Character and Word Classification in "Web Based Japanese Language Acquisition Support System OSARU"

（日本語教育支援システム「おさる」における文字・語彙分類）

Bhooshan Raj Neupane, Edson T. Miyamoto, Yuji Matsumoto

Graduate School of Information Science, Nara Institute of Science and Technology

E-mail: {neupa-ra, etm, matsu} @is.aist-nara.ac.jp

Abstract

我々はこれまで、日本語教育支援システム「おさる」の開発を行ってきた。「おさる」の特徴の一つは学習者に合った問題を自動生成することである。本研究では、「学習者に合った問題」を学習者のレベルや興味、バックグラウンドを考慮して作成された問題と捉え、日本語の文字・語彙を難易度別・分野別に分類することにより、問題を自動生成する手法を提案した。この手法を用いて、新聞記事中の文字・語彙を分類し、「おさる」による学生に合った問題の自動生成が可能になった。

We have been developing a web based Japanese Language acquisition support system "OSARU". One of the objectives is to automatically create tailor made questions to Japanese language learners. For this, we have to classify the characters and vocabulary in terms of difficulty and field in which they are more frequent. In this paper, we propose the classification method utilizing the classification of Japanese Language Proficiency Test and the frequency of words in different fields.

1. Introduction

One of the objectives of OSARU is to realize an environment for effective teaching and learning by generating and providing language exercises that are tailor made for each learner. With this end, multiple-choice questions are created based on each learner's background, language acquisition level, and field of interest. Thus, OSARU requires a linguistic database in which each character and word is assigned a set of values. The values are labeled according to their difficulty, (*Difficulty Value)*, and according to the field where they are most frequent (*Field Value)*. In this paper we discuss how to assign *Difficulty Value* and *Field Value* to Japanese characters and words.

2. *Difficulty Value* and *Field Value*

The *Difficulty Value* provides a guideline for the order in which characters and words should be introduced as the learner's knowledge of the language progress. The *Field Value* ranks characters and words according to the field of interest of the learner.

3. Character Classification

3.1 Related Research

There are about 6000 kanjis (i.e., characters of Chinese origin) being used in current Japanese. They are classified in various groups according to their use and importance. One classification from the Ministry of Education is the Kyouikuyou kanji [1] consisting of 1006 characters which are taught in the six years of elementary education. For Japanese learners, a widely used classification is the one provided by the Japanese Language Proficiency Test (abbreviated JLPT) [2] which divides 2036 characters (of which 1926 characters are from Jouyou list [1] and 110

[1] The Jouyou list was prescribed by the Japanese

characters are of Dainisuijyunn[2] list) into four levels. Level 4 of JLPT contains 100 characters, level 3 contains 300 characters, level 2 contains 1000 characters and level 1 contains 2036 characters. There are various studies that propose grouping kanjis according to their importance. An example is [3] which selects the kanjis that are used in basic Japanese words. Other proposals utilize JLPT's listing for kanji classification, (e.g., [4]).

We decided not to implement JLPT's classification as it is because of its few levels. We wanted a finer classification with fewer kanjis in each level, so that learner's progress could be better assessed. We were not able to use other classifications due to the following reasons: low number of kanjis (Kyouikuyou list), no available list classifies all the Jouyou kanjis.

3.2 Studies and Results

We conducted an experiment in order to determine the relation between frequency of characters in daily use and their classification in JLPT. We chose the text data of the Mainichi (9 years, 1991-1999) and Nikkei (10 years, 1991-2000) newspapers to calculate the frequency of the characters used. We found that all the characters in JLPT level 4 were among the top 760 most frequent in the Mainichi list except one, 雨 (rain) which was the 928 th most frequent, and the top 800 most frequent in the Nikkei list except six 左 (left),右 (right), 雨, 父 (father), 母 (mother), 毎 (every). All characters in JLPT level 3 were in the top 1200 of both newspapers except eight, in the Mainichi (茶,昼,弟,冬,肉,飯,勉,妹) and (漢,兄, 犬,姉,昼,弟,勉,妹) in the Nikkei. These characters even though not listed on the top 1200 most frequent are very common characters and have high familiarity level. Hence in general, there is a high correlation between the JLPT classification and frequency in printed media.

3.3 Classification of Characters

In our system we have classified characters into 16 difficulty levels as shown in Table 1. Level 0 includes all the non-kanji characters and special phonemic characters (i.e., ヶ, 々, ヵ) that are in use. Level 9 contains all the kanjis in level 1 of JLPT. We have adopted the levels of JLPT as a general rule. However we have utilized the listings of Kyouikyou levels and the information on frequency of characters in newspapers to divide level 1 and 2 of JLPT into smaller groups. In levels 10 to 15, characters are divided according to their frequency. The 16 difficulty levels for the characters are set as the *Difficulty Value* to be used in our system.

4. Vocabulary Classification

4.1 Related Research

There are a number of basic vocabularies for foreigners available, however most of them are developed with specific purposes and the list varies according to purpose. A study conducted by the National Institute of Japanese language used the database of seven different vocabularies listings for foreigners found that there were only 278 words, which were common to all lists, even though the number of words in each list varied from 500 to 5000 [5].

JLPT list of four levels is one important vocabulary list utilized for educating foreigners, however, due to its small size (about 10,000 in level 1), few levels and no field listing, we are not able to adopt it. There are few studies on

government as the characters permitted to be used for written Japanese of daily usage.

[2] The kanji codes enacted by JIS X 0208 are divided into two major groups. They are Daiichisuijyun (第一水準) and Dainisuijyun (第二水準). Daiichisuijyun contains 3489 characters of high frequency including 62 phonetic kana characters, 62 alphanumeric characters and 293 symbols and codes. Dainisuijyun contains the less frequent 3390 characters including the characters used for names of places and persons and old characters.

vocabulary classification based on field, and a number of books and reference materials also contain the important words in some fields.

