What affects the difficulty of Chinese syntax?

Yueming Du*, Lijiao Yang§
Institute of Chinese Information Processing, Beijing Normal University
UltraPower-BNU Joint Laboratory for Artificial Intelligence
Beijing, China
*Email: *ddddym@yeah.net, §yanglijiao@bnu.edu.cn*

***Abstract*—The traditional measurement of sentence difficulty only focuses on lexical features but neglects syntactic features. This paper takes 800 sentences in primary school Chinese textbooks published by People's Education Press as the research object and studies their syntactic features. We use random forest to select the top five important features and then employed SVM to do the classification experiment. The precision rate, recall rate and F-scored for the classification of 5 levels are respectively 50.42%, 50.40% and 50.41%, which indicates that the features we selected has practical value for the related research.**

Keywords-Chinese syntax; random forest ; classification; SVM;

I. INTRODUCTION

Linguistic complexity is a well-studied and multifaceted concept. Several measures are proposed in different frameworks, such as first and second language acquisition, typology and readability assessment. However, there exit two problems in previous studies. The first one is that most researches are based on the chapter level to study the difficulty of linguistic rather than based on a single sentence, which is the basic components of the text. Sentences are the basic unit of a text. Although, at present, there are many studies, which regard the length of sentences (Laughlin,1969), the use of simple sentences or complex sentences as important factors (Dechant,1961), the variables they measured were nothing more than vocabulary difficulty and the shallow sentence characteristics. The current situation, however, is that scholars study the super-outline words and phrases in sentences from the perspective of the overall number of texts, rather than as part of a sentence, let alone a specific sentence as a complete unit. The investigation of individual sentences is helpful to the development of AES and text readability. Second, researchers pay more attention to Chinese lexical features, but rarely to the features of Chinese syntax relate to the complexity. On the contrary, there are many studies on the syntactical difficulty in English. Hunt suggested that T-unit was a useful factor to exam English syntax（1965）. Norries and Ortega (2009) further summarized the multi-dimensional indicators of syntactic complexity into five sub-dimensions, including the use of subordinate structures, the overall complexity of sentences, the phrase expansion of clauses, the use of juxtaposition structures and the diversity, complexity and acquisition order of sentence structures. Lu selected 14 indicators to measure English syntactic complexity on the basis of quantitative analysis and qualitative analysis. However, whether the characteristics applied in English can be used in Chinese and what characteristics are effective in measuring the complexity of Chinese syntax remains to be studied.

Syntax refers to the rules of conjunction formation in linguistic units. With the development of second language teaching, syntactic complexity has become one of the most important indicators to measure learners' overall language proficiency. However, in the task of language difficulty analysis in the field of Chinese, there are few studies on the measurement of its syntactic complexity.

In our study, we randomly selected 800 sentences from the primary school Chinese textbooks of the People's Education Press as the research object, and used HanLP[1] as well as StanfordCorenlp[2] to extract their syntactic features automatically. At the same time, we asked three undergraduates with linguistic background to rate the syntactic complexity of the sentences and proofread the extracted syntactic features manually. On this basis, we further analyzed the contribution of each feature to the complexity of syntactic.

II. OUR METHOD

A. Preprocessing

The syntactic complexity explored in this paper is mainly in the field of modern Chinese, so classical Chinese and ancient poetry are not within the scope of this study. Because of the particularity of modern and contemporary poetry and nursery rhymes, their syntactic complexity is also beyond the scope of this paper.

We use primary school Chinese textbooks published by People's Education Publishing House (hereinafter referred to as "textbooks") as the basic corpus. After eliminating some texts that do not conform to the content of this study (ancient poems, classical Chinese, modern and contemporary poetry, nursery rhymes, etc.), we cut each text into sentences and randomly extract 800 sentences. At the same time, we use HanLP and StanfordCorenlp to make word segmentation, part-of-speech tagging and dependent grammar on the 800 sentences.

