Automatic Difficulty Assessment for Chinese Texts

John Lee, Meichun Liu, Chun Yin Lam, Tak On Lau, Bing Li, Keying Li
Department of Linguistics and Translation
City University of Hong Kong
jsylee@cityu.edu.hk, meichliu@cityu.edu.hk, mickey1224@gmail.com,
tolau2@cityu.edu.hk, bli232-c@my.cityu.edu.hk, keyingli3-c@my.cityu.edu.hk

Abstract

We present a web-based interface that automatically assesses reading difficulty of Chinese texts. The system performs word segmentation, part-of-speech tagging and dependency parsing on the input text, and then determines the difficulty levels of the vocabulary items and grammatical constructions in the text. Furthermore, the system highlights the words and phrases that must be simplified or re-written in order to conform to the user-specified target difficulty level. Evaluation results show that the system accurately identifies the vocabulary level of 89.9% of the words, and detects grammar points at 0.79 precision and 0.83 recall.

1 Introduction

Reading is critical to foreign language acquisition (Krashen, 2005). While language textbooks provide a convenient source of reading materials, these materials are limited in quantity and do not always match the language learners' interest. To supplement textbooks, teachers often utilize texts from other sources, such as newspapers, magazines and the web. Since they were not originally written for pedagogical purposes, these texts typically require adjustments: teachers must simplify or re-write difficult vocabulary items and grammatical constructions so that the text becomes "comprehensible input" (Krashen and Mason, 2015) to the learners; conversely, teachers might desire more advanced language usage to challenge the learners. This editing process can be time consuming and labor intensive.

To assist the editor, we built a web-based interface that automatically determines the difficulty level of Chinese texts. It detects vocabulary items and grammar points covered by the *Hanyu Shuiping Kaoshi* (HSK) guidelines, the official curriculum for Chinese as a foreign language (CFL) in mainland China. Furthermore, the editor can specify a target difficulty level, and ask the interface to highlight all words and grammatical constructions that must be simplified or re-written to reach the target level.

To the best of our knowledge, this is the first system that assists editors of CFL pedagogical material by explicitly pinpointing the words and grammatical constructions that exceed the target difficulty level in an official curriculum.

2 Previous Work

Most text difficulty assessment systems aim at native speakers, both for Chinese (Chen et al., 2013; Sung et al., 2015) and for other languages (Pitler and Nenkova, 2008; Sato et al., 2008). Among those that target language learners, most give a holistic score on the overall difficulty level of the text (François and Fairon, 2012; Pilán et al., 2014), but do not specifically indicate the difficult words or grammatical constructions. Hence, while these systems can help identify suitable reading material for language learners (Brown and Eskenazi, 2004), they are not designed to facilitate editing of language teaching materials, which is the goal of our system.

Targeting learners of English as a foreign language, *FLAIR* (Chinkina et al., 2016) can detect 87 linguistic forms in the official English curriculum in a German state. The system attains an average precision and recall of 0.94 and 0.90 in detecting grammar points. Most systems for CFL determine the difficulty level of a text on the basis of vocabulary difficulty alone. ChineseTA (Chu, 2005), for example, estimate vocabulary difficulty on the basis of word frequencies interpolated from var-

The Companion Volume of the IJCNLP 2017 Proceedings: System Demonstrations, pages 45–48,
Taipei, Taiwan, November 27 – December 1, 2017.

Sentence	据说， *jushuo* 'reportedly'	齐白石 *qibaishi* 'Qibaishi'	一开始 *yikaishi* 'at first'	画 *hua* 'paint'	的 *de* DE	虾 *xia* 'shrimp'	太 *tai* too	重 *zhong* 'emphasize'	写真 *xiezhen* 'realism'
	"It is said that realism was overly emphasized in the shrimps painted by Qibaishi in early times."								
Vocabulary	5	NR	6+	3	1	6+	1	5	6+
Grammar	5 Parenthetical expression	3 Relative clause with subject and predicate			-	-	1 Adverb of degree	1 Verbal predicate	

Table 1: Vocabulary and grammar difficulty level of an example sentence, according to the HSK scale. "NR" refers to a proper noun; 6+ is the vocabulary level attributed to words not found in the HSK vocabulary lists.

Lv	Vocab. items	Gram. points	Lv	Vocab. items	Gram. points
1	150	35	4	1200	38
2	150	58	5	2500	39
3	600	68	6	5000	28

Table 2: Number of vocabulary items and grammar points at each HSK level

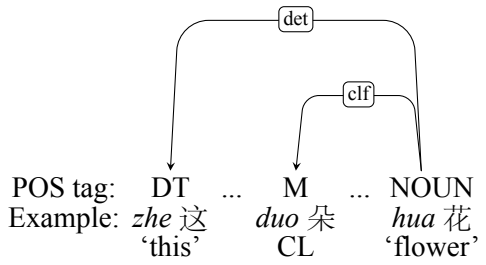

Figure 1: Parse tree pattern, in Stanford Dependencies for Chinese, for detecting the grammar point "Determiner and classifier"

ious corpora. The *Guidelines for CLT Materials Development* website (http://www.cltguides.com), the system that is most similar to ours, also concentrates on vocabulary assessment. It can detect a number of grammar constructions, but does not indicate their HSK levels or specific grammar points.

