Extending Japanese Grammar Ruleset in LanguageTool

Kaori Chiba s1200211 Supervised by Maxim Mozgovoy

Abstract

LanguageTool is an application that detects grammar errors in sentences of Japanese and other languages by looking for error patterns, and shows correct answer. It can react to errors with the help of numerous “error rules”. These rules are error descriptions made with patterns, representing language grammar errors. However, these rules are still under development and sometimes they do not react to errors, so they can be improved. In this study, some problems of the rules were found and improved. They were principally spelling mistakes and grammar mistakes. After improvement, the rules are able to detect Japanese language errors such as the negative conjunction of adjectives, several types of spelling mistakes, and certain homonym and punctuation errors.

1 Introduction

LanguageTool can find errors in sentences and display some detail of incorrect grammar. There are certain rules for detecting linguistic errors in the system. Rules can be written in Extensible Markup Language (XML) and Java [1]

However, these rules are not yet complete, so not all errors can be detected. When LanguageTool cannot react to incorrect grammar, it means that the rules do not have codes to detect such errors.

1.1 What is LanguageTool?

Figure 1 is a screenshot of LanguageTool screen. The top half of the screen is an input window with an example sentence. The bottom half window is an output. The output displays found errors, messages pointing the errors, and the number of errors found. LanguageTool can react to errors in languages other than Japanese if we choose another language in the menu bar at the bottom of the screen. Sentences are searched for errors automatically if the checkbox next to the menu bar is clicked.

LanguageTool is a proofreading program. It can handle more than 20 languages including English, French, German, Polish, and Japanese. Its current version is 3.1 and requires Java 7+ (as of 2015/12). This can be used with LibreOffice, OpenOffice,

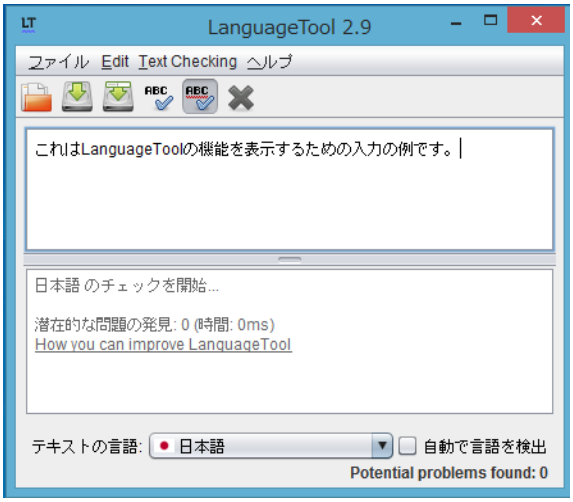

Figure 1. LanguageTool Main Screen

Firefox, Chrome and as a standalone PC product.

LanguageTool is an open source project, and it can be improved. Its developers invites everyone to improve the system by posting ideas and bugs. [1]

1.2 Current Issues

In LanguageTool, many grammatical errors are not detected. I divided found errors into the following list. In the present work, the following types of undetected grammatical errors are addressed:

- Negative conjugation of *i*/*na*-adjective.
- Spelling of di, zi, du, zu.
- Homonyms (21 items).
- Consecutive punctuation symbols.

Negative conjugations are typical grammar mistakes that occur due to incorrect change of Japanese adjective endings in negative form.

Example:

- *i*-adjective: 怖い + ない → 怖いない (Correct is 怖くない)
- *na*-adjective: 素敵な + ない → 素敵ない (Correct is 素敵ではない)

Spelling of di, zi, du and zu are common spelling mistakes. Essentially, in modern Japanese most words are spelled in accordance with present-day phoneme

as zi and zu although a few words are spelled with di and du [2].

Example:

- 一つづつ (Correct is 一つずつ)
- 恥ぢる (Correct is 恥じる)
- こずく(Correct is こづく)
- はなじ(Correct is はなぢ)

Homonyms are the words that have the same pronunciation but different meanings (and written differently in Kanji). Sometimes people choose improper spelling.

