The Companion Volume to the Proceedings of 41st Annual Meeting of the
Association for Computational Linguistics, July 2003, pp. 189-192.

Automatic Detection of Grammar Elements that Decrease Readability

Masatoshi Tsuchiya and **Satoshi Sato**
Department of Intelligence Science and Technology,
Graduate School of Informatics, Kyoto University
tsuchiya@pine.kuee.kyoto-u.ac.jp, sato@i.kyoto-u.ac.jp

Abstract

This paper proposes an automatic method of detecting grammar elements that decrease readability in a Japanese sentence. The method consists of two components: (1) the check list of the grammar elements that should be detected; and (2) the detector, which is a search program of the grammar elements from a sentence. By defining a readability level for every grammar element, we can find which part of the sentence is difficult to read.

1 Introduction

We always prefer readable texts to unreadable texts. The texts that transmit crucial information, such as instructions of strong medicines, must be completely readable. When texts are unreadable, we should rewrite them to improve readability.

In English, measuring readability as *reading age* is well studied (Johnson, 1978). The reading age is the chronological age of a reader who could just understand the text. The value is usually calculated from the sentence length and the number of syllables. From this value, we find whether a text is readable or not for readers of a specific age; however, we do not find which part we should rewrite to improve readability when the text is unreadable.

The goal of our study is to present tools that help rewriting work of improving readability in Japanese. The first tool is to help detect the sentence fragments (words and phrases) that should be rewritten; in other words, it is a checker of "hard-to-read" words and phrases in a sentence. Such a checker can be realized with two components: the check list and its detector. The check list provides check items and their readability levels. The detector is a program that searches the check items in a sentence. From the detected items and their readability levels, we can identify which part of the sentence is difficult to read.

We are currently working on three aspects concerned with readability of Japanese: kanji characters, vocabulary, and grammar. In this paper, we reports the readability checker for the grammar aspect.

2 The check list of grammar elements

The first component of the readability checker is the check list; in this list, we should define every Japanese grammar element and its readability level. A grammar element is a grammatical phenomenon concerned with readability, and its readability level indicates the familiarity of the grammar element.

In Japanese, grammar elements are classified into four categories.

1. Conjugation: the form of a verb or an adjective changes appropriately to the proceed word.
2. Functional word: postpositional particles work as case makers; auxiliary verbs represent tense and modality.
3. Sentential pattern: negation, passive form, and question are represented as special sentence patterns.
4. Functional phrase: there are idiomatic phrases works functionally, like "not only ... but also ..." in English.

A grammar section exists in a part of the Japanese Language Proficiency Test, which is used to measure and certify the Japanese language ability of a person who is a non-Japanese. There are four levels in this test; Level 4 is the elementary level, and Level 1 is the advanced level.

Test Content Specifications (TCS) (Foundation and Association of International Education, 1994) is intended to serve as a reference guide in question compilation of the Japanese Language Proficiency Test. This book describes the list of grammar elements, which can be tested at each level. These lists fit our purpose: they can be used as the check list for the readability checker.

TCS describes grammar elements in two ways. In the first way, a grammar element is described as a 3-tuple: its name, its patterns, and its example sentences. The following 3-tuple is an example of the grammar element that belongs to Level 4.

Name daimeishi 代名詞 (Pronoun)
Patterns kore コレ (this), sore ソレ (that)
Examples kore ha hon desu. これは本です。(This is a book.),
sore ha nōto desu. それはノートです。(That is a note.)

Grammar elements of Level 3 and Level 4 are conjugations, functional words and sentential patterns that are defined in this first way. In the second way, a grammar element is described as a pair of its patterns and its examples. The following pair is an example of the grammar element that belongs to Level 2.

Patterns ta tokoro ～たところ (when ...)
Examples sensei no otaku he ukagatta tokoro 先生のお宅へ伺ったところ
(When visiting the teacher's home)

Grammar elements of Level 1 and Level 2 are functional phrases that are defined in this second way.

