CSE567 Project

Simple Unsupervised Chinese Word Segmentation

Authors: Xingyu Yan, Yilei Zhao
The State University of New York at Buffalo

Abstract

In this project, we propose a joint model for unsupervised Chinese word segmentation. Unlike using complicated models to implement unsupervised Chinese word segmentation, our joint model combines two generative models, which are character-based hidden Markov model and Morfessor model. We optimize segmentation result with alternating and iterative executions on the two models. We conduct our experiments on PKU dataset. We use the corpus provided by the 4th SIGHAN Workshop to train and test the model. In current stage, the results are not promising. 200 tagged sentences are random sampled to test the model accuracy. The accuracy is around 54%, which means the method actually don't improve the segmentation.

Key words

unsupervised Chinese word segmentation, PKU dataset, tagging problem, probability based optimization algorithm, character-based hidden Markov model, transition probability, emission probability, morfessor, minimum description length, cost function, cookbook, iteration, converge, SIGHAN Workshop

Data and codes are available at https://github.com/i7242/CSE-567-Computational-Linguistic

Proposed Methods

Unlike using complicated models to implement unsupervised Chinese word segmentation, we propose a simple joint model which combines character-based hidden Markov model and Morfessor model. We optimize segmentation result with alternating and iterative executions on the two models. The steps in detail are listed below.

Steps of Method:

1.Segment the corpus by SA.

2.Using the segmentation result from SA, train the HMM model. The transition and emission probabilities will be calculated.

3.Using the trained HMM model to tag the same corpus. This gives some initial segmentation for CFM, even not good

4.The CFM reads the tagging result by HMM and generates a cookbook, then CFM segments the corpus based on this cookbook.

5.The HMM reads the segmentation result by CFM. From this result, we generate corresponding tags, and train the HMM using supervised training function. Notice that even the supervised training function is used, we have no idea about the tagging result from CFM. It is still unsupervised. After training, use the HMM to segment the corpus again.

6.Repeat Steps of 4 and 5, until the result converges.

HMM is one of the basic models which is frequently used in NLP. It contains two types of possibilities: the transition probability and the emission probability. In our case, we have only to states: N, S. For each state, we have emission possibility,which gives the most possible character in cur-rent state. Consider the HMM is probability based model, it may be difficult to establish these strong relations in an unsupervised situation. This is one of the reasons that we change from S, B, M, E tagging to N, S tagging.

Minimum Description Length Cost Function is used in Morfessor model. The Step 4 is to find the optimal segmentation of the source text into words. The cost function in Morfessor evaluates the cost of the segmentation strategy. If the cost increases, it is not suggest to segment the word, otherwise the word can be segmented.

One can think of this as constructing a model of the data in which the model consists of a vocabulary of words, i.e. the cookbook and the data is the sequence of text. We try to find a set of words that is concise, and moreover gives a concise representation for the data. This is achieved by utilizing a Minimum Description Length Cost Function. For the CFM, it is important to have one good cookbook initially. However, it is not available in our unsupervised method. We will assume out previous segmentation results from HMM or SA as a "cookbook"

Total Cost = Cost(Source text) + Cost(Cookbook)

The total cost consists of two parts: the cost of the source text in this model and the cost of the cookbook. The cost of the source text is thus the negative log-likelihood of the word, summed over all the word tokens that comprise the source text. The cost of the cookbook is simply the length in bits needed to represent each word separately as a string of characters, summed over the words in the cookbook.

Results

We use the corpus provided by the 4th SIGHAN Workshop to train and test the model. In current stage, the results are not promising. 200 tagged sentences are random sampled to test the model accuracy. The accuracy is around 54%, which means the method actually don't improve the segmentation. Since the test result is not good, we have nothing to say yet.