Performance Evaluation of English Part-of-Speech Tagging Based on Typical Parameter Smoothing Algorithm

ChaoYi Wang*
Jilin Communications Polytechnic, China, 130000

*Corresponding author e-mail: 49838708@qq.com

Abstract. Part-of-speech (POS) tagging for English is the basis for implementing English automatic correction. Although researchers have done a lot of useful studies on English POS tagging, most of them are aimed at users with English as the first language, while studies for users with English as the second language are few. For this purpose, manual tagging is performed based on the typical parameter smoothing algorithm. On this basis, a performance evaluation method for English part-of-speech tagging is proposed, which integrates the features of term clustering, non-tagged corpus statistics, word pronunciation, etc. The experimental results show that the algorithm can improve the performance of POS tagging effectively, with the tagging accuracy improved from 94.49% to 97.07%

Keywords: Part-of-speech Tagging, Student English, Features, Term Clustering

1. Introduction
Part-of-speech (POS) tagging is to judge the grammatical category of each word in a given sentence and determine the POS [1]. Researchers have done a lot of useful work on the research of English POS tagging, and many excellent POS tagging systems have emerged [2]. The tagging accuracy has reached 97% ~ 98%, which has reached people's expectation of automatic POS tagging. POS tagging was once considered a "solved" problem. However, the POS tagging When the tagging system is applied to another new text, the tagging accuracy will be significantly reduced [3]. Most POS tagging systems are trained based on official English news articles, such as Binzhou Tree Library [4]. Since spelling errors, loan words, proper nouns, fabricated words, and other unknown grammatical errors often occur in English, there will be many problems when these excellent POS tagging systems are applied to English.

In this paper, the English POS tagging is carried out manually, and a performance evaluation algorithm in English POS Tagging Based on the easy first strategy and two-way dependency network is proposed. The algorithm makes full use of large-scale non-tagged corpus [5]. Firstly, the non-tagged corpus is preprocessed to extract statistical information. Secondly, Brown term clustering algorithm is
used to cluster large-scale non-tagged corpus to obtain cluster information. Subsequently, Metaphone speech matching algorithm is used to generate pronunciation key values of words using [6]. Finally, experiments are conducted on the above-mentioned manual tagging corpus and official news corpus. The experimental results show that the proposed algorithm is very effective for POS tagging of English and official news.

2. Typical parameter smoothing algorithm

The typical parameter smoothing algorithm generally consists of two parts. One part is the structure of the typical parameter smoothing algorithm, which is a directed acyclic graph (DAG). Each node in the graph represents the corresponding variable, and the connection between nodes represents the conditional independence of the typical parameter smoothing algorithm. The other part is the conditional probability table (CPT), which is a series of probability values. If the typical parameter smoothing algorithm provides enough conditional probability value, it can calculate any given joint probability, which is called computable in this paper. For the hypothesis of vertex $x_i$, its parent node set is $Pai$, each variable ($X_i|Pai$) The conditional probability of is $X_i$, and $X = \{ x_1, x_2, L, x_n \}$. The joint probability distribution of is calculated as follows:

$$P(X) = \prod_{i=1}^{n} P(X_i | Pai)$$

The simplified joint probability formula in the typical parameter smoothing algorithm is as follows:

$$P(x_1, x_2, x_3, x_4, x_5, x_6) =$$

$$P(x_6 | x_5) gP(x_5 | x_2, x_3) gP(x_6 | x_1, x_2)$$

$$gP(x_3 | x_1) gP(x_2 | x_1) gP(x_1)$$

(2)

Once the correlation between propositions is represented by a directed arc, the conditional probability is represented by the arc weight. It represents knowledge about the relationship between propositional static structures. When obtaining new evidence, the possible value of each proposition should be comprehensively checked, and then the trust degree of each node should be defined as $B(x)$. The following can be specified:

$$B(x) = P(X = x_i | E)$$

(3)

It shows that under the current e condition, all facts and evidence are provided, proposition $x$ is the trust degree of $X_i$. Subsequently, $B(x)$ trust degree is calculated according to the evidence and facts. By using formula (2) and type joint probability formula, the calculation of $B(x)$ is greatly simplified

