Developing Japanese Grammar Checking Rules for LanguageTool

Hiroki Yatsu s1190229 Supervised by Maxim Mozgovoy

Abstract

LanguageTool application can find grammar and typographical errors, and can check many languages. However, at present, LanguageTool is very low accuracy. Therefore, I improved LanguageTool of Japanese. I increased seven hundred thirty numbers to be able to find mistakes a kind from forty three numbers, and I changed messages considering that user can understand why is sentences wrong, and I repaired that LanguageTool can find even if words and sentences somewhat change. As a result, LanguageTool of Japanese advanced precision just a little.

1 Introduction

It is a natural language that human have been used languages which occurred in natural generation like English and Japanese, and it is a natural language processing to analyze the natural language at computers [1]. If technology of the natural language processing develops, it is useful to read proofs and to research a feature of novels and an essays. In general, it is said that it is difficult to correctly conduct the natural language processing. That is because, the natural language change similar to wight, and new words and new usage is created constantly. Moreover, one word have many meanings and this meaning change [1].

Specially, it is troublesome to handle the Japanese language at computers. Words of Japanese isn't to divide like English. On that account, if we want to analyze Japanese, we will need to execute morphological analysis first of all. Also, it is confused by computers that Japanese have many homonyms too. However, at a same time, it is said that Japanese is very hard for nonnative who try to learn. Then, I thought that it is in demand to precisely point out something such mistakes as grammar and typographical errors to use computers.

Application which is called LanguageTool exist [2]. The main function of LanguageTool is proofreading of texts. If we input texts which we want to check, LanguageTool will output mistakes positions and output explanations why is sentences wrong. LanguageTool can find mistakes of many languages. Proofreading application is more. However, those cannot handle many languages, and cannot detect and several grammar problems.

Though LanguageTool is incomplete. In actually, at present, LanguageTool cannot find mistakes almost all. Therefore, I improved LanguageTool of Japanese. I want to stand on the start line to complete LanguageTool in the future.

2 LanguageTool

2.1 What is LanguageTool?

LanguageTool can find mistakes of words（grammar and typographical errors). It finds many errors that is a simple spell checker cannot detect and several grammar problems.

LanguageTool is continuing version up. New version is version 2.7（2014/12 in time）. Version 2.7 can handle thirty eight languages. Also, LanguageTool is a open source project. That is, everyone can improve.

2.2 How to use LanguageTool?

We can download from the top page of a site, and we can use for browser extension (only Firefox）, desktop, and OpenOffice. At first, we need to unzip a file. Next, we click the LanguageTool.jar file. Then, LanguageTool application will activate.

How to use is very easy. We choose language from a bar in a bottom. Moreover, we input text which want to check in a upper frame. Then, LanguageTool will output mistake portions, and output correct text of this in a lower flame. Also, mistakes portions is drew a wavy line, and LanguageTool output demonstrations why is sentences wrong.

3 Method

3.1 How to improve

I found mistakes of Japanese languages which is a frequent occurrence. Moreover, I made xml file is based on these mistakes. (XML is the Extensible Markup Language, a subset of the Standard General-

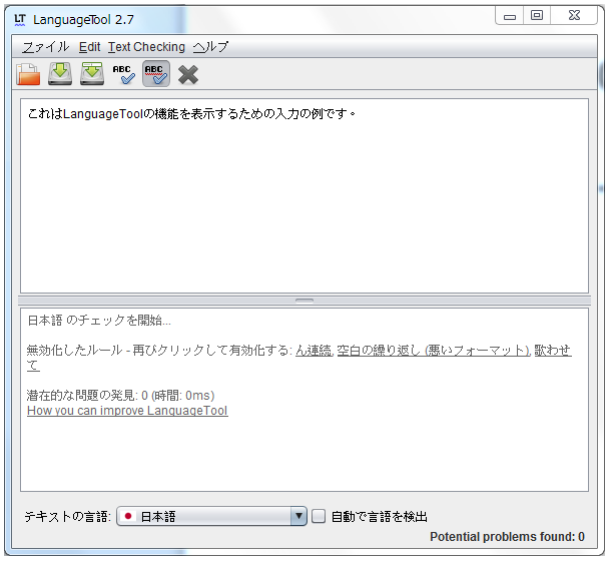

Figure 1: LanguageTool

ized Markup Language(SGML) and a companion to the HTML [3]). Then, I improved details because I wanted to be able to extract various mistakes and to comfortably use by users.

