A Basic Syntax for Mandarin Chinese

Yixiao FEI, Xin ZHANG

I Introduction

I.1 Goal

The project tends to implement a grammar for a new language other than English or French, which has been already demonstrated in the lab courses during the study. As native speakers of Mandarin, which is largely different with any western language, we consider it interesting to apply the syntax analysis to Mandarin.

Since the Prolog doesn't support Chinese character well enough, we decide to realize the application on Python, which support almost any character encoded in "utf-8". The environment we used is based on "python3.6+" and do not require any other external library.

I.2 Challenge for the Syntax Tree Implementation

As the only fully logogram writing system still in wide use, the parsing of the Chinese texts has many unique problems to solve. First there is no space between words (actually every Chinese character is a morpheme), we need to separate them before parsing. Secondly, Chinese doesn’t have strong and evident grammatical rules, as there is no conjugation, no gender, no tenses or even the distinction between verbs and nouns. But there are preferences and restrictions on how words should be placed and used. Thus, new grammar rules are required for the syntax analysis. The grammar we implemented is a simple one at A2 level, because it seems complicated to realize everything in a single project.

II Word Segmentation

II.1 FMM

The algorithm used for the word segmentation is FMM (Forward Maximum Matching). The algorithm begins with the first character of the left hand of the sentence. It tries to find a word, which begins with this character, in a simple dictionary defined by ourselves. The word should contain the largest length among all the possible matching words. Then the process of matching goes to the next character right after the matched word.

II.2 Limit

The exact implementation can be found in the source code. However, this algorithm may not be the optimal solution because sometimes the word segmentation after FMM may not be meaningful at all. Here the algorithm performs well enough because our dictionary is not so large, but if we add more words into the dictionary, we may need a more powerful algorithm. In this case, the grammar also need to be changed.

III Grammar Implementation

III.1 Basic Grammar

As an SVO (subject-verb-object) language, a Chinese sentence can still be separated to noun phrases and verb phrases.

For noun phrases, all modifiers, including adjectives, numbers and particles should be placed in front of the noun. Similarly, adverbs should also generally appear before the verb. One thing that requires particular attention is that quantifiers should be used between numbers and nouns. Additionally, there is no “link verb” between a noun and an adjective after it, but normally there should be a degree descriptor like “很(very)”, “挺(kind of)”, “有点(a bit of)” between them.

III.2 Particles

A major feature of the Chinese language is the use of “particles”. One is the connecter “的” which is used to put adjectives in front of nouns as well as declare possession, though it is regularly omitted. For example, “暴(brutal)虐(cruel)的君(lord)主(master)” and “暴(brutal)君(lord)” mean the same thing, but the latter one is mainly used as “tyrant” and the former one as a description. This is a common case in Chinses, likely, the majority of Chinese vocabulary are self-explanatory. Another group of particles are combined with verbs: After being added after the verb, “地” makes the verb an adverb,

“过” or “了” indicates the action is in a far and near past and “着” indicated the action is ongoing. Last but not least, there are also two accusative particles to connect two entities. “把” has the action performer in front and the action receiver close after (the action receiver is before the verb), “被” works inversely and makes a passive voice.

III.3 Limit

The dictionary we used is a small one which contains basic words for a A2-level speaker. Adding more words into the dictionary may cause cross ambiguity and combination ambiguity.

Many words can be regarded as different types of words, for example either a verb or a noun. Listing all possible rules in the grammar is tedious and may generate a wrong parse tree.

The implementation of a higher level of Chinese require more complicated grammars since the language is more meaning focus not structure focus. For example, a sentence often omits a verb but still can express a complete meaning.

IV Conclusion

In this implementation of a basic Mandarin Chinses parser, only the basic language rules, which correspond to an A1 level, are included. During the implementation, we have observed that though it is hard to apply a structure based on Western languages to Chinese, there are some conveniences like we can basically use Adv -> V particle_adv as a rule and use V -> V Particle_time to save time from the present and past tenses of verbs.

For further improvements, firstly we are considering to expand the grammar to a higher level by introducing more rules, but still it will be hard to push really far as high level Chinses involves classical Chinese, where the language becomes monosyllabic and really flexible in word orders like Latin. Secondly, semantic feature can be introduced tin order to ensure correct pairs of quantifiers and nouns and evite nonsenses.