The development of Chinese grammar checker website based on Natural Language Processing

J Wangi[1*], K Rosalin[1], Theresia[1]

[1]Chinese Department, Faculty of Humanities, Bina Nusantara University, Jl. Kemanggisan Ilir III No. 45, Jakarta 11480, Indonesia

*julita.wangi@binus.ac.id

Abstract. Research on artificial intelligence in learning Mandarin in Indonesia is still very rare. The research on developing a Chinese grammar checking website is aimed at helping Mandarin learners in mastering Mandarin grammar. This website can be a "learning companion" of basic-level Mandarin learners outside the classroom. This technology is still very new, so this research should be carried out immediately by researchers in Indonesia to master this technology. The method used in this research is a quantitative approach. The collection of grammar patterns that will appear on the chineseteak.net will be selected from several Chinese textbooks. Then the pattern will become Mandarin grammar patterns and will be entered the website auto-correct grammar. This website is designed using PHP and MySQL as the database. From the results of the questionnaire, we can find out that there are 7 points of Mandarin grammar that are still difficult for early learners.

1. Introduction

Today's Mandarin learning requires technology support that can be used anywhere and anytime [1]. Mandarin language learning that is considered difficult is learning grammar [2]. Grammar is a rule for the formation of a language, to master Mandarin [3], the error that commonly found in grammatical errors are addition, omission, ordering and selection [4].

In Indonesia, there is currently no website or application that can help elementary Chinese learners in checking Mandarin grammar. In previous studies, researchers have developed a website that can help students in checking Chinese grammar, the website chineseteak.net. However, the developed chineseteak.net website is still in the initial stages of development, after the chineseteak.net website was tested, researchers found that this website still has many shortcomings and is not yet suitable for use by Mandarin learners [5]. Therefore, in this advanced research stage, the researcher wants to improve the website chineseteak.net to be more user-friendly, the purpose of improving this website is also to be able to help the accuracy of the grammar patterns inputted by users of this website. That way this website can help Mandarin learners and instructors in learning and teaching activities so that it can facilitate Mandarin students in mastering Mandarin grammar.

2. Research method

The method used in this research is a qualitative approach. The website chineseteak.net is based on PHP with the Laravel framework and the guzzle library. The function of the guzzle is to scrapping data from the website http://nlp.standford.edu/software/localhost:8080.

The researcher collected this research data by collecting basic Mandarin grammar patterns that exist in the textbooks used in compulsory subjects namely Chinese Language I, which uses the book Er Ya Basic Chinese: Comprehensive Course (I) and Basic Chinese: Comprehensive Course (II), from chapter 9 to chapter 27. From these patterns, practice questions will be prepared grammar problems done by 54 the 2nd semester Mandarin Literature students, and from that, the grammar will choose the wrongest.

After that, the Chinese grammar pattern will be entered the website server chineseteak.net, by taking tagging each Chinese grammar pattern using the website http://nlp.stanford.edu:8080/parser/index.jsp. After taking the tagging of each Chinese grammar pattern, it is then entered the server.

On the website chineseteak.net, users can see the 7 most frequently wrong grammar, in which there will be an explanation of how to use the grammar. Users can also input the sentence they want to check, then click the "Check" button. Sentences will be sent to be processed in the controller (logic pattern matching) in Unicode form (non-Latin). In the controller, two methods are used, namely the library guzzle for scrapping data on the NLP, and the matching pattern of tagging that has been entered the server. Finally, the results of the check will be sent back to the website chineseteak.net. The flow chart of research method as shown in Fig 1.
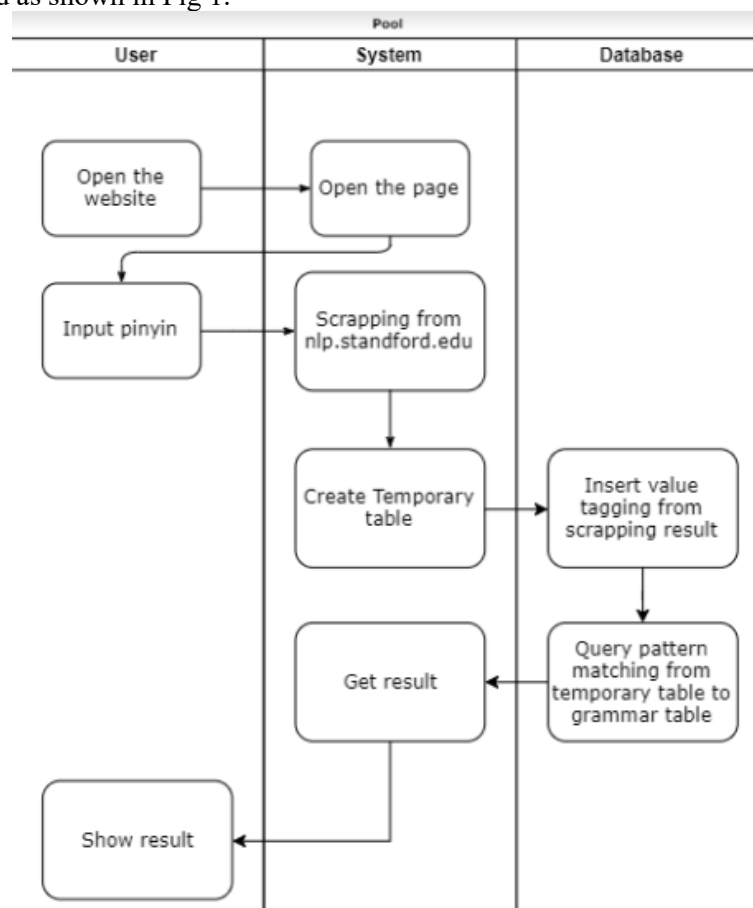