When I made XML to some extent, I input ”testrules.bat ja” from command prompt. In addition, I directly input in LanguageTool and checked whether or not to correctly output and be easy to look. I repeated these.

I thought the way which can be pointed out even if sentences change after I made basic structure of XML. Mainly, I add attribute of a token tag. The details will be described later.

3.2 How do you make xml?

Of course, we are possible to write XML ordinarily. However, There is a convenient tool in the LanguageTool site. It is ”LanguageTool Rule Editor”. This site automatically create XML when we input the requirement.

We select language which we want to check, and we input the wrong sentence and the corrected sentence. Then we push ”create initial error pattern”. We input requirement after the tool display ”Set the Error Pattern” and ”Set the Rule Details”. XML is created with this.

3.3 Basic structure of XML

```
<rule id="MOTOMERERU" name="求めれる">
  <pattern case_sensitive="no">
    <token regexp="yes">求め|もとめ</token>
    <token inflected="yes">れる</token>
  </pattern>
  <message>ら抜き言葉があります。
    <suggestion>求められる</suggestion>
  の間違いです。</message>
  <url>http://www3.kcn.ne.jp
  /~jarry/keig/c01c05.htm</url>
  <example type="correct">
    <marker>求められる</marker>。
  </example>
  <example type="incorrect">
    <marker>求めれる</marker>。
  </example>
</rule>
```

Japanese grammar.xml file is in the chart above wise. I wrote explanation of cardinal tags shown below.

rule: This tag appoint address name. We need to write address names in attribute like ”ID=””””. Address names must not use same name because LanguageTool output errors.

pattern: If we write in attribute like $case_sensitive$ = ”*yes*”, lower case letter differ from upper case letters of English and so on. I didn’t need to change this tag because I improved LanguageTool of Japanese. Also, we can use ”marker” of a sub element. If we use marker in texts, texts will be emphasized in a thick font character.

token: This tag is most important. If we want to be output that mistake, we need to use this tag. If we write attribute, this tag will be added various functions. I tried much attribute and wrote those later on.Also, we must take care when we use token tag. If we write according to a rule, LanguageTool cannot output. We need to be morphological analysis texts which want to check. Morphological is least a meaningful unit, and morphological analysis is to divide in meaningful words and to discriminate part of speech and content to use a dictionary. For example, In “I love you.”, “I”, “love”, and “you” is morphological [4]. LanguageTool have a morphological analysis function. We input text and click 「テキストにタグ付けをするを押す」.Then, we can be output as shown in the figure. There is more morphological analysis tool, but analysis result of these tool output different occasionally. Japanese is difficult to divide in morphological

because we don‘t write to divide words.

message: This tag output messages in LanguageTool. If we use suggestion of the sub element, the text will be emphasized and output in LanguageTool. I wrote how change messages later on too.

url: This tag output URL. We need to teach why is sentences wrong and where is evidence. If we don’t need to refer, we may delete this tag because this is the option tag.

example: We write example sentence in this tag. We need to write at least two. It is correct example and incorrect example. If we write correct example, we add attribute like type=”correct”. If we write incorrect example, we add attribute like type=”incorrect”. we must need ”marker” of the sub element. we write the positions of the errors in marker. If we input testrules.bat ja in command prompt, LanguageTool debug to input example. It is careful not to output example in LanguageTool.

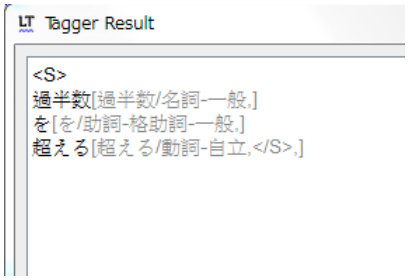

Figure 2: morphological analysis

4 Impovoed point

4.1 Add number

- 「lang-8」http://lang-8.com/

This site is very convenience.If we contribute texts which we want to correct in this site, we will be repaired from people who use the language of that text mainly. I researched how mistakes nonnative Japanese from this site.