We decided to use this example-based definition for the check list, because the check list should be independent from the implementation of the detector. If the check list depends on detector's implementation, the change of implementation requires change of the check list.

Each item of the check list is defined as a 3-tuple: (1) readability level, (2) name, and (3) a list of example pairs. There are four readability levels according to the Japanese Language Proficiency Test. An example pair consists of an example sentence and an instance of the grammar element. It is an implicit description of the pattern detecting the grammar element. For example, the check item for 'Adjective (predicative, negative, polite)' is shown as follows,

Level 4
Name Adjective (predicative, negative, polite)
Test Pairs
Sentence$_1$ kono heya ha hjroku nai desu. この部屋は広くないです。
(This room is not large.)
Instance$_1$ hjroku nai desu 広くないです
(is not large)

The instance 広くないです/hirokunaidesu/ consists of three morphemes: (1) 広く/hiroku/, the adjective means 'large' in renyo form, (2) ない/nai/, the adjective means 'not' in root form, and (3) です/desu/, the auxiliary verb ends a sentence politely. So, this test pair represents implicitly that the grammar element can be detected by a pattern "Adjective(in renyo form) + nai + desu".

All example sentences are originated from TCS. Some check items have several test pairs. Table 1 shows the size of the check list.

Table 1: The size of the check list

Level	# of rules
1	134
2	322
3	97
4	95
Total	648

3 The grammar elements detector

The check list must be converted into an explicit rule set, because each item of the check list shows no explicit description of its grammar element, only shows one or more pairs of an example sentence and an instance.

3.1 The explicit rule set

Four categories of grammar elements leads that each rule of the explicit rule set may take three different types.

- Type M: A rule detecting a sequence of morphemes
- Type B: A rule detecting a *bunsetsu*.
- Type R: A rule detecting a modifier-modifee relationship.

Type M is the basic type of them, because almost of grammar elements can be detected by morphological sequential patterns.

Conversion from a check item to a Type M rule is almost automatic. This conversion process consists of three steps. First, an example sentence of the check item is analyzed morphologically and syntactically. Second, a sentence fragment covered by the target grammar element is extracted based on signs and fixed strings included in the name of the check item. Third, a part of a generated rule is relaxed based on part-of-speech tags. For example, the check item of the grammar element whose name is "Adjective (predicative, negative, polite)" is converted to the following rule.

```
np( 4, 'Adjective
       (predicative,negative,polite)',
    Dm({ H1=>'Adjective',
         K2=>'Basic Renyou Form' },
       { G=>'ない/nai/',
         H1=>'Postfix', K2=>'Root Form' },
       { G=>'です/desu/',
         H1=>'Auxiliary Verb' }) );
```

The function `np()` makes the declaration of the rule, and the function `Dm()` describes a morphological sequential pattern which matches the target. This example means that this grammar element belongs to Level 4, and can be detected by the pattern which consists of three morphemes.

Type B rules are used to describe grammar elements such as conjugations including no functional words. They are not generated automatically; they are converted by hand from type M rules that are generated automatically. For example, the rule detecting the grammar element whose name is "Adjective in Root Form" is defined as follows.

```
np( 4, 'Adjective in Root Form',
    Db( { H1=>'Adjective',
          K2=>'Root Form' } ) );
```

The function `Db()` describes a pattern which matches a *bunsetsu* which consists of specified morphemes. This example means that this grammar element belongs to Level 3, and shows the detection pattern of this grammar element.