Example:

- この本は暑い (Correct is この本は厚い)
- このスープは厚い
  (Correct is このスープは熱い)

The words 熱い, 暑い and 厚い are read /atsui/. But, they have different meanings. 熱い means that a substance is hot. 暑い expresses hot air temperature. 厚い means that an object is thick. There are other homonym pairs such as 暖かい/温かい and 初めて/始めて. Also, there are some words written in different kanji characters but their meanings are the same such as 美味い and 旨い (both are pronounced as /umai/, and both spellings are acceptable)

Two consecutive punctuations, such as a period and comma, used in a sentence, are an error.

Example:

- このりんごはうまい。。
- したがって、。

2 Work Process

2.1 Finding Issues in LanguageTool

When incorrect sentences are input into LanguageTool, if it does not react to a linguistic error, existing rules should be changed or new rules should be added. The rules for the Japanese grammar are located in the directory `ja`. There is a file named `grammar.xml`, containing XML rules as mentioned above. After improving the rules, incorrect sentences are input again and checked if LanguageTool can react to them.

2.2 Improvement of LanguageTool

Since LanguageTool rules operate with individual morphemes, we had to make sure that the system's built-in tokenizer works well with our examples. We developed a number of test sentences and placed them into the file `test.txt`. Then we checked the quality of the tokenizer by executing the following command:

```
java -jar languagetool-commandline.jar
          -v -l ja test.txt
```

Then we carefully examined the output (see Figure 2) and used the morphemes identified by LanguageTool in the new grammar rules.

```
C:¥Users¥ADMIN>cd C:¥Users¥ADMIN¥Desktop¥LanguageToo
l-2.9

C:¥Users¥ADMIN¥Desktop¥LanguageTool-2.9>java -jar la
nguagetool-commandline.jar -v -l ja test.txt
Expected text language: Japanese (no spell checking
active, specify a language variant like 'en-GB' if a
vailable)
Working on test.txt...
1087 rules activated for language Japanese
<S> 怖い[怖い/形容詞-自立]な[だ/助動詞]い[い/名詞-一
般]。[。/記号-句点,</S>]
Disambiguator log:
```

Figure 2. Testing LanguageTool Tokenizer

Next, we created new rules and rule categories using XML language. The final ruleset is stored in the file `ja\grammar.xml`.

3 Improved Points

3.1 Defining Rules in XML

According to [3], "XML is a software and hardware independent tool for storing and transporting data." The forms of rules are defined in XML as the following example shows.

Example:

```
 <category name="同音異義語">
 <rule id="WOHAZIMERU" name="を初める">
   <pattern case_sensitive="no">
    <token skip="3">を</token>
    <token inflected="yes">初める</token>
   </pattern>
   <message>「初」は副詞です。
      <suggestion>始めて</suggestion>
             の間違いです。</message>
   <url>http://kanjibunka.com/kanji-
faq/old-faq/q0422/</url>
   <example type="correct">料理を
       <marker>始める</marker>。</example>
   <example type="incorrect">料理を
<marker>初める</marker>。</example>
  </rule>
</category>
```

Words surrounded by <> or </> are called elements. For example, `<category>` is called the element of

category.

As mentioned above, since the rules are incomplete in LanguageTool, there are some problems of the rules need to be resolved. In order to solve them, first the problems were sorted into several categories. Second, a rule was created to detect each linguistic error in each category and to suggest the correct form. Then, the rules were assigned a unique ID and name. Next, patterns were made to react to incorrect grammar uses by with the help of tokens, attributes, and sentences divided into morphemes. Finally, new error messages were developed. Sentences surrounded by elements of *message*, *suggestion*, or *URL* can be reported. Words surrounded by the elements of *suggestion* are regarded corrected words and indicated in bold.

When necessary, URLs are shown to reveal details. Then, examples of correct types and incorrect types are written using the XML elements of *example type* and *marker*. They are used to show the proper use of grammar in the current case and highlight certain words in example sentences.

3.2 Rule Attributes

Each *token* XML element has certain attributes, such as:

- skip;
- inflected;
- regexp;
- postag.

The attribute *skip=n* is used to detect patterns, separated by at most *n* morphemes. For example, the sentence “今日は熱い” (the correct form is 暑い) could be changed to some variations:

- 今日はとても熱い。
- 今日はいつもより熱い。

With additional words as shown above, those sentences are recognized as by LanguageTool. If we want to let the system react those sentences, we have to use *skip* attribute.

