Analyzing lecturer's speech and slide by estimated word difficulty

Yoshinobu Kito[a*], Kiyoaki NAKANISHI[a], Nobuyuki KOBAYASHI[b], Hiromitsu SHIINA[c] & Fumio KITAGAWA[c]
[a]*Graduate School of Informatics, Okayama University of Science, Japan*
[b]*Faculty of Human Sciences, Sanyo Gakuen University, Japan*
[c]*Faculty of Informatics, Okayama University of Science, Japan*
*sigel4280@live.com

Abstract: As research goal of providing the difficulty level of the Japanese used in lectures from the viewpoint of foreign exchange students and lecturers who are not native speakers of Japanese, we propose a system for providing Japanese word difficulty used in speech and slides of lecturers in lectures. The grades of the old version of Japanese Language Proficiency Test are used to determine the Japanese word difficulty. In the case that the Japanese word difficulty is already determined by materials, that Japanese word difficulty is used, and in the case of words with an unknown word difficulty, estimates reached by use of a support vector machine are used to arrive at a word difficulty. This system provides three kind of information of lecture by word difficulty.

Keywords: Lecture difficulty, Word difficulty, Support Vector Machine

Introduction

For foreign exchange students whose native language is not Japanese, it is not easy to attend lectures intended for Japanese people and they will have trouble comprehending such lectures unless they achieve a certain level of proficiency in Japanese. And yet, lecturers themselves do not understand the difficulty of speech and slides from the perspective of foreign exchange students. We are conceivable that it would be useful for both sides to be able to know the level of usage of Japanese words used in lectures. Students would be able to understand the difficulty of lectures and learn words that are necessary, while lecturers would be able to understand problems and ways to improve slides that they created and understand the nature of the spoken language used during the lecture with the insight gained from knowing the Japanese word difficulty used, and also update their slides and improve their lectures. In this research, we are developing a system that provides lecture difficulty on Japanese word difficulty using VOD lectures [1] which use an internet environment in which the speech and materials used in a lecture are saved as files.

1. System Summary

In this system, the data that is processed is subtitle data that is obtained from conversion of speech during the lecture into subtitles, as well as data that is extracted from the text portions of PPT files. The grade from the JLPT [2] which corresponds to each noun that is obtained from each of these sources is taken as the word difficulty and the Japanese word difficulty of a VOD lecture can thus be rated.

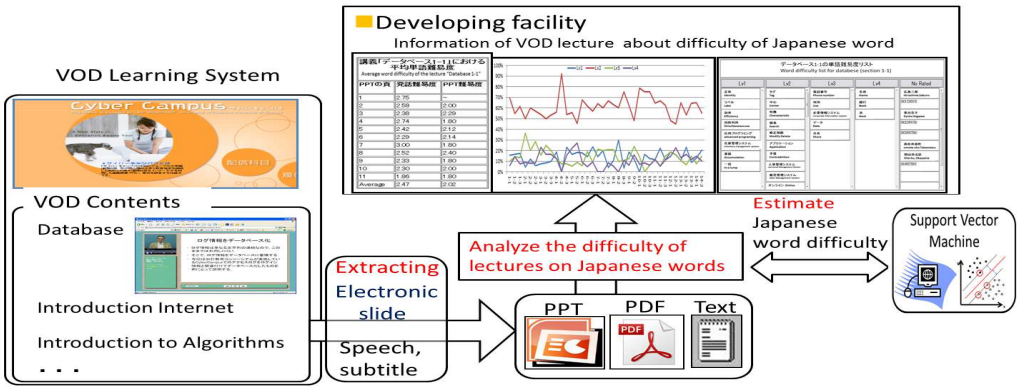

Figure 1 System Layout

2. Estimation of Japanese Word Difficulty

We use support vector machine (SVM)[3] to estimate Japanese word difficulty. It is a classification algorithm and the feature to assign each grade of JLPT [2] to classes.

2.1 Building Learning Parameters of SVM from Dictionary Data

The hypothesis surrounding the estimation of the grade of Japanese word difficulty using dictionary data is that with regards to the relation between entry words and their semantic descriptions, the grade of Japanese word difficulty used in semantic descriptions is equal to or easier than the entry word itself. Based on this hypothesis, we assumed a correlation between the grade of difficulty distribution according to the grade of the words expressed in the semantic description and the grade of difficulty for the entry word and generated the learning parameters. As evaluation experiment in this study, we used Tokuhiro's data [4] as the initial data, and Meikyo Japanese Dictionary [5] as entry words of dictionary data. In the case of the word “keiei (management)” in Figure 2, two Lv1 words, one Lv2 word, one Lv3 word and zero Lv4 words appear in the semantic description, and the word is trained using a combination of the learning parameter (1/4,3/4,0/4,0/4), and a Lv2 teaching signal.

