Performance Evaluation of English Part of Speech Tagging
Based on Multi-feature Knowledge Algorithm

Jian Li*
Foreign Languages department of Humanity School, Yantai Nanshan University, China, 265713

*Corresponding author e-mail: 1291855800@qq.com

Abstract. In order to solve the traditional English part-of-speech tagging methods, a variety of English part-speech tagging methods are introduced. This paper systematically studies the method of multi-feature knowledge algorithm to mark English part of speech, then demonstrates it theoretically, and gives some examples to illustrate the characteristics of this method, and then evaluates it objectively.

Keywords: Multi-Feature, Knowledge Algorithm, English Part Of Speech

1. Introduction
Part of speech refers to the syntactic functional categories of words. In natural language, a word often has multiple parts of speech[1]. The purpose of part of speech tagging is to uniquely determine the part of speech in a sentence by the context of the word. The marking of the corpus is the basis for the in-depth processing of the corpus[2]. Part of speech tagging, as the name implies, is to add tagging to the part of speech of each word to determine the lexical category to which the word belongs, such as nouns, verbs, pronouns, or other parts of speech. It is a process that can bring value-added effect to corpus, which greatly increases the depth of corpus retrieval and makes the use of corpus more diversified[3]. At the same time, part of speech tagging is also the basis of natural language processing, which is an indispensable link in the application of machine translation and information retrieval. To achieve more efficient English part of speech tagging, we must rely on high accuracy tagging method. It can be said that the accuracy of marking means is the cornerstone of language processing firm or not. For such two high-accuracy methods, this paper tries to compare the similarities and differences between English and Chinese, to see the difference between English and Chinese language behind the noun annotation, and then take the language difference as the guide to put forward the evaluation and suggestion of the multi-feature knowledge algorithm for English part of
speech annotation, so as to further improve the accuracy of coding.

2. Models and templates for multi-feature knowledge

2.1. Model implication

Conditional randomness is a discriminative graph-free probability model, which can calculate the conditional probability of the output node under the condition of a given input node[4]. Theoretically, describe the conditional independence in a marker sequence whose structure can be arbitrary. But the simplest and most important is the linear chain structure, which assumes that the sequence satisfies the first-order Markov attribute. Conditional randomness has two advantages in application: one is that there is no strict independence assumption for the observed values. all properties of the observations can provide the basis for the value of the state variable on any vertex of the decision[5]. The second is to establish the joint probability of the whole marker sequence, which cannot only avoid the mark bias problem such as maximum entropy, hidden Markov model, but also obtain the importance of expressing features under global conditions by global normalization of all features, so that the model can make full use of feature information. in view of this, conditional randomization can well overcome the limitations of the independence hypothesis and the Markovian hypothesis, thus effectively exploiting the contextual features in the text linear sequence[6]. Figure 1 is a schematic diagram of two different of the part of speech tagging model algorithms.
2.2. Feature selection and template determination

How to choose the appropriate template for part of speech tagging is the key task of multi-feature knowledge algorithm in part of speech tagging. The multi-feature algorithm template refers to the statistical information of the information set word structure such as the length of the sequence unit, the front and back words, the speech, the correlation collocation and the probability as the word combination for the training process of the conditional random model. There are many complex components in the sentences in English texts, and the phenomenon of one word bearing many parts of speech is more common. The same word in different positions in the word often reflects the word's part of speech. Based on this, the first and last words and their parts of speech are statistically calculated respectively, and the information is used as the feature item to provide the knowledge algorithm. The statistical formulas for calculating a word as a specific part of speech are as follows.

\[ P_{BE-j} = \frac{N_j}{\sum_{k=1}^{BE-k} N_i} \]  
(1)

Where \( P_{BE-j} \) denotes the probability that the word is \( j \) when it appears as a single word. Character \( N_j \) is the frequency of the word when the word is \( j \), and \( BE-k \) denotes the sum of all the parts of speech that can be represented when the word is a word. The first and last words are expressed as the statistical formulas for specific parts of speech as follows.

\[ P_{B-j} = \frac{N_j}{\sum_{k=1}^{BE-k} N_i} \]  
(2)

\[ P_{E-j} = \frac{N_j}{\sum_{k=1}^{E-k} N_i} \]  
(3)

PB-\( j \) and PE-\( j \) denote the probability that the word has a part of speech of \( j \) when it appears at the
beginning and end of the word. B-K and E-K denote the sum of all parts of speech when the word begins and ends.

