



Using machine learning to validate search query in GeoSearch

Chonlatit Yaovalap, Thanabordee Jumsai na ayudhya, Fon Vongvorachate and Sirisup Laohakiat*

Bachelor of Science Program in Computer Science

RELATED KNOWLEDGE

Named-Entity Recognition (NER) [1]

Named-entity recognition is a subtask of information extraction that seeks to locate and classify named entities mentioned in unstructured text into pre-defined categories such as person names, organizations, locations, medical codes, time expressions, quantities, monetary values, percentages, etc.

Recurrent Neural Network (RNN) [2]

Recurrent Neural Network is a generalization of feedforward neural network that has an internal memory. RNN is recurrent in nature as it performs the same function for every input of data while the output of the current input depends on the past one computation. After producing the output, it is copied and sent back into the recurrent network. For making a decision, it considers the current input and the output that it has learned from the previous input, with RNN there are many approach such as Long short-term memory (LSTM), Gated recurrent unit (GRU)

Word Embedding [3]

Word embedding is a dense representation of words in the form of numeric vectors. It can be learned using a variety of language models. The word embedding representation is able to reveal many hidden relationships between words.

Conditional Random Fields (CRF) [4]

Conditional Random Fields are a type of Discriminative classifier, and as such, they model the decision boundary between the different classes. Generative models, on the other hand, model how the data was generated, which after having learnt, can be used to make classifications.

K-Nearest Neighbors Algorithm (KNN) [5]

K-nearest neighbors algorithm uses 'feature similarity' to predict the values of new datapoints which further means that the new data point will be assigned a value based on how closely it matches the points in the training set.

Edit Distance [6]

Edit distance is a measure of the similarity between two strings, which we will refer to as the source string (s) and the target string (t). The distance is the number of deletions, insertions, or substitutions required to transform s into t.

ABSTRACT

Nowadays, location search using computer has become one of the important application in our life. However, if the user fails to input accurate searching phrase, the server cannot provide correct location and navigation for that query. In this study, we aim at developing a machine learning system for processing and correcting location searching query. Our system is developed for searching query in Thai with the target locations in Thailand. We adopt name entity recognition technique to tokenize location name using conditional random field. Then, we use neural network to correct the erroneous phrases by using Bi-LSTM model to perform this task. The experimental results show that the accuracy of our system is as high as 78 percent.

OVERALL OPERATION

Each proposed component of the project was implemented on Python. The performances were shown as following:

1. Named Entity Recognition

We use NER model to recognize the administrative division appear in address searching phrases. The accuracy of the model tested by different datasets are shown as follow:

Dataset	Accuracy of Model
Test Dataset 5000 rows	4213 (84.26%)
CDG Dataset 8069 rows	6676 (82.73%)

2. Word Tokenization Model

With different evaluation metrics, the results are as follow:

2.1 Precision: 99.8%

2.2 Recall: 92.61%

2.3 F1-score: 96.07%

The accuracy of the model tested by different datasets are shown as follow:

Dataset	Accuracy of Model
Test Dataset 5000 rows	4174 (83.48%)
CDG Dataset 200 rows	158 (79%)

3. Word Correcting Model

KNN model was used to correct the search phrase. Testing with two proposed method: edit distance and MaxMatchWord algorithms, the results are summarized as the table below.

Method	Result		
	Wrong Word	Number of Correction Word	Percentage
Edit distance	22,144	14,278	64.48
MaxMatchWord()	22,144	13,513	61.02

4. Bi-Directional Long Short Term Memory-Conditional Random Fields (Bi-LSTM-CRF)

The performance test of Bi-LSTM-CRF model was obtained with the following results.

	precision	recall	f1-score	support
ALLEY	0.64	0.54	0.59	7503
H_NAME	0.21	0.69	0.32	1271
PAD	1.00	1.00	1.00	155257
ROAD	0.81	0.71	0.76	5978
dzp	1.00	0.96	0.98	29991
accuracy			0.97	200000
macro avg	0.73	0.78	0.73	200000
weighted avg	0.98	0.97	0.97	200000

OPERATION PROCESS

1. Prepare data for training NER with CRF model
2. Prepare data for training RNN GRU model
3. Prepare data for training KNN model
4. Prepare data for training Bi-LSTM CRF model
5. Create function to fill word in missing section

ส่ง : นายชลทิศย์ เขียวลาภ ที่อยู่ : เลขที่ 465/114 ชั้น 18 ห้อง 184 คอนโดไอดีโอ เวิร์ฟ ราชปรารก ถนนราดปรารก มักกะสัน เขตราชเทวี กรุงเทพฯ โทร : 095-5971114 ส่งลงทะเบียน

Entity Identification (NER CRF)	465/114 ชั้น 18 ห้อง 184 คอนโดไอดีโอ เวิร์ฟ ราชปรารก ถนนราดปรารก มักกะสัน เขตราชเทวี กรุงเทพฯ
Splitting (GRU)	465/114 ชั้น18 ห้อง184 คอนโดไอดีโอเวิร์ฟราชปรารก ถนนราดปรารก มักกะสัน เขตราชเทวี กรุงเทพฯ
Correcting (K-NN)	465/114 คอนโดไอดีโอ เวิร์ฟ ราชปรารก ถนนราดปรารก มักกะสัน เขตราชเทวี กรุงเทพฯ
Tagging (Bi-LSTM-CRF)	465/114 คอนโดไอดีโอ เวิร์ฟ ราชปรารก ถนนราดปรารก แขวงมักกะสัน เขตราชเทวี กรุงเทพฯ
Fill in Part (DIST, ZONE, PROVINCE)	465/114 คอนโดไอดีโอ เวิร์ฟ ราชปรารก ถนนราดปรารก แขวงมักกะสัน เขตราชเทวี กรุงเทพฯ 10400

System Diagram

Example Diagram

REFERENCES

- [1] What is Named Entity Recognition (NER)? [Internet]. [cited 2020 Oct 10]. Available from: <https://www.techslang.com/definition/what-is-named-entity-recognition-ner/>
- [2] What is Recurrent Neural Network (RNN) and Gated Recurrent Unit (GRU) How to Create RNN GRU With Python – NLP ep.9 [Internet]. [cited 2020 Oct 10]. Available from: <https://www.bualabs.com/archives/3103/what-is-rnn-recurrent-neural-network-what-is-gru-gated-recurrent-unit-teach-how-to-build-rnn-gru-with-python-nlp-ep-9/>
- [3] What is Word Embedding and Word2Vec - lukkidd [Internet]. [cited 2020 Jun 11]. Available from: <https://lukkidd.com/word-embedding-and-word2vec-what-is-e60bdf6d78d3>
- [4] Conditional Random Fields for Sequence Prediction [Internet]. [cited 2020 Oct 10]. Available from: http://www.davidsbatista.net/blog/2017/11/16/Conditional_Random_Fields/
- [5] k Nearest Neighbor Classifier (kNN)-Machine Learning Algorithms [Internet]. [cited 2020 Oct 10]. Available from: <https://medium.com/@equiptelligence/k-nearest-neighbor-classifier-knn-machine-learning-algorithms-ed62feb86582>
- [6] Fuzzy string search [Internet]. [cited 2020 Jun 11]. Available from: <http://ntz-develop.blogspot.com/2011/03/fuzzy-string-search.html>

CONCLUSION

We proposed four models for processing the address searching phrase, namely NER model, word tokenization model, word correcting model and word tagging and final correction model. From the experiments, we found that the overall accuracy of the proposed system is as high as 78%.