S. L. Wong et al. (Eds.) (2010). Proceedings of the 18th International Conference on Computers in Education. Putrajaya, Malaysia: Asia-Pacific Society for Computers in Education.

A Statistical Approach on Automatic Passage Level Checking Framework for English Learner

Wasan NA CHAI, Taneth RUANGRAJITPAKORN, Thepchai SUPNITHI
Human Language Technology Laboratory, National Electronic and Computer Technology Center, Thailand Science Park, Pathumthani, 12120, Thailand
{wasan.na_chai, taneth.rua, thepchai.sup}@nectec.or.th

Abstract: In this paper, we develop a preliminary research on passage grading system. We propose an approach to examine an English reading passage that meets students' ability and level. CRF has been applied to create a level characteristic model from passage corpus. The system calculates by using three features; word, syllable and sentence complexity. The system does not require manual criteria to grade a passage, but passages are automatically graded by comparing their scores to a model. The output of the system shows a level of passage based on Thai academic school level.

Keywords: grading system, readability, English passage, characteristic model, CRF

Introduction

In ICT community, a large amount of interesting documents are provided in many locations, such as Wikipedia, knowledge sharing, and social network website. However, appropriate reading passages in English class are normally assigned by teacher's decision alone. From the reason, students do not gain their motivation and lose their eagerness to read those passages since the passages apparently do not meet their interest especially for non-native English learners such as Thai. To allow student to use their own passages, teacher otherwise loads more burden to approve an appropriation on those passage level to suit students' ability. Improving reading motivation is another difficult issue for Thai learner. This paper focuses on the questions "How to match between reading passage and student level if students want to select their own reading passage?"

Readability level checking is one of the most important issues in ESL and EFL. Many tools were implemented focusing on this topic such as Kincaid formula, SMOG-grading, Fox index and Flesch reading easy formula. The Flesch reading easy formula [1] was developed by Flesch in 1948 and it is based on school text covering grade 3 to 12. Unfortunately, it has not been updated for a decade. The Kincaid Formula [2] has been developed for grading Navy training manuals. It is accountable in technical document grading because it is based on adult training manuals rather than school book text. The SMOG-Grading [3] is a tool for grading English texts. It has been developed by McLaughlin in 1969. Its result is a school grade. The Fog index [4] has been developed by Robert Gunning. It especially concerns a proper name issue and handles it separately. All of those systems compute a readability score based on syllable, word, and sentence amount and their scores are graded by manually constructed criteria. The purposed system also use those three features but it differs from them in terms of we build a level model automatically with examples of reading passage provided in corpus. Therefore, we do not need to construct a criterion for grading manually.

1. System architecture

In this system, there are two main processes; training and testing process. For determination of passage level, three features of English information are focused in this work; syllable, word, and sentence complexity score. System overview is shown in Figure 1.

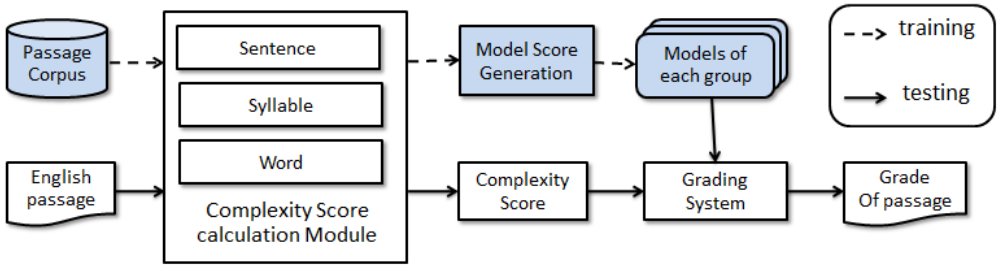

Figure 1. System overview of training characteristic model

For training, English passage corpus is separated by level regarding to Thai academic system (grade 1-12). Pre-processes, which are sentence segmentation and word segmentation, are required to handle data. Sentence list is computed for a sentence feature in sentence complexity module. Word list is calculated for a word and syllable feature in vocabulary and syllable complexity module respectively. A supervised learning method, called conditional random field (CRF) [5], is exploited to produce a characteristic model of each passage level with above mentioned features by (1). A result of training is characteristic models.

$$argmax\, P(Lev_i|Doc) = \lambda_1^i f_1 + \lambda_2^i f_2 + \lambda_3^i f_3 \quad (1)$$

where $P(Lev_i \mid Doc)$ the probability of document in each level is, λ_n^i is a feature constant generated from CRF and f_n is features score from the three sub modules.
Once characteristic models are obtained, they are used as a reference set for grading an unknown levelled reading passage. The system apparently compares the scores calculated upon the same three features with models, and it results an appropriate level that the given passage belongs to.

