Address Element Extraction using Embed-Encode-Attend-Predict Named Entity Recognition

This is an NLP study group task and should not be considered a legitimate research paper.

Vincent Michael Sutanto

Abstract

In Indonesia, e-commerce startups are growing rapidly. Their growth cannot be separated from the fact that every day there are many buying and selling transactions. However, the unstructured address format in Indonesia causes a reduced time efficiency in delivering the goods. Therefore, we need a way to extract points of interest and street addresses from raw shipping addresses. This study uses a neural-based Named Entity Recognition model that has an Embed, Encode, Attend, and Predict architecture. The model produces an accuracy rate of 64.96% and 61.57% on the train data and test data respectively. I also present an analysis of the weakness of the model built for the address extraction problem.

1 Introduction

The growth of e-commerce startups is growing rapidly, especially in Indonesia. For example, in terms of Gross Merchandise Value (GMV), Tokopedia and Shopee, the 2 largest startups in Indonesia, recorded a GMV value of US$14 billion and US$14.2 billion respectively in 2021 (Jayani, 2021). This indirectly implies that there has been an extraordinary process of buying and selling transactions in 2021.

However, the large number of transactions also has its own obstacles. One of the problems that e-commerce startups face is the unstructured delivery address in Indonesia. The non-uniform address structure causes congestion in the flow of goods delivery. The computer cannot automatically extract the Point of Interest and the actual address, even though these two items are needed to accurately determine the geographic position of the customer[1]. Therefore, we need a mathematical model that can extract the Point of Interest and address automatically and accurately.

Named Entity Recognition (NER) is one of the tasks in Natural Language Processing science that can be used to extract a word/phrase from a sentence and determine the class of that phrase/sentence. NER works by assigning a label to each token in a sentence, then the NER model will be trained to predict the specified label from a sentence that has never been encountered before. In general, NER detects tokens into pre-trained classes, for example PERSON to denote person, ORG to organization, and many others. Apart from these classes, the NER model can also be modified to detect custom classes as needed (Kleinberg et al., 2017; Popovski et al., 2020).

This study will use the NER technique, specifically the are Embed-Encode-Attend-Predict architecture to identify and extract Points of Interests and street addresses.

The rest of the section of this paper is compiled as follows: Section 2 explains previous related research; Section 3 contains an explanation of the dataset used; Section 4 contains methods and evaluation metrics; Section 5 contains the results of research and discussion about these results; and Section 6 concludes this paper.

2 Background and Related Works

(Kleinberg et al., 2017) used NER to detect verbal deception. This research uses the same technique as mine, namely Embed-Encode-Attend-Predict which is used in the Spacy library. This study also states that on custom domains, Spacy libraries can get better results than Stanford's.

[1] https://www.kaggle.com/c/scl-2021-ds/overview

Because of this research, I was inspired to use the same technique to solve my research problem.

(Popovski et al., 2020) used the NER technique to detect Adverse Drug Effect Classification and Extraction and Identification of professions and occupations in Spanish Tweets. This study applies a Neural-based NER model consisting of stacked embedding, Long-Short Term Memory, and Conditional Random Field. The author claims that by using this architecture, the model is able to produce competitive performance. This study used a technique similar to me. The difference is that there is an additional attention mechanism to the architecture I used.

3 Dataset

This study uses a dataset taken from the 2021 Shopee Code League competition [2]. The dataset is in the form of a full address along with the Point of Interest and address. The data is divided into 300,000 train data and 50,000 evaluation data. A snapshot of the dataset is presented in Table 1.

From table 1 it can be seen that not all addresses have point of interest and street address. Specifically, 178,509 data did not have a point of interest, 70,143 data did not have a street address, and 31,993 data did not have both. Data that has neither of these is removed because it has no impact on the model.

Raw Address	POI	Street
setu siung 119 rt 5 1 13880 cipayung	-	siung
toko dita, kertosono	toko dita	-
cikahuripan sd neg boj 02 klap boj, no 5 16877	sd negeri bojong 02	klap bojong
raya samb gede, 299 toko bb kids	toko bb kids	raya samb gede

Table 1: Snapshot of the Address Dataset.

4 Methods

4.1 Named Entity Recognition

Named Entity Recognition (NER) is a task in Natural Language Processing that aims to label words/phrases from a sentence into predefined classes, such as PERSON, ORGANIZATION, LOCATION, and others. NER is needed because in real cases, for example, Wendy's can be categorized as LOCATION, but it can also be stated as PERSON ownership.

In fact, NER that has been trained in one domain often performs unsatisfactorily when applied to other domains. Therefore, it is necessary to retrain the NER model according to the relevant domain.