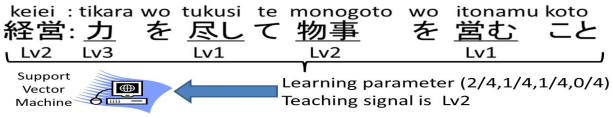

Figure 2 Building Learning Parameter

2.2 Difficulty Level Estimation of Compound Words using Web Search Data by SVM

There are many Japanese compound words that are not entried in dictionaries. Because of this we use descriptions obtained from web searches to estimate word difficulties. The method used for estimation is to enter the compound word into a web search, and then use the distribution of grade of word difficulty among the words near the compound word in the search results. In Figure 3, the learning parameter (1/2, 1/2, 0/2, 0/2) is created from the description of web search results, then SVM estimates the grade of difficulty is Lv2.

Figure 3 Estimation of Difficulty for Compound Words

3. Analysis of Lecture difficulty on Japanese Words

In this research, we are developing analyze system of lecture difficulty of Japanese word. It provides information related to Japanese word difficulty from speech and PPTs used in a lecture and the variation for each lecture. In particular, we provide the average word difficulty level for each lecture, word difficulty ratio distribution and word difficulty list.

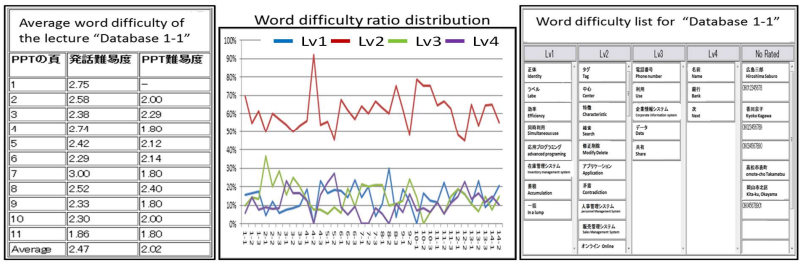

Figure 4 Information of Lecture Difficulty

The survey on the comprehensiveness of the word difficulty list in Table 1 shows the number and percentage of words in each grade that were judged to be Lv1 words. And it shows that almost all the difficult words are included up until Lv2, and also that the ratio of words judged to be more difficult by the Chinese foreign exchange student was low.

Table 1 Survey on Comprehensiveness of Word Difficulty Level List

Estimated grade of Word	Average for Japanese students	Chinese student
Lv1 (8)	3.333 (0.416)	2 (0.25)
Lv2 (36)	7.666 (0.213)	2 (0.053)
Lv3 (5)	0.333 (0.067)	0 (0.0)
Lv4(3)	0.000 (0.0)	0 (0.0)

4. Future works

This research proposed various types of information about word difficulty in order to help lecturers and lecture takers to understand the situation of lectures. The word difficulty of speech and PPT show a trend whereby the word difficulty used in PPT has a higher difficulty than those used in speech. Also, since lectures proceed by following a PPT, we think that a lecturer's speech is somewhat dependent on the contents of the PPT. We believe that this means that it can be useful for lecturers to be careful of the distribution of word difficulty levels in PPT and when displaying of words with high difficulty. As future works, we will consider developing of learning system that allows for preparatory study of the Japanese needed to understand a lecture. We also have plans to create a general Japanese difficulty level index that takes into account not only word difficulty level, but that also indexes the complexity of a sentence according to construction.

References

[1] Kitagawa, F., Onishi S. (2007). An Experiment on selective Course with Face-to-Face or/and E-learning, and Students Behavior (in Japanese), *Japan Society of Educational Information*, *22*(3), 57-66.

[2] Japanese Language Proficiency Test, http://www.jlpt.jp

[3] Vapnik,V. (1998). Statistical Learning Theory, *Springer*.

[4] Tokuhiro, Y. (2008), Kanji2100 Listed according to Frequency and Familiarity, *Sanseido*.(in Japanese)

[5] Kitahara,Y. (2010), Meikyo Japanese Dictionary Second Edition, *Taishukanshoten*. (in Japanese)