The attribute *inflected* can react to words in different forms. In the example below, the correct forms should be 始める・始めた as a verb, so the examples with 初 as an adverb are incorrect:

- 初める
- 初めた

If the *inflected* attribute is not used, LanguageTool reacts to 初める, but does not react to 初めた. Since the attribute can react to different word forms, there is no need to write separate patterns for each case.

The *regexp* attribute is similar to OR statement of C or Java programming langauges. LanguageTool recognizes the examples below as different sentences:

- ちいさいない
- 小さいない

However, are were divided into morphemes in a similar way:

- ちいさい/な/い
- 小さい/な/い

Therefore, we can use the following pattern to detect them:

```
<token regexp="yes">小さい|ちいさい<token>
```

Different types of morphemes (parts of speech) can be gathered by using the *postag* attribute. For example, the adverb 初めて could modify a verb. The correct form is 初めて in the following example:

- 始めて作る。
- 始めて料理する。

If rules were made to react to each possible verb, the file data would be massive because the number of verbs is enormous. Verbs could be unified and matched by using the pattern `<token postag="動詞.*" postag_regexp="yes"/>`. [4]

3.3 Reacting to Errors

LanguageTool reaction to different grammatical errors is shown in Figure 3.

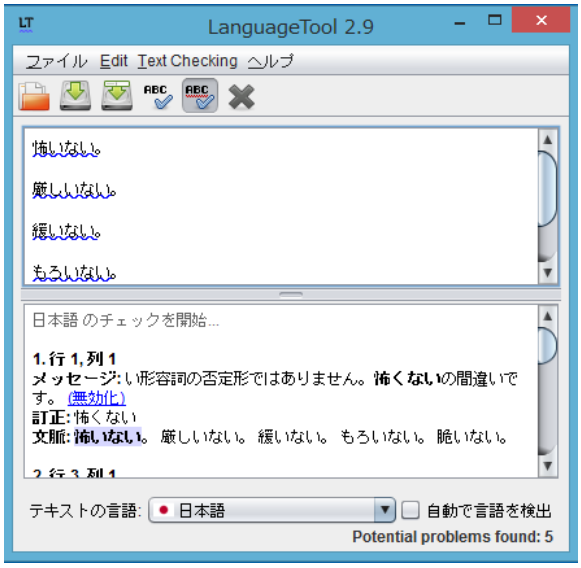

Figure 3. Reacting to Errors

The system uses the following messages to react to the errors of our error categories:

- *i*-adjective
  大きいない → 大きくない

Common message:
い形容詞の否定形ではありません。

- *na*-adjective
  完璧ない　→　完璧ではない

Common message:
な形容詞の否定形ではありません。

Adjectives were collected from [5]. However, some of *na*-adjectives such as 元気ない are not actually incorrect because many people use this word. LanguageTool did not have rules for them.

- Spelling of di and zi
  いちぢく　→　いちじく（無花果）
  あさじえ　→　あさぢえ（浅知恵）

Common message:
「ぢ」ではなく「じ」を用います。
or
「じ」ではなく「ぢ」を用います。

- Spelling of du and zu
  あづける　→　あずける（預ける）
  あいそずかし　→　あいそづかし
  （愛想尽かし）

Common message:
「づ」ではなく「ず」を用います。
or
「ず」ではなく「づ」を用います。

These words were collected from [2] and [6]. LanguageTool shows the URL [2] that provides detailed information.
Some words with either du and zu are correct, such as きづく（気付く）and きずく（築く）. LanguageTool will not react to them.

- Homonyms
  この本は暑い　→　この本は厚い

Common message: 同音異義語です。

Currently, the system reacts to 21 homonyms, and some words are reacted if they set a very likely context, such as 今日は or スープ, but it will not react for other words, where the context is not so clear. For 始める and 初めて, the URL [7] appears. This website provides a detailed explanation about the difference between 始める and 初めて.