3. Performance evaluation algorithm in POS tagging

The algorithm flow is shown in Figure 1
Figure 1. performance evaluation algorithm flow in English oriented POS tagging

1) After preprocessing the unmarked corpus, we get the formatted data, and then cluster the words to get the cluster information of each word;

2) The POS tagging of standardized data is carried out, and the possible tagging and frequency of each word are calculated;

3) Match the pronunciation of the words on the labeled data to obtain the statistical information of the pronunciation of the words;

4) Based on the features of cluster information, statistics information, word pronunciation information, etc. are tagged based on the easy first tagging strategy, according to the evaluation function (see formula (1)), the POS tagging of the unmarked words is carried out. The algorithm is as follows. To fully leverage the tagging information of the left and right sides of Ti, the calculation is based on the bidirectional dependency network

\[
\text{score}(w_i, t_i) = p(t_i | t_{i-2}, t_{i-1}, t_{i+2}, w_1, \cdots, w_n, C_w_i, M_w_i)
\]  

(4)

\(w_i\) represents the word at the position I in the sequence, \(t_i\) represents the POS of the word at the I position in the sequence, \(C_w_i\) represents the term clustering feature related to \(w_i\), \(S_w_i\) represents the statistical information characteristics of non-tagged corpus related to \(w_i\), \(M_w_i\) represents the pronunciation features of the words related to \(w_i\).

Algorithm 1: easy-first sequence tagging

input: \(W = w_1w_2\cdots w_n\)

output: \(T = t_1t_2\cdots t_n\)

1. \(\text{tagset} = \{t_1, t_2, \cdots, t_I\}\)

2. \(\text{unlabeledset} = \{w_1, w_2, \cdots, w_i, \cdots, w_n\}\)
Finally, and discussion.

Street tagging, "???".

PIGAI tagging the semantically (source represents network, every relatively difficult easy"

Upon the tagging of wi, the context information of wi is broken down into many situations to select the most suitable situation for each wi in the process of tagging, and avoid the repeated calculation of tagging information.

3. while unlabelset ≠ "do
4. \( w_j^*, t_j^* \) ← \( \arg \max \{ \text{score}(w_j, t_j); w_j \in \text{unlabeledset} \} \);
5. remove(unlabeledset, wi)
6. return T

Easy first tagging strategy first finds the word “easiest” in the sentence, then finds the word “next easy” for tagging, until all the words are tagged, so we should postpone the tagging of the most difficult words to the last. When tagging the most difficult words, the POS information of the relatively easy tagging words on both sides of the word has been obtained. In this way, it can be used to provide auxiliary information for current word tagging, to ensure the correctness of tagging results, and the propagation range of tagging error can also be suppressed to a certain extent. The “easy” degree is determined according to the evaluation function, which not only selects a reasonable tag for every single word but also determines the tagging order of each word.

Figure 2 shows a simple dependency network with two nodes. Figure 2 (a) and (b) are easy to understand and semantically clear graph models. Figure 2 (c) contains rings, not a standard Bayesian network, but a more comprehensive two-way dependency network. Each node in Figure 2 (c) represents a random variable, representing a conditional probability model with all incoming edges (source variables). Therefore, two-way dependency networks and standards Bayesian networks are semantically consistent.

Upon the tagging of wi, the context information of wi is broken down into many situations to select the most suitable situation for each wi in the process of tagging, and avoid the repeated calculation of tagging information.

Figure 2. Simple dependency network

4. Experiment

To make the study more scientific, the tagging process is as follows:

1) Corpus preprocessing. 15000 sentences were randomly selected from 52115096 sentences of PIGAI ③. After filtering out the nonstandard sentences that are excessively short/long, the remaining 14115 sentences were divided into one mark with incorrect punctuation, such as "------" "----" "!!!!" "??!". OpenNLP ④ was used for corpus tagging

2) Manual tagging. Four taggers corrected the above tagging results, recorded the ambiguous tagging, and discussed and determined the final tagging results by referring to the tagging in the Wall Street Journal corpus

3) Review of the tagging results. 2 taggers review and revise all sentences marked in the first stage, record the dissenting tagging, and determine the final tagging results through tagging group discussion. The third tagger relabels 300 sentences without referring to the tagging results in step 2), and the tagging consistency rate is 93.79%, which indicates that the tagging results are effective. Finally, the fourth tagger scans the whole corpus to correct the errors in manual tagging and further
improves the consistency of corpus tagging.