This study uses the Spacy library as the NER model (Figure 1). The NER model used is composed of 4 main parts, which are Embed, Encode, Attend, and Predict sequentially (Honnibal, 2016). Embed aims to map each word into a latent space so that words that have similarities have similar encodings. Furthermore, the encode stage works to capture the relation of sequential tokens and issue a context-sensitive matrix as output. The algorithm used at this stage is CNN / LSTM. Then, the attend stage receives the output from the Encode stage and summarizes it according to the Query. This process produces a one-dimensional output called global-problem specific representation. Finally, a simple Multi Layer Perceptron is used to classify the vector into a predetermined named entity class.

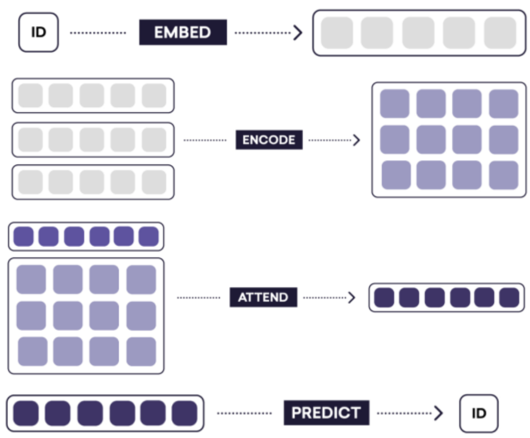

Figure 1: Embed, Encode, Attend and Predict (Honnibal, 2016)

4.2 Evaluation Metrics

The evaluation metric used in this task is the accuracy score. An address is declared to have the correct label when its POI and street address are correct. The calculation of the evaluation value is formulated in the equation:

[2] https://www.kaggle.com/c/scl-2021-ds/data

$$Accuracy = \frac{Number\ of\ correct\ sample}{Number\ of\ all\ sample} \quad (1)$$

5 Results and Discussion

In this study, the NER model has been trained as much as 100 iterations with Spacy libraries. The results of the accuracy of data training are 64.96% and 61.57% in test data.

I tried to predict all data trains to analyze things that block the model to produce a good performance. Of the 300,000 data trains, 16.307 data is predicted not to have point of interest and street address. This is a bad sign considering the metrics used are very sensitive and maybe it will immediately calculate the results of this prediction as wrong.

For some prediction results, there are differences between raw addresses with the point of interest and street addresses. For example, in Table 2, "rumah makan pela" was repaired to "rumah makan pelangi", "pp minhajutt" became "pp minhajutthollab" and "toko bang ajs" turned into a "toko bangunan ajs".

The correction of the word is not bound by certain rules, so it is impossible to improve the word prediction results. There are no specific rules that change the word "pela" to "pelangi" (in English: Rainbow). Likewise, with "minhajut" being "minhajutthollab", which in Indonesia is more often called "minhajut thullab" (with spaces and letter o turn into u). Finally, "toko bang ajs" can be translated into English as "Uncle AJS' Store", which is a characteristic of a common store in Indonesia. But the results of the growth change to "toko bangunan ajs", which can be translated as "AJS Building Shop".

Raw Address	Actual	Prediction
rumah makan pela, raya jomb,	rumah makan pelangi/raya jomb	/raya jomb
pp minhajutt, kh abdul manan, sumberberas muncar	pp minhajutthollab/kh abdul manan	/kh abdul manan
toko bang ajs	toko bangunan ajs/	toko bang ajs/

Table 2: Comparison of actual and predicted addresses

6 Conclusion

This research has successfully implemented Named Entity Recognition to extract the Point of Interest and Street Address from the address of raw shipping in Indonesia. By implementing Embed-Encode-Attend-Predict architecture, the model recorded an accuracy value of 64.96% for train data and 61.57% for test data.

I also investigated the weaknesses of our proposed method, such as how the model did not issue any predictions and how the model did not have the ability to correct the predicted named entity tokens.

I suggest further research to use additional techniques that can do text correction. In addition, ensemble techniques may also be applied to reduce the probability of the model does not issue any predictions.

References

Dwi H. Jayani. 2021. *Persaingan Dua Raksasa E-Commerce di Indonesia*. Katadata.co.id, Indonesia.

Bennett Kleinberg, Maximilian Mozes, Arnoud Arntz, and Bruno Verschuere. 2017. Using Named Entities for Computer-Automated Verbal Deception Detection. *J Forensic Sci*, 63:714-723. https://doi.org/10.1111/1556-4029.13645.

Gorjan Popovski, Barbara K. Seljak, and Tome Eftimov. 2020. A Survey of Named-Entity Recognition Methods for Food Information Extraction. *IEEE Access*, 8:31586-31594. 10.1109/ACCESS.2020.2973502.

Matthew Honnibal. 2016. *Embed, encode, attend, predict: The new deep learning formula for state-of-the-art NLP models*. Explosion.ai, Berlin, Germany.