SHOPEE CODE LEAGUE 2021

Guidelines for Approaching the Address Elements Extraction Problem

Competition Details

This competition was launched on **13 March 2021.**

Please refer to the problem statements here.

Task:

In this competition, you'll work on addresses collected by us to build a model to correctly extract Point of Interest (POI) Names and Street Names from unformatted Indonesia addresses.

Participants are expected to build their own model for this competition, submissions by teams which directly call any third party APIs on the test set will not be taken into consideration.

Approach

There are various ways to solve this problem. But in general, there will be two main parts. First part is to develop a Name Entity Recognition model to recognize the class of each token. The second part is to develop a spelling correction model to correct the labelled token which is not complete.

Preprocessing:

Restoration of Data:

As mentioned in the problem statement, there are cases where POI/street elements are not complete in data. The first step is to restore the raw addresses based on the labelled complete POI/street pair. You can also label the incomplete raw addresses directly and train the model based on it.

Model Training:

Model selection

Models such as BERT are the most recommended as it is already pre trained with large text corpus. Pretrained models using Bahasa Indo tend to give better performance in this task. But if you use a multi-language version, a decent result can also be achieved. If you are trying statistical methods, conditional random fields (CRF) is recommended, as most of the addresses are short and follow certain written patterns. With creative features and post-processing, CRF can achieve a similar score as BERT.

Pretrained Model

With pretrained models, there are two ways to frame this problem and finetune the model based on the specific tasks. First one is to frame it as a tokenization classification problem. Another one is to frame it as an extractive QA problem.

CRF Model

To achieve similar performance as BERT, a great number of features need to be constructed for the CRF model. There are some tricks that can be used to boost accuracy. First is to use 2-Gram or 3-Gram features to consider nearby words information. The second one is to use W2V + PCA + softmax to consider higher dimensional features. You can also use BERT + CRF or LSTM + CRF.

Spelling Correction

Based on raw addresses along with complete POI/street pairs in the training set, you can build a mapping between incomplete components and its restored version. This dictionary can be used to restore the incomplete addresses before/after the NER task. In the case where the same incomplete component can be mapped into different complete pairs, the easiest way is to use the one that has the highest frequency. To have a better performance, 2-Gram, 3-Gram, and Noisy Channel can also be applied here to consider more of the nearby words information.