Jieba- A Chinese Text Segmentation Module

1. Prelude

In case you, the reader, are not familiar with Chinese and its complexity, let's start with a scenario:

Suppose you can read Chinese, and you just borrowed a book written in Chinese, here comes the first line:

ISWORDSEGMENTATIONNECESSARYFORDEEPLEARNINGOFCHINESEREPRESENTATIONS?

2. Introduction

In Chinese, there are no separators between words in a sentence. The reason why it makes sense is that the number of Chinese characters is far more than in English, it would be relatively easy to pick up the "distributed patterns" in a sentence.

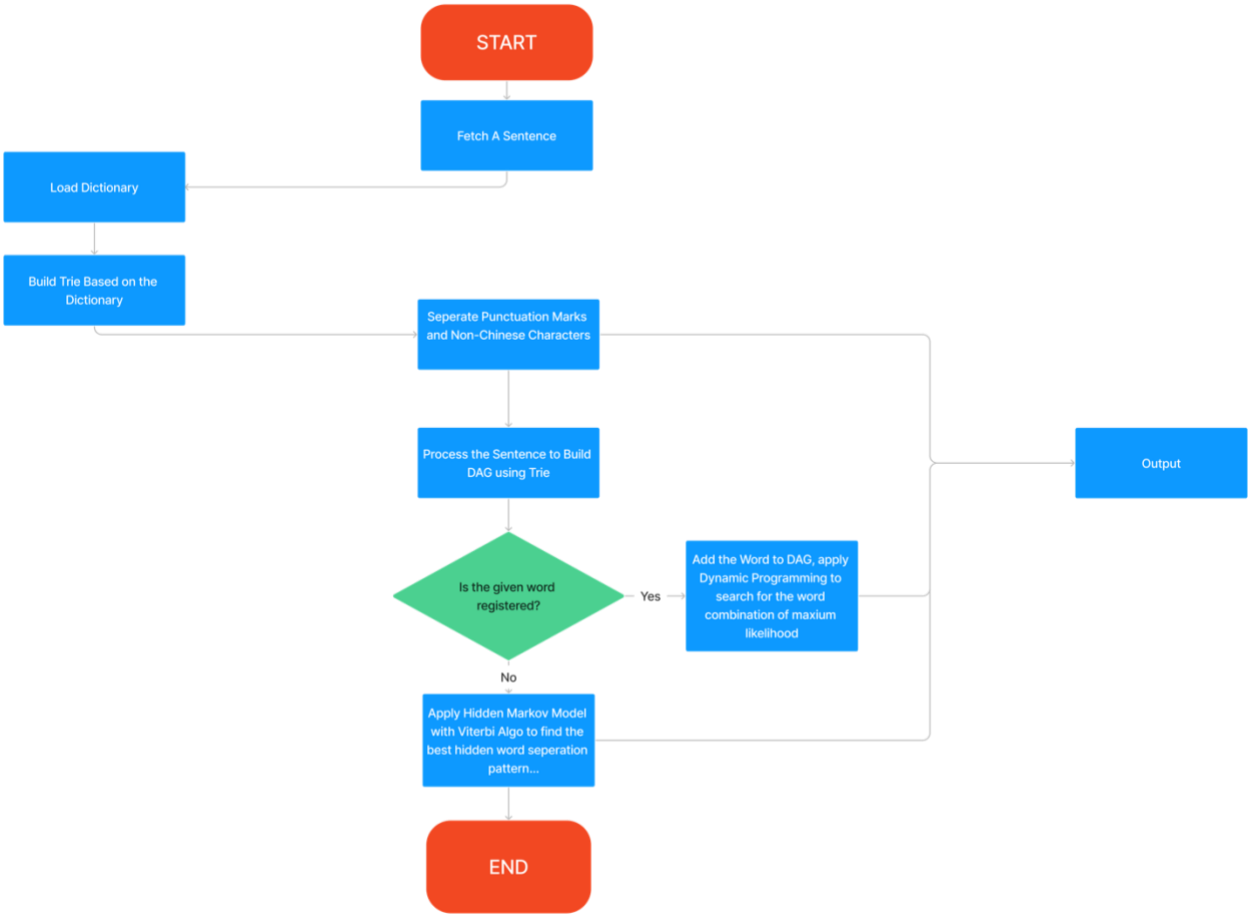
The problem arose when natural language processing is applied to Chinese: when we talk about syntactic analysis, we would always assume that words can be directly extracted from a sentence if separated by explicit characters (spaces or commas). But this is not the case for the Chinese, our humans can handle word segmentation empirically, but when it comes to machines, we would have to "help" machines tackle this problem.

That is what Jieba does: Text Segmentation. The name Jieba comes from the phonetic representation in Chinese describing people who suffer from a stammer. And the toolkit is designed to preprocess Chinese sentences into meaningful and vectorizable segments.

In fact, Chinese Segmentation is far more complicated than described (e.g. Entity Names Recognition, Word Ambiguation, Multiple Meanings in a word) and is still a very active research topic.

3. Mechanisms

The structure of Jieba can be demonstrate as follows:



(1) Hidden Makrov Model

The main idea is to use the Hidden Markov Model to predict the word separation. To be more specific, each Chinese character in the sentence would be regarded as one of the four types: B (beginning of a word), M(Middle), E(End), and S(Word of a single character). For example, “W” in the word “Word” should have the label “B”, and “r” should have the label “M” instead. For each word:

$$P(B) + P(M) + P(E) + P(S) = 1$$

There are also probabilities indicating the relationship between adjacent characters, for example:

$$P(E|B) = 0.851821$$

Such a number shows that in general, a word is very likely to end at the 2nd character, which exactly corresponds to Chinese: Most words consist of only 2 characters.

The rest of the work is to calculate the segmentation pattern that has the biggest probability using Viterbi Algorithm, which would be returned as the result. It's worth noticing that the probabilities are fetched directly from a saved file and can be edited based on needs.

(2) Trie

Jieba uses Trie to save indexed dictionaries because the number of Chinese characters can be insanely large. Compared to other options, the trie is fast and space-friendly.

(3) Dictionary-Assisted Algorithm

Indeed, the HMM is expected to perform well on meaningful sentences, but it doesn't work on entity names and non-separatable long sentences(such as idioms). Therefore, Jieba hereby introduces a dictionary to filter out the entities and sentences. Once a group of characters matches the pattern in the dictionary, it will be also considered as one valid candidate.

4. Evaluation

Jieba has now been released for 3 years and is still considered one of the good approaches to text segmentation. However, after reviewing the implementation, I think there are several things worth discussing:

1. The basic approach for Jieba still uses static probabilistic and therefore may not perform well on new words. To address that, the parameters clearly need to be updated routinely.
2. The model may over-rely on the dictionary. Since all words that are not registered would be sent to the probabilistic model, it is crucial to choose a dictionary that covers most of the existing entities.
3. However, the built-in dictionary is already large(160MB) and requires plenty amount of time in the loading process. If we increase the size, things will get worse.

To sum up, jieba is an excellent toolset for Chinese text segmentation that supports multiple languages. It is super-fast and can perform comparably well.

5. Reference

1. Fxsjy (no date) FXSJY/Jieba: 结巴中文分词, GitHub. Available at: https://github.com/fxsjy/jieba (Accessed: November 6, 2022).

2. 无敌小想法 (no date) Chinese word segmentation mechanisim explanation+jieba Introduction(2nd Part), zhihu. Available at: https://zhuanlan.zhihu.com/p/66904318 (Accessed: November 6, 2022).