We recruited 3 undergraduates with Chinese linguistic backgrounds and asked them to rate the syntactic complexity of the above data on a 5- point scale where 1 means "very easy" and 5 "very difficult", the distribution of syntactic levels is shown in Figure I. In addition, we asked them to check the results of word segmentation, part-of-speech tagging and dependency grammar analysis above, and to mark the inter-sentence relationship of each sentence.) We computed the Kappa coefficient reliability corresponding to the number of annotators who assigned the same judgment and obtained a reliability of 29%.

Corresponding author: yanglijiao@bnu.edu.cn

[1] https://github.com/hankcs/HanLP
[2] https://stanfordnlp.github.io/CoreNLP/

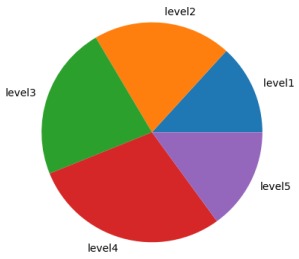

Figure I. Distribution of syntactic levels

B. Raw Sentence Feature

1）*sentence length*, i.e. average number of words per sentence (n_tokens). The basic unit of text is sentence, but the length of sentence is limited. Most researchers believe that sentence length can reflect the complexity of syntax, that is, the longer a sentence is, the more complex components it contains, and the more complex it tends to be. Therefore, we calculated the average sentence length in each level and the statistical results. We find that with the increase of the difficulty of the syntax, the average sentence length basically shows an upward trend.

2) *Punctuation*, the number of punctuation marks in a sentence. The more punctuation the sentence has, the more fragmented it is and the more difficult it is to understand. At the same time, punctuation to a certain extent reflects that the more insertions the sentence has, the more difficult it will be. The result shows the distribution of punctuation symbols.

C. Ssyntactic Features

1) *The max_depth of the whole parse tree*, a syntactic tree is a graphical representation of a sentence's syntactic structure. It shows us the complex nesting relationships among the syntactic components in the form of a tree graph. Therefore, the depth of the syntactic tree reflects the complexity of the sentence to a certain extent.

The methods for calculating the depth of the syntax tree are as follows: first, the dependency grammar analysis of sentences is performed by Stanford Core NLP; secondly, the syntax tree is transformed into a binary tree to calculate the number of nodes from the root node to the farthest leaf node; finally, the maximum depth of each syntax tree is obtained.

In order to visually reflect the corresponding relationship between the maximum depth of the syntactic tree and the syntactic difficulty, we calculated the maximum depth of the syntactic tree of each corpus in turn. As can be seen, the maximum depth of the parse tree increases with the increase of the syntactic difficulty.

2) *Modifier:* we will calculate the length of modifier and the ratio of length of modifier to sentence length. We label the modifier of 800 sentences manually and count the length of the modifier by rule method. On this basis, we get the values of these two features under different syntactic complexity. Both of the two features are positively correlated with text difficulty, but the second feature, i.e. distribution of the ratio of length of modifier to sentence length, shows an upward trend of fluctuation with the increase of text difficulty.

3) *Number of inter-sentence relationship types.* The relationship between sentences involved in this study is different from the meaning types of complex sentences defined in Modern Chinese. We believe that nesting is one of the characteristics of Chinese, and this kind of nesting also reflects the complexity of syntax to a certain extent.

Therefore, in order to better carry out the study, we will also include the form of complex sentences and the structural relationship of phrases involved in a single sentence. When it comes to specific classification, we still use the definition of the relationship between complex sentences in modern Chinese. We can see that with the increasing of the syntactic complexity, the number of inter-sentence relationship types also goes on.

4) *Number of the special phrases:* prepositional phrase and pivotal phrase. Complex sentences are usually longer, structurally intense, and impose a higher cognitive burden on the reader, so we developed phrasal features that measure structural complexity, including the proportions of noun phrases, prepositional phrases. Generally speaking, the two features also go on with the increase of syntactic difficulty, but the upward trend is not obvious.

5）*Special parts of speech:* Nouns and verbs are two important notional words. Their density will affect the reader's cognition to a certain extent. We calculate the densities of nouns and verbs in sentences and the ratio of noun verbs to verbs. From the result, we reach the same conclusion, that is to say, both of the two features are positively correlated with text difficulty.