Figure 1. Research method's Flow Chart

3. Finding and discussion

3.1. The grammar patterns

In general, the target of research outcomes conducted during 2019 is a grammar checker website that is helpful for elementary Chinese learners. The grammar pattern that will be checked, is taken from the results of the questionnaire include 61 questions to the 54 2nd semester Chinese Literature students as shown in Fig 2, there are 7 Chinese grammar that is most wrong:
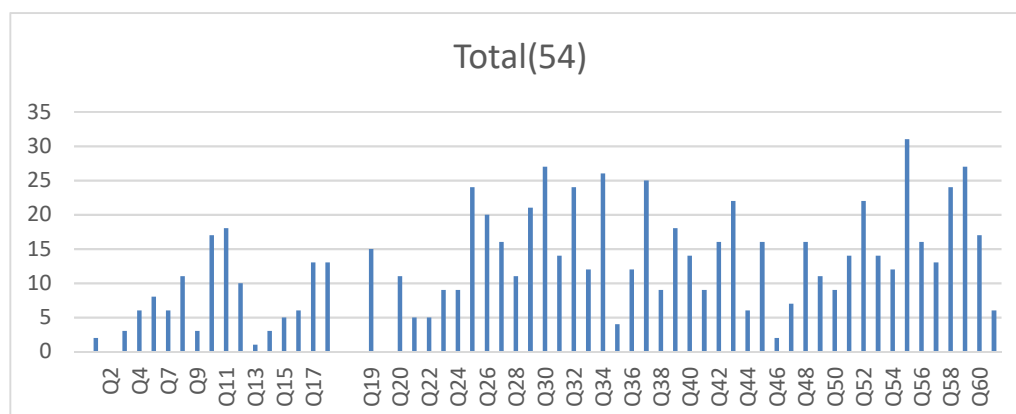
Figure 2. Bar chart from questionnaire

Evident in the bar chart here the most wrong grammar is Q55 (31 students), then Q30 (27 students), Q59 (27 students), Q34 (26 students), Q37 (25 students), Q32 (24 students) and Q58 (24 students)

3.1.1. *Complement of duration (时量补语)*
Complementary words used to express the duration of an action or status (状态).

3.1.2. *Complement of direction (趋向补语)*
After the predicate verbs, words, and phrases that show the action trend added, they are called complementary directions.

3.1.3. *To be doing something with "着" (V+着)*
The Continuation of State or the Progression of an action.

3.1.4. *Adverbials and the structural particle "地"*
The structural particle "地" is often used after an adjective as the indicator of an adverbial.

3.1.5. *The "Ba" sentence ("把"字句)*
The 把 (bǎ) sentence is a useful structure for constructing longer sentences that focus on the result or influence of action. One of its key features is that it brings the object of the verb closer to the front of the sentence and precedes it with a "把".

3.1.6. *The passive voice – The "Bei" sentence ("被"字句)*
The "Bei" Sentence often emphasizes the agent of an action, indicating disposal. It is often used to imply unhappiness of dissatisfaction.

