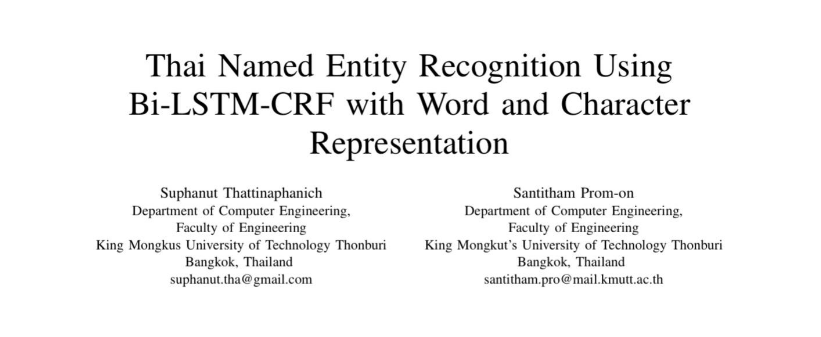
2 Contributors · 1 Issue · 180 Stars · 45 Forks

Literature review

- **What model should we use?**
  a. **LSTM**
  b. **BiLSTM**
  c. **CRF**
  d. **BiLSTM-CRF**

Literature review

- **What model should we use?**
  a. LSTM
  b. **BiLSTM**
  c. CRF
  d. **BiLSTM-CRF**



Thai Named Entity Recognition Using Bi-LSTM-CRF with Word and Character Representation

Suphanut Thattinaphanich
Department of Computer Engineering,
Faculty of Engineering
King Mongkus University of Technology Thonburi
Bangkok, Thailand
suphanut.tha@gmail.com

Santitham Prom-on
Department of Computer Engineering,
Faculty of Engineering
King Mongkut's University of Technology Thonburi
Bangkok, Thailand
santitham.pro@mail.kmutt.ac.th

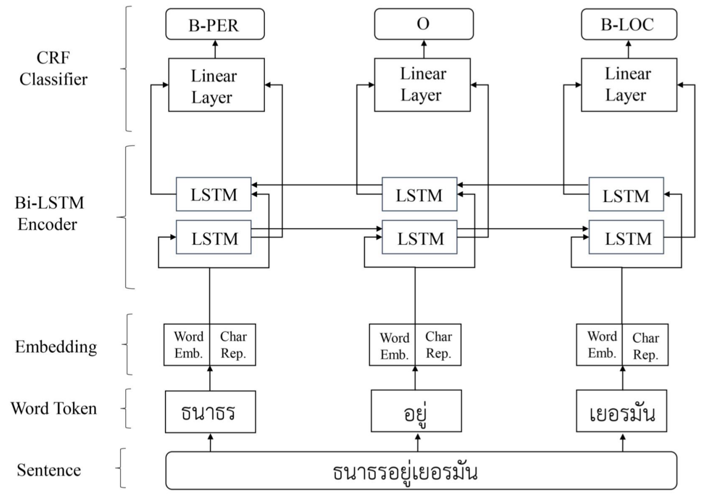


Fig. 2. Bi-LSTM CRF with Word/Character representation Architecture

https://www.researchgate.net/project/Thai-Named-Entity-Recognition-UsingBi-LSTM-CRF-with-Word-and-CharacterRepresentation

CONTENTS

- Task
- Data Exploration
- Literature Review
- **Experiment**
  1. ***BiLSTM - Thai2fit Embedding***
  2. ***BiLSTM - Bert Embedding***
  3. ***BiLSTM - CRF - Thai2fit Embedding***
- Result and Evaluation
- Future Improvement

Experiment

- **BiLSTM -Thai2fit**
  Input → Embedding(Thai2fit) → BiLSTM → Dense-softmax

- **BiLSTM -BERT**
  Input + Attention mask → Embedding(Bert) → BiLSTM → Dense-softmax

- **BiLSTM-CRF -Thai2fit**
  Input → Embedding(Thai2fit) → BiLSTM → CRF

- **Training 15 epochs**
- **Adam Optimizer (learning rate 0.01)**
- **Categorical Cross Entropy Loss**

CONTENTS

- **Task**
- **Data Exploration**
- **Literature Review**
- **Experiment**
  1. *BiLSTM - Thai2fit Embedding*
  2. *BiLSTM - Bert Embedding*
  3. *BiLSTM - CRF - Thai2fit Embedding*
- **Result and Evaluation**
- **Future Improvement**

Results and Evaluation

F1 Accuracy from Kaggle

- BiLSTM	**: 95.5966%**
- BiLSTM-CRF	**: 95.5902%**

01

BiLSTM
Thai2fit

Evaluation F1
(Micro F1)
88.32

02

BiLSTM
BERT

Evaluation F1
(Micro F1)
32.44
****something wrong****

03

BiLSTM-CRF
Thai2fit

Evaluation F1
(Micro F1)
88.40

Results and Evaluation

03

BiLSTM-CRF
Thai2fit

Evaluation F1
(Micro F1)
88.40

```
             precision  recall   f1-score   support

B_BRN          0.5909   0.2889   0.3881      45
B_DES          0.9304   0.9207   0.9255     1815
B_DTM          0.9126   0.8496   0.8800     1942
B_LOC          0.8744   0.8571   0.8657     4214
B_MEA          0.7960   0.9154   0.8516     3074
B_NUM          0.7737   0.6485   0.7056     1286
B_ORG          0.8352   0.8049   0.8198     4496
B_PER          0.9375   0.9336   0.9355     4112
B_TRM          0.6159   0.5776   0.5962      161
B_TTL          0.9821   0.9806   0.9813     2012
E_BRN          0.3333   0.3333   0.3333       6
E_DES          0.9034   0.8600   0.8811      250
E_DTM          0.9044   0.8644   0.8839     1718
E_LOC          0.8684   0.9063   0.8869     2038
E_MEA          0.8422   0.7711   0.8051      817
E_NUM          0.8411   0.8491   0.8451      106
E_ORG          0.8438   0.8374   0.8406     2361
E_PER          0.9435   0.9803   0.9615     1824
E_TRM          1.0000   0.3333   0.5000      21
E_TTL          0.9870   0.9383   0.9620      81
I_BRN          0.3333   0.4000   0.3636       5
I_DES          0.6734   0.7422   0.7061      225
I_DTM          0.9547   0.9331   0.9437     5394
I_LOC          0.8756   0.8162   0.8448     1708
I_MEA          0.8717   0.8049   0.8370     1030
I_NUM          0.8608   0.9653   0.9101      173
I_ORG          0.8709   0.8797   0.8753     3681
I_PER          0.9281   0.9883   0.9573     2052
I_TRM          1.0000   0.3056   0.4681      36
I_TTL          0.9231   0.9449   0.9339      127

micro avg      0.8874   0.8807   0.8840    46810
```

CONTENTS

- **Task**
- **Data Exploration**
- **Literature Review**
- **Experiment**
    1. *BiLSTM - Thai2fit Embedding*
    2. *BiLSTM - Bert Embedding*
    3. *BiLSTM - CRF - Thai2fit Embedding*
- **Result and Evaluation**
- **Future Improvement**

Future Improvement

- Upsampling Data
- Character Embedding
- Error Analysis to improve accuracy
- Ensemble Method

THANKS for Listening

SuperAI2-513
Airin Intaratat - Gaem