N-012

Inferring CEFR Reading Comprehension Index Based on Japanese Document Classification Method Including Pre-A1 Level

My Nguyen Tra Huynh† Yoshinori Miyazaki† Seiji Tani‡

1. INTRODUCTION

The Common European Framework of Reference for Languages (CEFR) [1] is an international standard for describing second language proficiency developed by the European Council. The framework includes four language skills with six stages of A1 to C2 level, and series of description statements (Can-Do Statements, hereafter called CDS) that are described for each level indicating what can be done. Moreover, each CDS can be used together with concrete example sentences for utilization. Hence, CEFR can serve to evaluate language proficiencies of learners in testing.

Regarding the utilization of CEFR for Japanese language education, Japanese CEFR-compliant text corpus has not been created, and only limited studies of CEFR for learners of Japanese have been conducted. Takada et al. [2] studied the semi-automatic classification of Japanese example sentences corresponding to reading comprehension indices (CDS) for the creation of Japanese CEFR-compliant text corpus using machine learning. In the study, “technicality”, “length”, and “document type” were chosen as features in CDS classification, but the derivation of “document type” was manually operated. Therefore, Hirakawa et al. [3] conducted a research to automatically estimate document types. In addition, [3] worked on the improvement of the accuracy of “technicality”, and carried out the experiment using these estimation results. [2] and [3] covered 27 CDSs excepting C1, C2 proficient levels and a CDS at B2 independent level which focuses on vocabulary ability rather than reading comprehension skills.

In 2017, the CEFR Companion Volume with New Descriptors was published to be intended as a complement to the CEFR [4]. The focus in the project was on updating the CEFR illustrative descriptors. Among the updates, this study will focus on Pre-A1 level, which supports novice learners of Japanese. We add new seven CDSs of Pre-A1 level in the Companion Volume with corresponding example sentences. To cope with the accuracy of classification of new CDSs, we extended to four features for inference, incorporating Kanji rate as one of the vital elements in reading comprehension. Further, instead of conventional 7 document types, we divided sentences into 8 types.

2. PRE-A1 LEVEL

The scale of CEFR is based on two levels (A1, A2) of “Basic language users” two levels (B1, B2) of “Independent language users” and two levels (C1, C2) of “Proficient language users”. However, even at the most fundamental A1 level, proficiency is too high for novice learners of foreign languages; as a result, Pre-A1 level before reaching A1 level has been complemented in the CEFR Companion Volume.

Pre-A1 level is a band of proficiency at which the learner has not yet acquired a generative capacity, but relies upon a combination of words and formulaic expressions [4]. At this level, learners are the beginners, who do not have a vocabulary structure yet and know the simple words they learned in class. As is appropriate for learners of Pre-A1 level, reading comprehension tasks focus on reading short sentences and recognizing words. Longer tasks are mainly based on simple stories, so learners should be provided as much opportunity as possible to read and enjoy stories at their level. In addition, Pre-A1 level produces simple utterances, and generally responds at word or phrase but may also produce some longer utterances. To support for these requirements, the CEFR Companion Volume has provided seven descriptors (CDSs) in terms of reading comprehension for learners of Pre-A1 level.

3. CLASSIFICATION FEATURES

3.1 Document Type

In document type estimation, from ten types used in Takada [2], Hirakawa [3] selected seven document types from CDSs because some of the ten types were similar and undistinguishable. Seven document types were “articles + news”, “newspaper articles”, “public documents”, “signs + posters”, “communication statements”, “instructions” and “others”.

In regard to the Pre-A1 level descriptors, most of them emphasize the language in daily use with very short, simple words and sentences. In addition, example sentences of Pre-A1 level can be seen in everyday signs; posters, flyers and notices; materials illustrated by pictures; letters, cards or email, notes or text messages; and instructions. However, “articles + news”, “newspaper articles”, and “public documents” in [3] are not focused much in these descriptors because example sentences are longer and require the knowledge in certain fields.

On the other hand, posters, flyers and notices are one of the main document types used in everyday context, while signs mainly emphasize the recognition of simple and familiar words. Moreover, when collecting the example sentences of Pre-A1 level, most of them belong to type of notices that provide much information about places, times and prices, rather than in posters and flyers. As a result, to improve the accuracy in estimation of document types of example sentences of Pre-A1 level, we divide the last four types used in [3] into five ones. “signs + posters” is divided into “signs” and “notices”, in which “notices” includes the example sentences can be seen in posters, flyers, notices, etc. In addition, example sentences belonging to menus, picture books, stories, etc., with visual illustration are listed as “others”. Consequently, there are eight document types used in this study: “articles + news”, “newspaper articles”, “public documents”,

† Shizuoka University
‡ Tokoha University

"signs", "notices", "communication statements", "instructions", and "others".

The data set used for the experiment on document type estimation includes 1,423 example sentences collected in [3], and 149 new example sentences categorized in level Pre-A1. These example sentences were collected by ten collaborators who are currently teachers of Japanese as a foreign language (eight of them are Vietnamese, two of them are Japanese). In total, 1,572 example sentences were used in this experiment.

From [3], fastText was considered as the most effective method to estimate document types. Therefore, we used fastText with the same parameters proposed in [3] to carry out the experiment. On average, the accuracy reached 77.48%.

