Extended Named Entity Recognition using Bidirectional LSTM CRF Model

Seong-Won Na 1 and Kyoungro Yoon*2
1, 2 Konkuk University, Republic of Korea
(E-mail:securityin4@naver.com, yoonk@konkuk.ac.kr)
* Corresponding author

ABSTRACT

One important field of natural language processing is the Named entity recognition. In named entity recognition, various nouns are recognized as several types of semantic role, such as person, location, organization and others. However, various names are still not recognized as persons, and we want to extend the named entity recognition so that they can be properly recognized as persons. To do this, we modified the NER dataset to fit our purpose and we created our own test dataset to evaluate the model. We trained and evaluated the most commonly used Bidirectional-LSTM-CRF models and the CoNLL2003 dataset in the NER problem, so that more nouns can be recognized as persons.

Keywords: Named Entity Recognition, Bidirectional-LSTM-CRF Model, CoNLL2003 Dataset

1. INTRODUCTION

Named Entity Recognition, one of the tasks of natural language processing, recognizes nouns with unique meanings in sentences and can be divided into Person, Location, Organization, and MISC[1]. The Bidirectional-LSTM-CRF model, which is one of the most frequently used models to solve NER problem, is one of the recurrent neural networks and shows strong performance in Sequence problems. In addition, this model can predict more effectively by replacing the conditional random field which can use sentence level information with the last layer. In this paper, we used the model described above for the sentence analysis and conducted the experiment using the extended CoNLL 2003 NER dataset.

2. DATASET

CoNLL 2003 NER dataset is a set of sentences. Words contained in a sentence are paired with a label and are used as input of the network. However, since words such as Woman, Man, and Player, which are referring a type of person, are not labeled as PER, we have modified the labels of necessary words. Since the purpose of our system is to recognize people, other labels are used without modification.

3. BIDIRECTIONAL LSTM CRF MODEL

RNN is an effective network for processing sequence data. In theory, it can learn long dependencies, but it actually tends to be biased towards the most recent input. LSTM is designed to solve this problem through memory cells and enables long-distance dependency capture. Figure 1 shows the structure of the bidirectional LSTM CRF model. Unlike the existing LSTM CRF, it is designed to learn in both directions, so that both of the previous word and the next word are considered in the current label decision.
4. EXPERIMENT
In this paper, Collobert & Westone word representation is used for English word embedding used in Bidirectional-LSTM CRF model and dropout rate of 0.5 is applied to mitigate overfitting. The LSTM hidden layer dimension is set to 100 and the max epoch is set to 30, and it is learned by using the Stochastic Gradient Descent algorithm. For evaluation, the model were evaluated using a new dataset created by us other than the CoLL2003 NER test dataset. This dataset was collected on the Internet and the sentences to be evaluated used human-contained images as inputs to the CNN-RNN architecture and used output sentences after processing.

<table>
<thead>
<tr>
<th>Time</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
</tr>
</thead>
<tbody>
<tr>
<td>Word</td>
<td>A</td>
<td>Woman</td>
<td>is</td>
<td>Playing</td>
<td>Tennis</td>
<td>On</td>
<td>A</td>
<td>Tennis</td>
<td>Court</td>
</tr>
<tr>
<td>Before tag change</td>
<td>O</td>
<td>O</td>
<td>O</td>
<td>O</td>
<td>O</td>
<td>O</td>
<td>O</td>
<td>B-LOC</td>
<td>I-LOC</td>
</tr>
<tr>
<td>After tag change</td>
<td>O</td>
<td>B-PER</td>
<td>O</td>
<td>O</td>
<td>O</td>
<td>O</td>
<td>O</td>
<td>B-LOC</td>
<td>I-LOC</td>
</tr>
</tbody>
</table>

5. CONCLUSION
We learned and experimented using the Bidirectional-LSTM-CRF model for recognize Person in the extended NER dataset. A high accuracy of 93% was measured because the label was modified only for a specific word. If the dataset gets bigger, the recognition accuracy will be lower, but it will be possible to recognize a wider range of persons.

ACKNOWLEDGMENT
This work was supported by Institute for Information & communications Technology Promotion(IITP) grant funded by the Korea government(MSIP) (No.2017-0-00224, Development of generation, distribution and consumption technologies of dynamic media based on UHD broadcasting contents)

REFERENCES