Project report for COMP5221

MAO Zhili 20169034

December 13, 2013

1 Project proposal

Translating "pinyin" into right Chinese sentences.

1. Create a very large parallel corpus where Language 0 is real Chinese character sentences, and Language 1 is the corresponding pinyin sentences. This would be easy to produce automatically, since you could automatically produce pinyin given Chinese character text.

2. Get as large as possible a dictionary of Chinese-characters-to-pinyin. These two together would be the training corpus, and could be directly submitted to SMT training.

2 Method

I assume the "pinyin" is a sequence of observations and the Chinese characters is a sequence of output. Between them the POS-tag is the hiden states. Besides the dictionary of Chinese-characters-to-pinyin, I used a corpus of the Chinese-vocabulary with POS-tag to built the pinyin-POStag-pinyin-Chinese model. In this model I have four Matrix as the defult input.

1. Transition_matrix: This matrix stores the transition probability of transiting from state to state, in another word, tag to tag.

2. Initial_matrix: This matrix stores the probability of a certain state is the first state in a sentence.

3. Emission_matrix: This matrix stores the probability of a certain observing to a certain state.

4. Vocabulary_matrix: This matrix stores the probability of a certain observing to a certain output Chinese.

The main task of this project is using Viterbi algorithm to segment the a sequence of observations what is discussed in the tutorial. The raw structure shows as fig.1

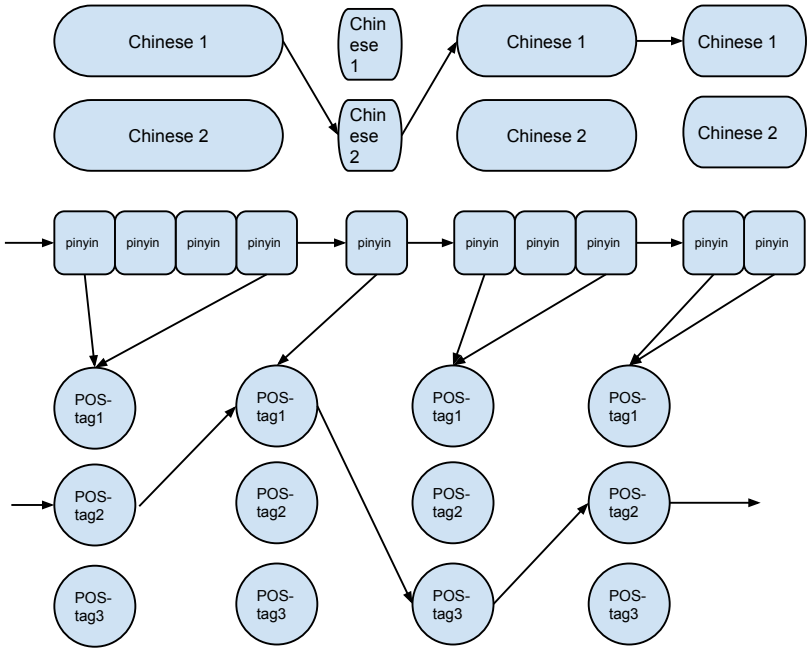

Figure 1: Structure

3 Result

The result is not very good. I guess it is due to the corpus. Run the demo.py and demo2.py could see the result.

4 Future work

1. Numbers and punctuation mark need to be considered.

2. More test and program improve to get the statistical result of this model.
3. Improve the model with new corpus or the algorithm.