Also, I thought that some mistakes is universalization. For example, mistakes as 「東京が行く」. Even if you 「東京」 change other counties, this sentence have mistakes. Therefore, I repaired XML that LanguageTool can find even other counties.

- 「言葉の誤用」http://starscafe.net/kotoba/misuse/

People collected frequent mistakes in this site. Mainly, these mistakes is used Japanese peoples. Moreover, a part of these mistakes is used at TV. I refer to this site too.

In addition, I research ’ra’-removed word,etc. I wrote those later on.

4.2 Sort category and add message and URL

I divided mistakes each category. I wrote explain of tags shown below. Moreover, I was output message and URL dependent on categorys because I was understand by users why is sentences wrong.

- grammatical mistake(message:文法ミスがあります。)

Grammar is misake. Example, 見るのをためらう→見るをためらう

- typographical error(message:誤字があります。)

Word is mistake. Example, スリルを味わわせるジェットコースター。→スリルを味あわせるジェットコースター。

- misconversion(message:誤変換です。)

Trasformation is mistake. Example, 奥義をきわめる→奥技をきわめる。

- collocation(message:連合関係がおかしいです。)

Connection of word is mistake. Example, 合いの手を入れる→合いの手を打つ

- overlap(message:重複しています。)

Word is overlap. Example, 満天の星→満天の星空

- ’ra’-removed word(message:ら抜き言葉があります。)

Example, ”ご飯を３杯くらいは余裕で「食べられる」→ご飯を３杯くらいは余裕で「食べれる」. ’ra’-removed word is phenomenon when we express a possible meaning. This is an error in linguistics but there is also an opinion which is linguistic evolution [5] . However, I improved LanguageTool to be able find ’ra’-removed word.

- ’re’-inserted word(message:レタス言葉があります。)

Example, ”ジュースを飲める→ジュースを飲めれる”. ’ra’-removed word is Phenomenon when we express the possible meaning.

- ’sa’-inserted word(message:さ入れ言葉がありま す。)

Example, ”読まさせていただく→読ませていただく”. A word processor cannot find ’sa’-inserted word unlike ’ra’-removed word [6] .However, this mistake is very. This mistake happen because we want to be polite.

- ’ra’-inserted word(message:ら入れ言葉がありま す。)

Example, ”切れる→切られる”. This mistake happen because we want to avoid ’ra’-removed word [7] . ’ra’-inserted word is not famous but this occasionally catch sight of too.

4.3 Add atiitbute of token tag

I added some attribute. LanguageTool was able to correspond to change words and a sentences. I used attribute and these features was shown below.

- regexp=”yes”

If hiragana change kanji, LanguageTool can find.
Example,
よい事（こと）だらけ→よく事尽くめ

```
<token>良い</token>
<token regexp="yes">事|こと</token>
<token>だらけ</token>
```

- skip=”10”

If words lie between sentences, LanguageTool can find. LanguageTool can indicate even if two other words is inserted when we write ”skip=”2””. I wrote skip=”10” except particular a case because I cannot know how many user insert words.
Example,
合いの手を（二人とも）打つ→合いの手を入れる

```
<token>合いの手</token>
<token skip="10">を</token>
<token>打つ</token>
```

私の彼女の机の中→私の彼女が持ってる机の中

```
<token skip="1">の</token>
<token skip="1">の</token>
<token>の</token>
```

- inflected=”yes”

If a verb change, LanguageTool can find.
Example,
愛苦しい（苦しく）→愛くるしい

```
<token>愛</token>
<token inflected="yes">苦しい</token>
```

- postag="SENT_START"

If we want to designate end of sentences, we use this tag.
Example,
。っかれたから寝る。→　。つかれたから寝る。

```
    <token postag="SENT_START" />
<or>
    <token>っ</token>
    <token>や</token>
    <token>ゆ</token>
    <token>よ</token>
    <token>ん</token>
</or>
```

- postag="SENT_END"

If we want to designate end of sentences, we use this tag.
Example,
したがって。→したがって、

```
    <token>したがって</token>
    <token postag="SENT_END" />
```