- Consecutive punctuations
  それから、、　→　それから、

Common message:
句点・読点は連続で使いません。

4 Summary of Results

Table 1. Summary of added rules

Category name	New or existing category	New rules	Total Rules
い形容詞　＋　否定形 (*i*-adjective)	New category	142	142
な形容詞　＋　否定 (*na*-adjective)	New category	64	64
「ず」と「づ」の使い分け (Spelling of di and zi)	New category	59	63
「じ」と「ぢ」の使い分け (Spelling of du and zu)	New category	55	56
同音異義語 (Homonyms)	New category	20	20
文法 (Consecutive punctuations etc.)	Existing category	1	79

As a result, I added several new categories of rules and significantly increased the total number of rules in LanguageTool. These changes are shown in the Table 1. Furthermore, some rules were transferred from their previous categories to new categories, as shown in the Table 2.

Table 2. Changes in categories

Rule	Former category	New category
こじんまり (Correct is こぢんまり)	誤字	「じ」と「ぢ」の使い分け
ずらい (Correct is づらい)	文法	「ず」と「づ」の使い分け
つくずく (Correct is つくづく)	文法	「ず」と「づ」の使い分け
つれずれ (Correct is つれづれ)	文法	「ず」と「づ」の使い分け
基ずく (Correct is 基づく)	文法	「ず」と「づ」の使い分け

5 Conclusion

To summarize, LanguageTool can react to more errors than before. Five new categories were created, and 346 rules were added. The number of categories became 15 and the total number of error patterns is about 10,713. However, some problems are not

resolved yet.

6 Future Work

LanguageTool has many problems related to Japanese grammar rules.

First, some negative adjectives are divided into morphemes in a wrong way. For example, ない is used when a word is changed to negative form. However, the word is divided as な/い occasionally, which is wrong. For example, 暑いない should be divided as 暑い/ない, but LanguageTool divided it as 暑い/な/い.

Table 3. Homonym reaction words

Word	Pronunciation	Reaction words
暑い・熱い 厚い	/atsui/	本・今日 スープ
優しい・易しい	/yasashii/	問題・性格
速く・早く	/hayaku/	起きる 走る
暖かい・温かい	/atatakai/	食べ物 気温
上手い・巧い・美味い 旨い	/umai/	食べ物 やり方
拙い・不味い	/mazui/	食べ物 行動
硬い・堅い	/katai/	口・石
軟らかい 柔らかい	/yawarakai/	布団・肉

Second, if conjunction is attached to some words, LanguageTool divides into morphemes unusually. For example:

- は/なじ
- は/な/じ/が

The correct form of the above example is はなぢ. However, if we attach a particle が, LanguageTool will not recognize はなじ.

Third, LanguageTool can react only to certain homonym words. For example, if 暑い, 厚い and 熱い are used with a noun other than 本, 今日, スープ, the system will not react. As mentioned above, 熱い means that a substance is hot, 暑い expresses hot air temperature, and 厚い means that an object is thick. However, LanguageTool cannot not determine each noun associated with each adjective. It reacts only to the words listed in the Table 3.

Fourth, LanguageTool cannot react the sequence of two punctuation marks “。、”, though they are incorrect according to the grammar.

Besides that, LanguageTool cannot react to certain errors not addressed in this research.

Acknowledgement

I would like to thank Professor Maxim Mozgovoy and Hiroki Yatsu for their comments and supports for my graduation thesis.

References

[1] LanguageTool Style and Grammar Check https://languagetool.org/

[2] 現代仮名遣い：文部科学省 (Modern kana usage: Ministry of Education, Culture, Sports, Science and Technology) http://www.mext.go.jp/b_menu/hakusho/nc/k19860701001/k19860701001.html

[3] XML Introduction - W3Schools www.w3schools.com/xml/xml_whatis.asp

[4] LanguageTool Wiki Open Source proof-reading tool http://wiki.languagetool.org/development-overview#toc0

[5] みんなの日本語８～５０課形容詞一覧表 – N2et (Everyone’s Japanese 8~50 adjective lessons of list) http://jn2et.com/Kyouan/L51-2han-keiyousilist.html

[6] 気になる言葉館 (Hall of words causing difficulties) http://kotoba.merrymall.net/af00_00.html

[7] 漢字文化資料館 (Museum of kanji culture) http://kanjibunka.com/kanji-faq/old-faq/q0422