Chinese Word Segmentation Based on Maximum Matching

1800017821 李云济

November 15, 2020

Abstract

In this paper, we report the results of a Chinese word segmentation task. We propose a novel method with the Maximum Matching algorithm as the base, and some artificial restrictions are attached. Experiments on dev set show that our model can complete the task decently and achieve competitive or better performances than some previous methods.

Key Words: Chinese word segmentation, maximum matching

1. Introduction

In the field of Chinese language processing, the importance of word segmentation can never be stressed enough. However, unlike English text in which sentences are sequences of words delimited by white spaces, in Chinese text, sentences are represented as strings of Chinese characters without similar natural delimiters. Therefore, the first step in a Chinese language processing task is to identify the sequence of words in a sentence and mark boundaries in appropriate places. This may sound simple, but in reality, identifying words in Chinese is a non-trivial problem that has drawn a large body of research in the Chinese language processing community.

2. Overview

Recently, neural models have been widely used for NLP tasks, for their ability to minimize the effort in feature engineering and their outstanding results. However, we are short of the help of some important machine learning libraries, such as TensorFlow and PyTorch, and we can clearly foresee the arduousness of constructing a machine learning model out of NumPy. Thus, in this report, we chose to accomplish this task using some traditional methods.

Our dataset contains a train, a dev, and a test, total about 70,000 segmented sentences in the train file, 17,000 in dev, and 4,000 unsegmented sentences in test. One of the most widely used methods in Chinese word segmentation is Maximum Matching, which will utilize the train file to build a vocabulary and match those words with the test lines. In our task, we will make a little improvement to MM method and apply it to our data.

3. Our method

1. Basic method

Firstly, we extract a word list, the vocabulary, from the training set, and filter out those one-character words and words whose length exceeds 30 characters. To increase the search efficiency, we group them by word length. The total words number is around 84,000. Then like the MM method, for each line unsegmented, we maintain a temporary word list for the sentence and scan it from left to right, each time trying to find the longest word in the vocabulary that match the forwarding characters in the sentence and append it to the temp word list.

For example, we have a sentence, “青年人在确定自己的人生坐标时”. “青年” and “青年人” may both dwell in the vocabulary, and since “青年人” has a longer length, we add “青年人” to the temp word list and turn the unsegmented line to “在确定自己的人生坐标时”. Finally, we will get a list of [“青年人”, “在”, “确定”, “自己”, “的”, “人生”, “坐标”, “时”], and by joining the words together with the delimiter “ ”, we will obtain the segmented sentence.

2. Problems and Improvements

The basic method gets a 0.95 F-score in the dev set, which is a considerably satisfying result. However, by checking the outputs of our model, we found some notable problems. Some single characters will misconnect with the following word or character, leading to an incorrect segmentation. Examples abound. “好学生”, which should divide into “好” and “学生” , because of the existence of the word “好学”, will become “好学” and “生” in our outputs.

To solve the problem, we first give Reversed Maximum Matching a try, which matches words from the end of a sentence to its beginning. To our surprise, the F-score drops by 0.01 because some new problems arise, such as “感受到” will be segmented to “感 受到”, which is unacceptable. Therefore, we decide to build a frontCh list manually, which contains some common front characters in Chinese. Every time one of the characters appears in the front place of a word, we will examine the possibility of an incorrect segmentation by checking if the back part of the word and the following character of the word can still form a new word. For instance, if a part of a sentence is separated into “完美” “地表” “现”, since “地” is in the frontCh list, we will check the existence of “表现”, and as a result, we will switch the segmentation to “完美” “地” “表现”, which fixes the problem.

To prevent excessive correction, such as “一个 人” to “一 个人” if “一” is in the frontCh list, we also maintain a backCh list, which contains some common ending characters, such as “了”. If a word ends with a character in backCh, it will not be changed even if it begins with a character in frontCh. We hope the two extra methods will keep a balance and help improve the result together.

With the improvements aforementioned, the F-score in the dev set increases to 0.986, which amazed us a lot. We reckon this incredibly high score results from the resemblance of the training set and the dev set, which means they may have similar themes of sentences and similar vocabulary.

To get a higher score on the test set, we combine the words in both training set and dev set together, to build the whole vocabulary, and we will use this vocabulary to do the segmentation of the test set. Although we notice that there is a word.txt in our evaluating file and it indeed contains the

vocabulary of both training set and dev set, we still finish the work of building the vocabulary from scratch. We also remove a tiny number of extremely rare and ambiguous words from the vocabulary to prevent mismatches, such as “面的” and “大西”, which we cannot tell the exact meaning and are absent from the test set.

4. Result

Our model gets 0.985 recall, 0.986 precision and 0.986 F-score on the dev set respectively, and in test set, we didn’t find any obvious error.

It is worth mentioning that if we add all the words with format of numbers or dates, such as “1533 亿美元” or “1997 年 11 月”, the result of test set will undoubtedly improve a little more, because these words are not supposed to be separated and some of them are not included in the vocabulary. However, we consider this feature engineering work too tedious and unworthy and it is not consistent with the purpose of automatic word segmentation, so we don’t do this work and just keep the original vocabulary without added words.

5. Conclusion

This report presents a novel model for the task of Chinese word segmentation, which adopts Maximum Matching as the foundation and adds some modification to it. The result on dev set is quite satisfactory. Nevertheless, the model still cannot handle the unknown words and absent words, such as some names and rare words. Much work needs to be done to evaluate this approach more thoroughly. We hope the model will gain a proper score on test set.