III. EXPERIMENT AND RESULT ANALYSIS

A. Feature Selection

In this paper, we chose random forest to make feature selection. Because of the inherent randomness of stochastic forests, the model may give different importance weights for each feature. However, by training the model many times, that is, by selecting a certain number of features and reserving the intersection of the last feature each time, we can cycle a certain number of times, so that we can finally get a certain number of features that have an important contribution to the impact of classification tasks.

We use random forests to rank the importance of the features mentioned in the previous section, and the results are shown in Figure Ⅱ. 0-9 in abscissa is average sentence length, average punctuation marks, max_depth of the pare tree, the length of modifier, the ratio of length of modifier to sentence length, inter-sentence relationship types, number of pivotal phrases, number of prepositional phrases, the densities of nouns and verbs.

As can be seen from the figure below, the top five features are average sentence length, the length of modifier, inter-sentence relationship types, the ratio of length of modifier to sentence length, and the density of verbs.

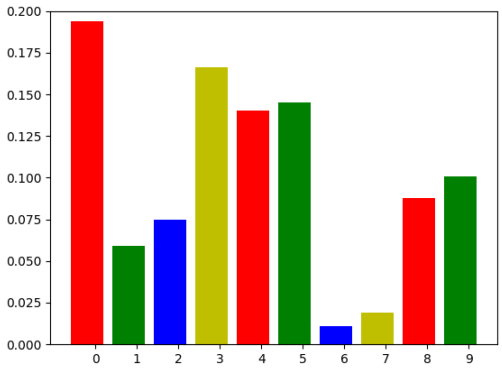

Figure Ⅱ. The importance of features

B. *Process and result of the experiment*

800 sentences are divided into training set and testing set according to 7:3 ratios. In order to ensure the accuracy of training and avoid over-fitting, we use 5-fold cross validation within the training set beforehand. Subsequently, we use Support Vector Machine (SVM) as the classifier. adding the top five features in the experiment process successively. The best experiment results are shown in Table I.

Table I: Result of the feature experiment

	P(%)	R(%)	F1(%)
F0	32.19	33.75	32.95
F0+F3	39.58	39.58	39.58
F0+F3+F5	44.16	43.10	43.62
F0+F3+F5+F4	45.25	45.17	45.21
F0+F3+F5+F4+F9	50.42	50.40	50.41

As Table I shown, five features that have been added successively make the precision rate, recall rate and F-score increase gradually. This illustrates that five features have different degrees of influence on classification of Chinese syntactic complexity.

We plotted the confusion matrix of the sixth classification experiment, as shown in Figure Ⅲ. It can be seen that the classification of level5 level1 and level4 reached the best results.

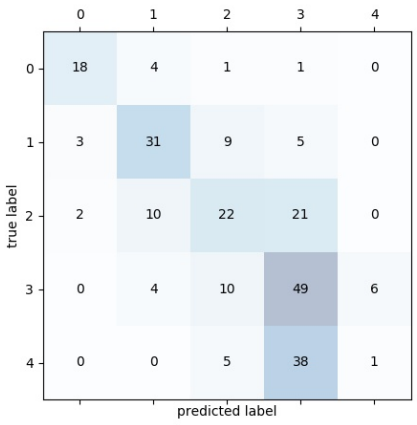

Figure Ⅲ. the confusion matrix of the experiment

IV. Conclusion

In this paper, we take Chinese sentences as the research object and study the Chinese syntactic complexity. Firstly, we chose so many features affecting sentence complexity. Then, we propose to use random forest algorithm to extract features. Finally, SVM classification algorithm is employed to verify the validity of the above features to the syntactic complexity. The precision rate, recall rate and F-scored for the classification of 5 levels can reach over 50 percent, which indicates that the features we selected has practical value for the related research.

In our further study, we will expand the scale of corpus and explore more effective features in order to better explore the factors affecting the syntactic complexity of Chinese.

References

[1] Laughlin G H M. SMOG Grading-A new readability formula[J]. Journal of Reading,1969,12(8):639-646.

[2] Dechant, E.V.& Smith, H. P. Psychology in Teaching Reading[M]New Jersey: Prentice-Hall, InC.,1961:134-150.