3 System Description

For Chinese as a foreign language, the two major assessment scales are the *Test of Chinese as a Foreign Language* (Zeng, 2014) and the *Hanyu Shuiping Kaoshi* (HSK) (Hanban, 2014). Both contain six levels and can be mapped to the Common European Framework of Reference for Languages, a global standard for measuring foreign language proficiency. Our system adopts HSK, the more widely used of the two in mainland China.

Upon input of any Chinese passage, the system performs word segmentation, POS tagging and dependency parsing using the Stanford Parser (Manning et al., 2014). It then offers difficulty assessment in terms of vocabulary and grammar (Section 3.1), and guides the user in editing the sentence towards the target difficulty level (Section 3.2).

3.1 Difficulty assessment

The HSK guidelines provide a vocabulary list and a set of grammar points for each level; as shown in Table 2), there are a total of 9,600 vocabulary items and 266 grammar points. For vocabulary assessment, the system matches each word with these lists, but does not assess the difficulty level of proper nouns, except those included in the HSK scheme. Table 1 shows an example sentence; the vocabulary difficulty levels of its word range from level 1 (e.g., *tai* 'too') to 6+ (e.g., *xiezhen* 'realism'); *Qibaishi*, a proper name, is not assigned any level.

For grammar assessment, we manually crafted parse tree patterns for the grammar points. A pattern may contain a combination of constraints in lexical, POS and dependency features. Figure 1 shows the pattern for the grammar point "Determiner and classifier" (指示代词和量词), requiring a noun to have two modifiers in the 'det' and 'clf' relations. The system performs dependency parsing on the input text, and then searches for matching parse tree patterns. In Table 1, the sentence exhibits four grammar points, the highest of which is the use of "Parenthetical expression" (*jushuo*, 'reportedly'), at level 5.

Most grammar points in the HSK guidelines provide concrete examples. The only exception is the grammar point for quadrasyllabic idiomatic expressions (成語), for which we use a list of about 1,000 expressions collected from Wiktionary. Further, three grammar points — semantic passive (意

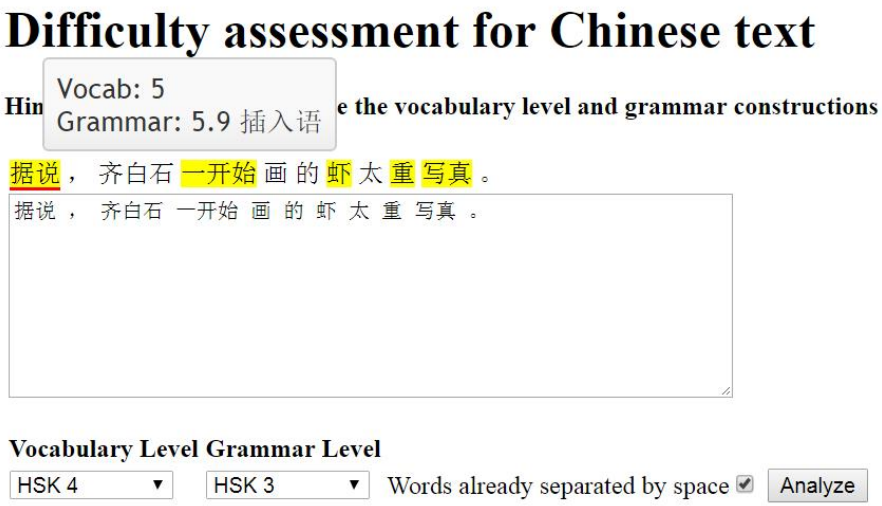

Figure 2: Screenshot of the system on input of the Chinese sentence in Table 1, with level 4 as the target vocabulary level, and level 3 as the target grammar level. The interface (i) highlights in yellow all words (*jushuo* 'reportedly', *yikaishi* 'at first', *xia* 'shrimp', *zhong* 'emphasize', and *xiezhen* 'realism') that exceed level 4; and (ii) underlines in red all grammatical points (*jushuo*) that exceed level 3.

义上的被动词), rhetorical questions with interrogative pronouns (用疑问代词的反问句) and directional complement (趋向补语) — require deeper semantic analysis, and thus have not been implemented.

3.2 Editing

If the user specifies the target vocabulary level and grammar level, the interface highlights in yellow all words that exceed the target level, and underlines in red all words participating in grammar structures that exceed the target level. For detailed information, the user can mouse over each word to view the vocabulary level detected, as well as the name of the grammatical structure detected (Figure 2). The user can edit the text accordingly, then re-submit the updated version for assessment, in an iterative manner until the text reaches the desired level of difficulty, or when the percentage of words exceeding the level falls below an accepted threshold, as shown by the distribution of statistics at the bottom of the page.

In case the system's word segmentation is inaccurate, the user may correct it and re-submit the text with the option "Words already separated by space", thereby asking the system to adopt the manual segmentation.