3. Examples of performance assessment for part of speech labeling

3.1. Number of nouns

English is a language with twists and turns, but Chinese is not. English nouns are divided into countable nouns, uncountable nouns and proper nouns. The proper noun is a kind of special vocabulary, often refers to the person name, the place name or the organization name and so on, marks relatively easy. But in the common noun class, there is a single plural. Countable nouns like book are not grammatical to appear in the text alone. In other words, such separate entities are determined by counting. This property, in turn, constrains the use of such words in the form of articles or twists and turns in the plural. And uncountable nouns, such as furniture, do not apply to the constraints of countable nouns. This kind of noun belongs to the collective noun, no zigzag change, no plural suffix, only the corresponding article or adjective to modify or limit it. In English, in addition to the obvious single complex number, there is also single complex number homomorphism. These nouns can only be judged by modifying a qualifier or verb. However, on the other hand, Chinese is not completely independent of the single plural language, but most nouns determine the single complex by modifying the qualifier. Chinese does have such plural expressions as "children" and "sheep ""We" and "group" can correspond to the plural implementation form "-s" in English, and also act as a noun plural suffix, and the articles and modifiers that appear before these nouns are usually also plural, such as “these, those ”. Therefore, it can be said that the Chinese part of the single complex form can also be identified by suffixes or qualifiers. Through the comparison of single-complex forms, it can be seen that although the classification given by the two label sets is different, there are still similarities between the two languages. Therefore, the Chinese tagging set can also try to distinguish the single plural when classifying nouns. By classifying the ends of these plural words, we can better distinguish the semantics and refine the annotation classes.

3.2. The division of quantifiers

The division of quantifiers is one of the obvious differences between the two label sets. Through this difference, we can see the different division and usage of the same concept in Chinese and English. Chinese numerals must be quantifiers to modify nouns. Only in idioms or sayings can numerals directly modify nouns, and this is a small number of special cases, which are not applicable to the general majority of cases. For example, there are numerals and quantifiers in Chinese, while in English there are only numerals and boundless parts of speech. The expression of the concept of quantity in English often refers to the corresponding nouns, which constitutes the difference of the concept of quantity between English and Chinese in the form of words and the structure of quantitative phrases. The concept of quantity in English is often expressed with the help of corresponding nouns. Some scholars also believe that the concept of expression in English is implicit in the deep semantics of this construction. It divides numerals into cardinal words and ordinal words, but at the same time they also contain cardinal parts of speech, which will undoubtedly affect the accuracy of tagging parts of speech, and may appear the situation of repeated tagging.
3.3. Time, place and position
Under the proper noun classification, four of the six categories are nouns for week and month. This division also stems from the difference in the way the two languages express their time. The time noun is an indispensable word to indicate the specific time, and must be collocated with the appropriate preposition to express the complete semantics. Therefore, it is reasonable to divide time-marked nouns under proper nouns. However, it doesn't seem reasonable to classify other time markers as other categories if only the nouns indicating the week and month are listed. In addition to weeks and months, there are nouns for time. There are also some words that do not need to be used with prepositions, but are also time-marked words, so it doesn't seem to make much sense to mark them only as common nouns. Of course, this is related to the way Chinese expresses time. The way of expressing time in Chinese is flexible, so it can be added preposition or omitted in the use of time-marked nouns. In this way, it is more efficient to label such time morphemes uniformly. The place word and the direction word also have the same situation to express the place, the position noun all under the noun classification. The labeling of the former is more convenient for the corpus observer than the classification of the former under the noun because of the different marks.

4. Conclusions
In the practice of automatic correction of English articles, the traditional method of part of speech tagging has encountered the problem of low correct rate in English articles. The author makes a thorough and detailed analysis of this, puts forward an English part of speech tagging algorithm based on multi-feature knowledge algorithm, and evaluates it through examples.

References
[1] Julie, Meng Yao, Zhao Tiejun. Performance evaluation of typical parametric smoothing algorithms in part of speech tagging [C ]// Language computing and content-based text processing - Proceedings of the Seventh National Joint Academic Conference on Computational Linguistics. 2003.
[2] Tan Yongmei, Wu Kun. An Algorithm for Part of Speech Marking for English Articles [J]. Journal of Beijing University of Posts and Telecommunications,2014,37(6):120-124.
[3] Wang Dongbo, Huang Shuiqing, he Lin. A study on automatic tagging of pre - Qin books based on multi - feature knowledge [J].
[4] Su Yulong. textual information analysis based on deep learning [D]. Guizhou University.
[5] Tan Yongmei, Wu Kun. An Algorithm for Part of Speech Marking for English Articles [J]. Journal of Beijing University of Posts and Telecommunications,2014,37(6):120-124.
[6] Fu Huaiqing. Cognition of the Pronunciation of Ci in Modern Chinese Dictionary [J]. Language and Text Applications ,2004(2):41-43.