[3] Hunt, K. Grammatical structures written at three grade levels[J]. Elementary Secondary Education. 1965.

[4] Norris J M, Ortega L. Towards an Organic Approach to Investigating CAF in Instructed SLA: The Case of Complexity[J]. Applied Lin-guistics,2009, 30(4):555-578.

[5] Lu X. Automatic analysis of syntactic complexity in second language writing[J]. International Journal of Corpus Linguistics,2010,15(4):474-496.

[6] Caylor John S, Methodologies for determining reading requirements of military occupational specialities [J]. Adult Literacy,1973:81

[7] Kincaid J P,Fishburn R P,Chisson B S.Derivation of new readability formulas for navy enlisted personnel[J].Adult Basic Education,1975:49.

[8] Cortes C,Vapnik V. Support-vector networks[J].Machine Learning,1995,20(3):273-297.

[9] Petersen S E,Ostendorf M.A machine learning approach to reading level assessment[J].Computer Speech & Language,2009,23(1):89-106.

[10] Alusio S,et al. Readability assessment for text simplification[C]// NAACL Hlt 2010 15th Workshop on Innovative Use NLP for Building Educational Applications. Association for Computational Linguistics,2010:1-9.

[11] Schwarm S E, Ostendorf M. Reading level assessment using support vector machines and statistical language model[C]// Meeting on Association for Computational Linguistics. Association for Computational Linguistics,2005:523-530.

[12] Vogel M, Washburne C. An objective method of deterring grade placement of children's reading material[J]. Elementary School Journal，1928，28(5)：373-381.

[13] Ortega, L. Syntactic Complexity Measures and their Relationship to L2 Proficiency: A Research Synthesis

of College-level L2 Writing[J]. Applied Linguistics, 2003, 24(4):492-518.

[14] Bachman, L.F. Fundamental considerations in language testing. Oxford: Oxford University Press. 1990.

[15] Wolfe-Quintero,K.,S. Inagaki & H.Kim. Second Language Development in Writing: Measures of Fluency, Accuracy and Complexity [M]. Honolulu：University of Hawaii, Second Language Teaching and Curriculum Center, 1998.

[16] Betts， E．A．Difficulty：Its Application to the Elementary School [J]． Journal of Educational Research, 1949(42)438—59．

[17] Hunt, K. Grammatical structures written at three grade levels[J]. Elementary Secondary Education. 1965.

[18] Dechant, E.V.& Smith, H. P. Psychology in Teaching Reading[M]New Jersey: Prentice-Hall, InC.,1961:134-150.

[19] Lu X. Automatic analysis of syntactic complexity in second language writing[J]. International Journal of Corpus Linguistics,2010,15(4):474-496

[20] Norris J M, Ortega L. Towards an Organic Approach to Investigating CAF in Instructed SLA: The Case of Complexity[J]. Applied Linguistics,2009, 30(4):555-578.

[21] Biber D, Gray B, Poonpon K. Should We Use Characteristics of Conversation to Measure Grammatical Complexity in L2 Writing Development[J]. Tesol Quarterly, 2012, 45(1):5-35.

[22] Dominique Bruna et al. Is this Sentence Difficult? Do you Agree? [C]//Proceedings of Conference on Empirical Methods in Natural Language Processing, 2018,2690-2699.

[23] Jin, H G. Syntactic maturity in second language writings: A case of Chinese as a foreign language(CFL)[J]. Journal of the Chinese Language Teachers Association, 2007, 42(1):27-54. 2007.

[24] Jiang W, Measurements of Development in L2 Written Production: The Case of L2 Chinese[J]. Applied Linguistics, 2012, 34(1):1-24.

[25] McNamara, D. S. et al. Automated Evaluation of Text Discourse with Coh-Metrix[M]. Cambridge: Campbridge University Press,2014

[26] Sheehan K M, Kostin I, Napolitano D, et al. The TextEvaluator tool: Helping teachers and test developers select texts for user in instruction and assessment[J]. Elementary School Journal,2014,115(2):184-209.