Level	# sentences	# words	# grammar points
1	18	105	69
2	51	407	296
3	52	639	403
4	60	1241	540
5	65	1211	577
6	85	1970	801

Table 3: Statistics of the evaluation dataset

4 Evaluation

In order to evaluate system performance, we harvested sentences from sample HSK exams from levels 1 to 6, obtained from the chinesetest.cn website. Our dataset contained a total of 331 sentences, including all sentences in the "Reading" sections of the examination papers for levels 1 to 4, and all sentences from reading comprehension exercises for levels 5 and 6. We performed manual word segmentation on these sentences, and annotated the HSK levels of each individual word and grammatical construction; Table 3 shows statistics of this dataset.

We evaluated system performance on both vocabulary and grammar assessment on this dataset; Table 4 presents the results according to HSK level. For vocabulary assessment, using automatic word segmentation, the system correctly recog-

Level	Vocabulary	Grammar	
	Accuracy	Precision	Recall
1	0.810	0.747	0.812
2	0.958	0.962	0.865
3	0.890	0.960	0.896
4	0.895	0.649	0.778
5	0.898	0.739	0.842
6	0.891	0.670	0.777

Table 4: System accuracy on vocabulary assessment, and precision and recall on grammar point detection

nized overall 89.9% of words and their vocabulary level. Most errors are due to word segmentation errors during automatic parsing, or misrecognition of proper names.

The average precision and recall of grammar points are 0.788 and 0.828. The system performs best in categories involving lexical features with unambiguous POS, such as "Pronouns" (人称代词), and worse in categories that requires accurate dependency parsing, such as double object (双宾语). Errors in recall were mostly due to the non-exhaustive nature of the examples in the HSK guidelines. Precision is most challenging for grammar points that can be disambiguated only through semantic analysis, for example between the use of *hui* (会) to express ability vs. prediction.

5 Conclusions and future work

We have presented a web-based interface that automatically assesses the difficulty level of a Chinese text. The system indicates the vocabulary level and grammar level of specific words and grammatical structures according to the HSK scale, and highlight those that need to be simplified or re-written in order for the text to conform to the target level. We have also reported the performance of the system on vocabulary and grammar level assessment.

In future work, we plan to estimate the overall difficulty level of a sentence; to offer suggestions for lexical simplification; and to extend the scope to other linguistic features, beyond the HSK guidelines, that can help estimate the difficulty of a text.

References

Jonathan Brown and Maxine Eskenazi. 2004. Retrieval of authentic documents for reader-specific lexical practice. In *Proc. InSTIL/ICALL Symposium*. Venice, Italy.

Yu-Ta Chen, Yaw-Huei Chen, and Yu-Chih Cheng. 2013. Assessing chinese readability using term frequency and lexical chain. *Computational Linguistics and Chinese Language Processing* 18(2):1–18.

Maria Chinkina, Madeeswaran Kannan, and Detmar Meurers. 2016. Online Information Retrieval for Language Learning. In *Proc. ACL System Demonstrations*.

Chengzhi Chu. 2005. ChineseTA 1.1. Beijing Language and Culture University Press, Beijing, China.

Thomas François and Cédric Fairon. 2012. An "AI Readability" Formula for French as a Foreign Language. In *Proc. EMNLP-CONLL*.

Hanban. 2014. *International Curriculum for Chinese Language and Education*. Beijing Language and Culture University Press, Beijing, China.

S. Krashen. 2005. Free voluntary reading: New research, applications, and controversies. In G. Poedisoedarmo, editor, *Innovative approaches to reading and writing instruction, Anthology Series 46*. SEAMEO Regional Language Centre, Singapore, pages 1–9.

S. Krashen and B. M. Mason. 2015. Can Second Language Acquirers Reach High Levels of Proficiency through Self-selected Reading? *International Journal of Foreign Language Teaching* 10(2):10–19.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP Natural Language Processing Toolkit. In *Proc. ACL System Demonstrations*. pages 55–60.

Ildikó Pilán, Elena Volodina, and Richard Johansson. 2014. Rule-based and Machine Learning Approaches for Second Language Sentence-level Readability. In *Proc. 9th Workshop on Innovative Use of NLP for Building Educational Applications*.

Emily Pitler and Ani Nenkova. 2008. Revisiting readability: a unified framework for predicting text quality. In *Proc. EMNLP*.

Satoshi Sato, Suguru Matsuyoshi, and Yohsuke Kondoh. 2008. Automatic Assessment of Japanese Text Readability based on a Textbook Corpus. In *Proc. LREC*.

Yao-Ting Sung, Ju-Ling Chen, Ji-Her Cha, Hou-Chiang Tseng, Tao-Hsing Chang, and Kuo-En Chang. 2015. Constructing and validating readability models: the method of integrating multilevel linguistic features with machine learning. *Behavior Research Methods* 47:340–354.

Wenxuan Zeng. 2014. Huayu baqianci ciliang fenji yanjiu 华语八千词词汇分级研究 (Classification on Chinese 8 000 Vocabulary). *Huayu xuekan* 华语学刊 6:22–33.