Level	#kanji	JLPT	Kyouikuyou	Description
0				hiragana, katakana, symbols and alphabets + (ヶ, 々、カ)
1	80	4		kanjis on JLPT level 4
2	165	3		kanjis on JLPT level 3
3	196	2	1,2,3	kanjis on JLPT level 2 and Kyouikuyou level 1, 2 and 3
4	291	2	4,5	kanjis on JLPT level 2 and Kyouikuyou level 4 and 5
5	268	2		remaining kanjis from JLPT level 2
6	178	1	2,3,4,5,6,	kanjis on JLPT level 1 and Kyouikuyou level 2,3,4,5 and 6
7	280	1		most frequent 280 of JLPT level 1(not on lower lists)
8	280	1		281st to 560th most frequent kanji from JLPT level 1
9	302	1		remaining kanjis from JLPT level 1
10	285			Jinmei kanjis
11	500			most frequent 500 kanjis from Non-Jouyou-Jinmei list
12	500			501st to 1000th most frequent kanjis from Non-Jouyou-Jinmei list
13	500			1001st to 1500th most frequent kanjis from Non-Jouyou-Jinmei list
14	500			1501st to 2000th most frequent kanjis from Non-Jouyou-Jinmei list
15	791			2001st to 2791st most frequent kanjis from Non-Jouyou-Jinmei list
Total of 5116 characters are listed on 15 levels according to difficulty				

Table 1.Difficulty level s adopted for Japanese characters.

Numerous basic vocabulary lists have been proposed, however there are few classifications of basic vocabulary in terms of difficulty. Basic vocabulary in any language can be classified according to frequency and to familiarity. The frequency classification approach is rather simple and involves counting the frequency of words in a large corpus and ordering the words according to their frequency [6]. By performing a psychological test the most familiar words can be determined [7].

4.2 Studies and Results

To classify the vocabulary according to field we performed a study on frequency of vocabulary in different fields. For our study, we used text data from 1991 to 1999 of the Mainichi newspaper. The Mainichi Newspaper is classified in 14 different fields (e.g., front page, international, economics, science and sports). Although in our system we plan to use a smaller number of fields, for initial evaluation we used all the fields as they are listed in the newspaper.

After extracting the text data from the corpus, we processed it using the morphological analyzer Chasen [8], whose output was then passed into a word counting program that calculated the frequency of each word in each field. As the frequency of each word depends upon the size of the text used, we calculated the relative frequency of each word in each field to assign a score that could be utilized in our system. We call this score *Rf* and define it as follows.

Rf = Frequency of each word / number of words

4.3 Classification of Vocabulary

In our system, we plan to utilize the difficulty of characters in order to classify vocabulary in terms of difficulty. We used the *Difficulty Value* of characters in each word and set *Difficulty Value* of the word as the *Difficulty Value* of its character with highest value. For example in Table 2, the word “回収“ has *Difficulty Value* 5 because the *Difficulty Value*

of the character “収” is 5.

Word	Kanji Level	Level	Field Value		
			Sports	Economics	Science
回収 (Recovery)	回:3 収:5	5	0.50	28.72	38.38
銀行 (Bank)	銀:3 行:1	3	1.37	202.38	2.28
運命 (Fate)	運:2 命:3	3	0.42	0.55	0.38

Table 2. Classification of vocabulary according to difficulty and field.

To classify the words according to *Field Value*, we used the Rf score of each word. The Rf score for each word is assigned as its *Field Value*. For example, in Table 2, the word “銀行” (bank) is more important in economics than the word “運命” (fate) although both of them belong to the same difficulty level. By assigning each vocabulary with *Difficulty Value* and a *Field Value*, we are able to select the important vocabulary in the field in which the learner is interested.

5. Conclusion

Utilizing the JLPT’s classification of kanji characters, Kyoikuyou kanji classification and the frequency of characters in print media, we were able to assign each character in Japanese writing a *Difficulty Value*. And utilizing the *Difficulty Value* and the frequency of characters in different fields, we were able to determine the most frequent words in each difficulty level and in various fields.

The user of the system selects the language level and the field of interest. While generating questions, the system selects the words from the users level that carry the highest frequency value for that field. Multiple-choice questions for the selected words are created. Highest priority is given to the sentences that are from the same field when preparing multiple-choice questions.

As usually seen in other classifications, the levels determined in our system does not follow a general concept of raising the number of characters per level as level increases. We adopted the JLPT classification as much as it was possible in our classification. In the case of classifying vocabulary according to the field, the fields we selected were based on the fields of the Mainichi newspaper. However, we need to collect big corpora from different fields and levels in order to efficiently evaluate the *Difficulty Value* and *Field value* of characters and words.

References

[1] 文部科学省、学習指導要領, 文部省告示第１７５号

[2] 日本語能力試験出題基準､国際交流基金、日本国際教育協会、凡人社、1999

[3] 田村真奈美、基本語彙のための漢字群、Mathematical Linguistics, Vol. 20 No.5, pp197-204, 1996

[4] Reading Tutor, Tokyo International University, Kawamura Yosiko, 2002

[5] 国立国語研究所､“日本語教育基本語彙七種比較対照表”日本語教育指導参考書９、大蔵省印刷局、1987

[6] 天野､近藤 “日本の語彙特性７頻度”、三省堂、2002

[7] 金杉友子、笠原要、稲子望、天野成昭、単語親密度に基づく基本的語彙の設定､自然言語処理 150-18、pp119-124, 2002

[8] 松本ら、形態素解析システム「茶筌」、2000