3.1.7. *The Existential Sentence (存现的表达——存现句)*
An existential sentence denotes the existence of something or somebody. The object of an existential sentence usually refers to something or somebody indefinite.

3.2. Analysis and design
For developed the website, we have some requirements that must be fulfilled.

3.2.1. System specification

Hardware system specifications in the development and use of web chinesegrammarchecker.com there are two specifications that must be considered:

Operating system   : Windows 10 Education
Processor   : Intel(R) Core (TM) i5-2450M CPU @ 2.50GHz
Memory   : 8,00 GB
VGA   : NVIDIA GeForce GT 630M
Software   : Web Browser

There are also server requests that must be fulfilled, it is strongly recommended to use the Laravel Homestead virtual machine as a Laravel development environment. But if you don't use Homestead, there will be several server requests that must be fulfilled:

PHP >= 7.1.3
OpenSSL PHP Extension
PDO PHP Extension
Mbstring PHP Extension
Tokenizer PHP Extension
XML PHP Extension
Ctype PHP Extension
JSON PHP Extension
BCMath PHP Extension

3.2.2. System implementation

This research is aimed at developing a user-friendly Chinese grammar checking website for elementary Mandarin learners. Chinese grammar checking website developed with the domain name chineseteak.net is a website that can be used by learners to check sentences made whether they are in accordance with Chinese grammar patterns.

This website uses the PHP programming language with the Laravel Framework and the Guzzle library for scraping data from the web nlp.standford.edu/software/localhost:8080. Users open the website chineseteak.net, then the display will appear as shown in Fig 3, Fig 4, Fig 5, and Fig 6.

Figure 3. This is the homepage of chineseteak.net   **Figure 4.** Grammar page of chineseteak.net

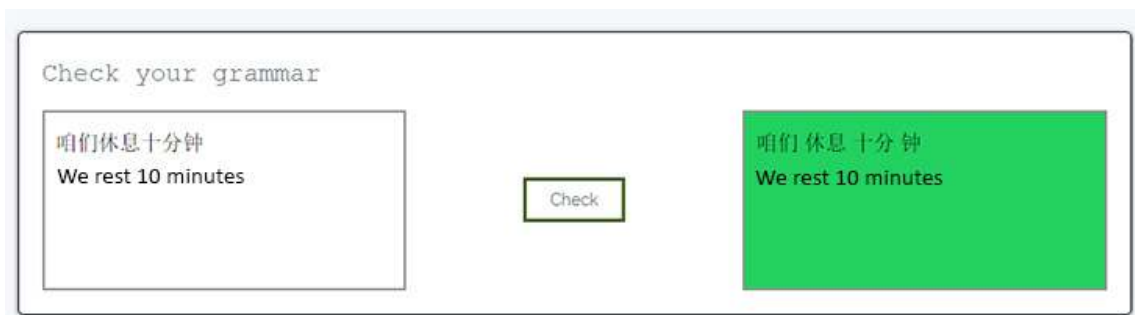**Figure 5.** Complement of duration page

Figure 6. Checking the sentence, example from right sentence

After logging into the page, the user can select the grammar to check, then type the sentence in Chinese and click "Check", to check whether the Chinese grammar is correct or not, as shown in Fig 7.
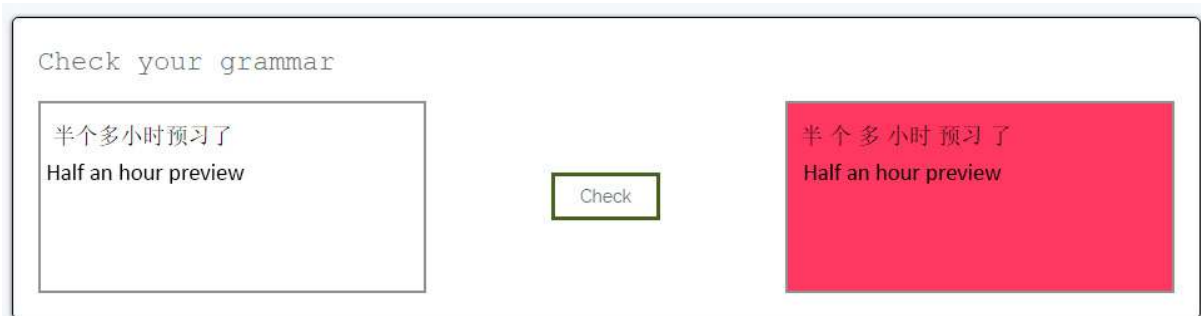

Figure 7. Wrong sentence display

In the display grammar checking results will display the results of checking and tagging per word in sentences. If the Chinese grammar pattern on the sentence entered the checkbox is correct, the results of checking the sentence will be green. If the Chinese grammar pattern on the sentence entered the checkbox is wrong, the results of checking the sentence will be red.