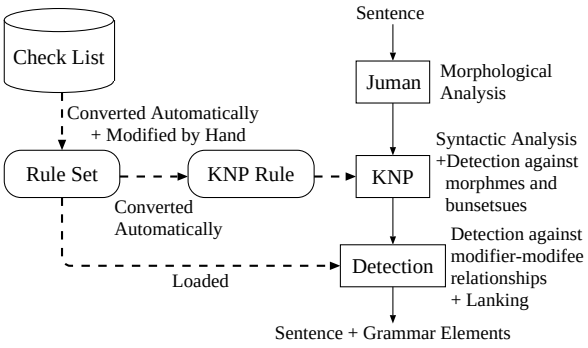

Figure 1: System structure

Type R rules are used to describe grammar elements that include modifier-modifee relationships. In the case of the grammar element whose name is "Verb Modified by Adjective", it includes a structure that an adjective modifies a verb. It is impossible to detect this grammar element by a morphological continuous pattern, because any *bunsetsu*s can be inserted between the adjective and the verb. For such a grammar element, we introduce the function `Dk()` that takes two arguments: the former is a modifier and the latter is its modifee.

```
np( 4, 'Verb Modified by Adjective',
    Dk( Db({ H1=>'Adjective',
             K2=>'Basic Renyou Form' }),
        Dm({ H1=>'Verb' }) ) );
```

3.2 The architecture of the detector

The architecture of the detector is shown in Figure 1. The detector uses a morphological analyzer, Juman, and a syntactic analyzer, KNP (Kurohashi and Nagao, 1994). The rule set is converted into the format that KNP can read and it is added to the standard rule set of KNP. This addition enables KNP to detect candidates of grammar elements. The 'Detection' part selects final results from these candidates based on preference information given by the rule set.

Figure 2 shows grammar elements detected by our detector from the sentence "地図(chizu)はおろか(oroka)、略図(ryakuzu)さえも(sae mo)配られ(kubarare)なかった(nakatta)。" which means "Neither a map nor a rough map was not distributed."

4 Experiment

We conducted two experiments, in order to check the performance of our detector.

Fragment		Name	Level
chizu 地図	(a map)	-	-
ha oroka はおろか	(neither)	～ha oroka はおろか (neither ...)	1
、	(,)	読点 (comma)	4
ryakuzu 略 図	(a rough map)	-	-
sae さえ	(even)	～sae さえ (even ...)	2
mo も	(nor)	も!副 (huku postpositional particle means 'nor')	4
kubarare 配られ	(distributed)	Ⅴreru レル (passive verb phrase)	3
nakatta なかった	(was not)	～nai ない (predicative adjective means 'not')	4
。	(.)	句点 (period)	4

Figure 2: Automatically detected grammar elements

The first test is a closed test, where we examine whether grammar elements in example sentences of TCS are detected correctly. TCS gives 840 example sentences, and there are 802 sentences from which their grammar elements are detected correctly. From the rest 38 sentences, our detector failed to detect the right grammar element. This result shows that our program achieves the sufficient recall 95% in the closed test. Almost of these errors are caused failure of morphological analysis.

The second test is an open test, where we examine whether grammar elements in example sentences of the textbook, which is written for learners preparing for the Japanese Language Proficiency Test (Tomomatsu et al., 1996), are detected correctly. The textbook gives 1110 example sentences, and there are 680 sentences from which their grammar elements are detected correctly. Wrong grammar elements are detected from 71 sentences, and no grammar elements are detected from the rest 359 sentences. So, the recall of automatic detection of grammar elements is 61%, and the precision is 90%. The major reason of these failures is strictness of several rules; several rules that are generated from example pairs automatically are overfitting to example pairs so that they cannot detect variations in the textbook. We think that relaxation of such rules will eliminate these failures.

References

The Japan Foundation and Japan Association of International Education. 1994. *Japanese Language Proficiency Test: Test content Specifications (Revised Edition)*. Bonjin-sha Co.

Keith Johnson. 1978. Readability. `http://www.timetabler.com/readable.pdf`.

Sadao Kurohashi and Makoto Nagao. 1994. A syntactic analysis method of long Japanese sentences based on the detection of conjunctive structures. *Computational Linguistics*, 20(4).

Etsuko Tomomatsu, Jun Miyamoto, and Masako Waguri. 1996. *Donna-toki Dou-tsukau Nihongo Hyougen Bunkei 500*. ALC Co.