1.1 Word Complexity Module

When word list is sent to this module, it is classified into two groups, content word and function word. Content words are words that have a stable lexical meaning, such as noun, verb, adjective. Function words are words that have little lexical meaning, but instead serve to express grammatical relationships with other words. In this module, content words are extracted into their lemma for checking their level. Lemma extraction using in this work is *morpha* [6][7] which is an open source tool. All words are matched to assign a level with reference word list collected from training corpus. The frequency of each word is also accumulated. Finally, word complexity score is calculated by (2).

$$WordComplexityScore = \frac{\sum_{n=1}^{N}\left[(Lv_{n_C} \cdot \beta_n \cdot n_n) + (Lv_{f_C} \cdot \beta_f \cdot n_f)\right]}{\sum_{n=1}^{N}(Lv_{W_C} \cdot W_i)} \quad (2)$$

where Lv refers to a level of a word in reference list, C indicates a content word, f is a function word, β is a parameter, n is a frequency, W_i is the frequency of i^{th} word.

1.2 Syllable Complexity Module

Syllable is a measure to point out a difficulty of a word. The more syllable a word has, the more difficult level it could be. All words are applied to (3).

$$SyllableComplexityScore = \frac{\sum_{t=1}^{W}(n_{syl_t})}{W} \quad (3)$$

where n_{syl_t} the number of syllable of word t^{th} and W is the total number of words in a passage.

1.3 Sentence Complexity module

In general, there are four main types of sentence: simple sentence, compound sentence, complex sentence and combination between compound and complex sentence. We calculate sentence types by (4).

$$SentenceComplexityScore = S \cdot X^{n_X} \cdot P^{n_P} \quad (4)$$

where S refers to a simple sentence, X indicates a complex sentence and P is a compound sentence. N_x is a number of a recursion of a complex sentence and N_p is a iterative number of compound sentence.

2. Conclusion and future work

We present a system framework to grade a level of English reading passage that student personally chooses by their own since the passage tends to enthuse student to read it excitingly and joyfully. In this work, three features, which are word complexity, syllable complexity and sentence complexity, are set to represent language phenomena. The system begins with training a level characteristic model from given English passage corpus using CRF. The trained model is used to compare with the test passage to grade it.

In the future, we plan to compare other supervised learning methods, such as neural network and expectation maximisation, to find the best supervised learning suitable to the system. Moreover, other significant features, such as proverb and idiom usage, domain specific vocabulary, are planned to include in the system.

References

[1] Flesch, R. (1948). "A New Readability Yardstick," Journal of Applied Psychology 32: 221-233.
[2] Luo, S. and Callan, J. (2001). A Statistical Model for Scientific Readability. In Proceedings of the tenth international conference on Information and knowledge management.
[3] McLaughlin, G. Harry (1969). "SMOG Grading — a New Readability Formula". Journal of Reading 12 (8): 639-646.
[4] Fuller, S., Horlen, C., Cisneros, R., and Merz, T. (2007). "Pharmacy Students' Reading Ability and the Readability of Required Reading Materials". American Journal of Pharmaceutical Education 2007; 71 (6) Article 111.
[5] Lafferty, J., McCallum, A., & Pereira, F. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In Proceeding of 18th International Conf. on Machine Learning.
[6] Minnen, G., Carroll, J. and Pearce, D. (2000). Robust, applied morphological generation. In Proceedings of the 1st International Natural Language Generation Conference, Mitzpe Ramon, Israel. 201-208
[7] Morphological and Orthographic Tools for English (morpha): available at http://www.informatics.sussex.ac.uk/research/groups/nlp/carroll/morph.html