4. Conclusion

From the results of the questionnaire, we can find out that there are 7 points of Mandarin grammar that are still difficult for early learners. That is: the complement of duration, the complement of direction, to be doing something with "着", adverbials and the structural particle "地", the "把" sentence, the passive voice and the existential sentence.

During the process of developing this website, several things became difficult, the first being dependence on standalone NLP tagging, which was not specifically made for Mandarin. As a result, the sentences made must be following what has been entered into the server, the impression is too stiff and less agile.

For example, tagging from the sentence "他们休息十分钟" is 他们/PN 休息/VV 十分/AD 钟/NN, if we change the subject to "玛丽休息十分钟", the tagging will change to 玛丽/NR 休息/VV 十分/AD 钟/NN, even though pattern grammar is not change. This website checking the sentence by tagging not by syntactical functions (词性).

We make this research to try the water of the Chinese grammar checker website, we hope that in future there is a much more Chinese word library system that can make the website more effective. We hope that this website can assist elementary Mandarin learners in mastering these 7 Mandarin grammars and can facilitate Mandarin language teachers to check sentences in learning and teaching activities. In addition, the website will also be one tool that can help self-learning Mandarin learners.

In the next research, we will develop chineseteak.net, not only can check 7 Mandarin grammars, but we will check another Mandarin grammar, and will provide some interactive Mandarin grammar exercises in chineseteak.net.

Acknowledgments

References

[1]   Ying Y, Lin X, Go Y and Mursitama T N 2017 Mobile learning based of Mandarin for college students: A case study of international department sophomores *11th International Conference on Information & Communication Technology and System (ICTS)*

[2]   Ying Y, Mursitama T N and Novianti N 2018 Sustability of Textbook for the Improvement of Linguistic Competence in Chinese by International Relations Students in Indonesia *Pertanika J Soc Sci & Hum 26 (3)* **1241-1252**

[3]   Go Y 2014 Error Analysis of Chinese Word Order *Humaniora* **1140–1145**

[4]   Rosalin K 2014 A Brief Analysis on Error in Indonesian Beginner Level Students Chinese Composition *Humaniora* **960–967**

[5]   Wangi J, Rosalin K and Theresia 2018 Chinese Grammar Auto-Correct Website chinesegrammarchecker.com Development Model *Business Economic, Communication, and Social Sciences (BECOSS) Journal* **137-146**

[6]   Bhirud N S, Bhavsar R P and Pawar B V 2017 Grammar Checkers for Natural Languages: A Review *International Journal on Natural Language Computing* **1-13** https://doi.org/10.5121/ijnlc.2017.6401

[7]   Hiemstra R 2006 Self-Directed Learning *The International Encyclopedia of Education* **1–11**

[8]   Liu Y, Han Y, Zhuo L and Zan H 2016 Automatic Grammatical Error Detection for Chinese based on Conditional Random Field *Proceedings of the 3rd Workshop on Natural Language*

Processing Techniques for Educational Apllications **57–62**

[9]   Rohman A 2014 Mengenal Framework " Laravel " (Best PHP Frameworks For 2014)

[10]   Sutami H 2007 Kekhasan Pengajaran Bahasa Mandarin di Indonesia *Wacana Jurnal Ilmu Pengetahuan Budaya* **222–237**

[11]   Otwell T 2013 Laravel: The PHP framework for web artisans

[12]   Wong K, Li W, Xu R and Zhang Z 2010 Introduction to Chinese Natural Language Processing. *Computational Linguistics* **778–780**

[13]   Xiang Y, Wang X, Han W and Hong Q 2015 Chinese Grammatical Error Diagnosis Using Ensemble Learning *Proceedings of The 2nd Workshop on Natural Language Processing Techniques for Educational Applications* **99–104**

[14]   Yu L, Lee L and Chang L 2014 Overview of Grammatical Error Diagnosis for Learning Chinese as a Foreign Language *Proceedings of the 22nd International Conference on Computers in Education* **43–47**

[15]   Yudhanto Y and Prasetyo H A 2018 Panduan Mudah Belajar Framework Laravel (Jakarta: PT Elex Media Komputindo)