The experimental data are 14022 sentences randomly selected from the above corpus (Section 3.1), including 10000 sentences in the training set, 1780 in the development set and 2242 in the test set. The unmarked corpus is 51840035 sentences randomly selected from PIGAI. The main experimental data statistics are shown in Table 1, and the algorithm tagging results (accuracy) of different feature combinations are shown in Table 2.

Table 1. Statistical information of the main experimental data

| Experiment setup | Number of sentences | Number of words | Words not logged in |
|------------------|---------------------|-----------------|---------------------|
| train            | 10000               | 241178          | 0                   |
| Development      | 2416                | 56684           | 3634                |
| test             | 1346                | 32987           | 2363                |

Table 2. Algorithm tagging results of different feature combinations (accuracy)%

| Feature combination          | Development set | Test set |
|-----------------------------|-----------------|----------|
| basic feature               | 95.95           | 95.99    |
| Term clustering             | 96.40           | 96.40    |
| statistical information     | 96.95           | 97.03    |
| Word pronunciation          | 96.21           | 96.28    |
| Term clustering + statistics| 97.00           | 97.04    |
| Term clustering + word pronunciation | 96.48 | 96.51 |
| Word pronunciation + statistics | 96.92 | 97.02 |
| All characteristics         | 97.01           | 97.07    |

According to the experimental results in Table 2, the POS tagger with all the features has the best performance, and the accuracy reaches 97.07% in the test set. A series of feature reduction experimental results show that statistical information is an active embodiment of vocabulary knowledge, term clustering has an excellent performance on vocabulary knowledge, and word pronunciation has a specific performance on vocabulary knowledge.

5. Conclusions

In the practical application of English automatic correction, the traditional POS tagging machine faces the problem of low accuracy in English tagging. The author conducted a thorough and detailed analysis of this problem, combined the existing corpus resources to perform manual English tagging, and proposed a performance evaluation algorithm for English POS tagging. This algorithm has integrated the features of term clustering, non-tagged corpus statistics, word pronunciation, etc., which is experimented on the tagged corpus. The experimental results have verified that the algorithm can improve the tagging accuracy effectively. Finally, the POS tagging for English is applied to the English intelligent assessment system. The feedback from users shows that the assessment system shows a good improvement on the grammar check of some typical wrong sentences.

References

[1] Sun, W., & Wan, X. (2016). Towards accurate and efficient chinese part-of-speech tagging. Computational Linguistics, 42(3), 1-30.
[2] Senouci, F., Chioukh, N., & Dris, E. A. (2019). Performance evaluation of bottom-standing
submerged breakwaters in regular waves using the meshless singular boundary method. Journal of Ocean University of China, 18(4), 823-833.

[3] Bohoua, G., & Veton Këpuska. (2018). Performance of wuw and general asr speech recognition systems in different acoustic environments. The Journal of the Acoustical Society of America, 143(3), 1758-1758.

[4] Masuda, H. (2016). Misperception patterns of american english consonants by japanese listeners in reverberant and noisy environments. Speech Communication, 79, 74-87.

[5] Milvae, K. D., Alexander, J. M., & Strickland, E. A. (2016). Investigation of the relationship between cochlear gain reduction and speech-in-noise performance at positive and negative signal-to-noise ratios. The Journal of the Acoustical Society of America, 139(4), 1987-1987.

[6] Rezai Mohammad Javad, & Mosavi Miangah Tayebeh. (2016). Farsitag: a part-of-speech tagging system for persian. Digital Scholarship in the Humanities(3), 3.