3.2 Kanji Rate

The most common Japanese writing system is based on the mixture of Kanji and Kana (Hiragana and Katakana). Kana has clear and direct character-to-sound correspondences where each Kana represents Japanese mora. In contrast, Kanji (originally derived from Chinese characters) is commonly used for writing content words – most of nouns, verbs and adjectives are written in Kanji – and Kanji characters, alone or combination with other characters, represent whole words [5]. It implies that reading Kanji ability may require different reading strategies or different cognitive skills to acquire reading comprehension of Japanese.

In regard to the applications of learning foreign languages, "readability" is used to define how difficult learners evaluate a reading text. Functions which define readability are called readability formulas, and there have been many readability formulas developed for learning Japanese. Morioka et al. [6] and Yasumoto et al. [7] proposed formulas using the sentence length measured in letters, words and the percentage of Kanji characters for estimating the difficulty of the vocabulary. It is said that a text with longer sentences is estimated as difficult, and a text with more Kanji characters is also estimated as difficult.

On the other hand, the number of Kanji characters which students have to master is specified in each grade affecting the reading comprehension of language learners. According to the curriculum for the school subject "Japanese language" by the Ministry of Education, there are six Grades 1-6 of Kanji acquisition. However, the number of Kanji taught in Grade 1 is limited to 76 characters, students in Grade 1 can only achieve the accuracy of around 80% in reading. The accuracy is suggested to improve when learners are taught more Kanji characters for higher grades [8]. Acknowledging the important role of Kanji characters in reading comprehension, especially at low levels, we decided to use Kanji rate as a feature of CDS classification to improve the accuracy of classification in Pre-A1 level.

Moreover, besides the number of Kanji characters, the degree of difficulty of each Kanji characters affects reading ability of learners. The more Kanji characters appearing in a reading text at low levels, the more readable the text becomes [9]. In this study, we use four lists of Kanji characters of Japan-Language Proficiency Test (JLPT). The four lists are named JLPT level 4, JLPT level 3, JLPT level 2 and JLPT level 1 [10]; in which the difficulty rises from level 4 to level 1. In addition, in case a Kanji character appearing in a reading text does not belong to any of the four lists, it will be added to the list named "Other". Consequently, there are five Kanji lists to be estimated in this study.

In this experiment, we used 370 example sentences used in [3] and 149 example sentences of Pre-A1 level collecting this time as was mentioned in 3.1; in total 519 example sentences. Although the data was not sufficient in quantity, machine learning was attempted to carry out the experiment with the Kanji rates calculated for each level.

4. CDS CLASSIFICATION EXPERIMENT

Regarding the CDS classification experiment, we continue to use "length", "document type" and "technicality" proposed by Takada [2], and a new feature "Kanji rate".

As in [2], we assumed multi-label classification corresponding to multiple CDSs in an example sentence. For the data set of the experiment, we used 519 example sentences with multi-label information collected from 10 experienced Japanese educators with knowledge in CEFR. The average number of CDSs corresponding to one example sentence was about 2.82. 34 binary classifiers (SVM) were used for the multi-label classification method, and the cross validation with three divisions was performed for the evaluation.

The classification results for all example sentences were about 70.29% positive recall, about 38.39% positive precision, about 49.66% positive F-value, about 89.82% negative recall, about 97.10% negative precision, and the negative F-value was about 93.32%. Although it was confirmed that the positive recall, negative recall and precision ratio were maintained at a relatively high level, the accuracy of the positive precision rate was low. One of the possible reason was that the positive predicted number for one example sentence was about 5.15 on average.

ACKNOWLEDGEMENT

This work was supported by a Grant-in-Aid (KAKENHI) for "Scientific Research (C) (18K00722)."

REFERENCES

[1] Council of Europe, "Common European Framework of Reference for Languages: Learning, Teaching Assessment," Cambridge University Press (2001).

[2] 高田宏輝, 宮崎佳典, 谷誠司, "韓国人日本語学習者のための CEFR 読解指標に基づく例文分類", 韓國日本學會第 94 回國際學術大會, pp. 299-303 (2017).

[3] 平川遼汰, 宮崎佳典, 谷誠司, "日本語例文自動分類による CEFR 読解指標推定支援 Web アプリケーションの開発", 情報処理学会第 80 回全国大会, pp. (4)-635-636 (2018).

[4] Council of Europe, "Common European Framework of Reference for Languages: Learning, Teaching. Companion Volume with New Descriptors" (2017).

[5] Alexandra S. Dylman, Mariko Kikutani, "The role of semantic processing in reading Japanese orthographies: an investigation using a script-switch paradigm", Reading and Writing, Vol.31, No.7, pp. 503-531 (2018).

[6] 森岡健二, "ことばの教育", 明治書院 (1988).

[7] 安本美典, "説得の文章技術", 講談社現代新書 (1983).

[8] Katsuo Tamaoka, "A Japanese Perspective on Literacy and Biliteracy: A National Paper of Japan," The Reading Research Symposium (1996).

[9] Komori Saeko, "A Study of Kanji Word Recognition Process for Japanese as a Second Language", Tokyo: Kazama Shobo (2009).

[10] http://kanjicards.org/kanji-lists.html (Reference date: 2018.6.20).