- postag=”(品詞名）.*”

If we want to designate a part of speech, we use this tag. For example, if we write postag=”動詞.*”, LanguageTool can find all verb. Also, We must add *postag*$_r$*egexp* = ”*yes*” attribute too.
Example,
描くのでは→描くでは

```
   <token postag="動
詞.*" postag_regexp="yes"/>
    <token>で</token>
    <token>は</token>
```

5 Conclusion

I increased seven hundred thirty numbers to be able to find mistake a kind from forty three numbers, and I changed messages considering that user can understand why is sentences wrong, and I repaired that LanguageTool can find even if words and sentences somewhat change.

I checked improvement factor of LanguageTool. I tried to utilize twitter. I input about one hundred fifty thousand numbers of characters from twitter. Therefore, LanguageTool of before improvement output, and in contrast, LanguageTool of after improvement could output sixteen mistakes.

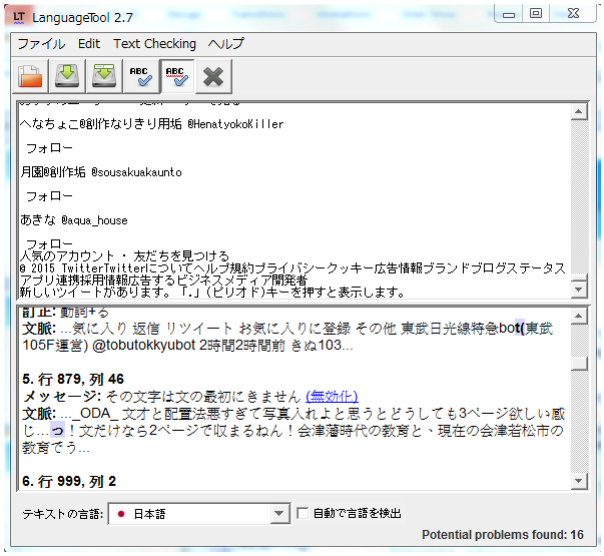

Figure 3: Potential problems found:16

6 In future

LanguageTool need to be further improved after this is broadly three.

First, we must increase to kind of grammar mistakes which can be extracted. Literal and typographical errors can find as such. However, grammar mistakes is few still more. The best feature of LanguageTool is which can find grammar mistakes. If grammar mistakes increase, LanguageTool will increase existence value.

Second, we must change messages to understandable and clear still more. Messages can output in view of category like「誤字があります」and「文法ミスがあります」. However, we want to improve so that LanguageTool can output minutely and intelligibly. Also, it is a problem that a correcting sentence output only one kind too. For example, if user input 「東京が行く」, LanguageTool output "「東京に行く」の間違いです". However, 「東京へ行く」 is the correcting sentence too. Therefore, this sentence must output in the future.

Third, we must be corresponded to various words. For example, 「成分は野菜です (材料は野菜です)」 can be found. However, 「成分は果物です (材料は果物です)」 cannot be found. I want to indicate all foods, if possible. If we can such a thing, LanguageTool will develop rapidly.

Acknowledgement

I would like to thank my supervisor Professor Maxim Mozgovoy. I thank everyone in my laboratory, too.

References

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

[2] 自然言語(日本語)処理. http://www.sist.ac.jp/ kanakubo/research/*natural*$_l$*anguage*$_p$*roces*

[3] Alex Homer and Alex Horner. *IE5, XML AND XSL Programmer's Reference*. Wrox Press Ltd., 1999.

[4] 形態素解析とは. http://e-words.jp/w/E5BDA2E6858BE7B4A0E8A7A3E69E90.html.

[5] あはは 敬語 ら抜き言葉. http://www3.kcn.ne.jp/ jarry/keig/c01c05.htm.

[6] あはは 敬語 さ入れ言葉. http://www3.kcn.ne.jp/ jarry/keig/c01c06.htm.

[7] ら入り言葉(られ入り言葉?):ら抜きの過剰訂正?ゆれる日本語の可能形活用もっと学ぼうニッポン:ブログ時代の日本語学習/ウェブリブログ. http://nihon.at.webry.info/200608/*article*$_1$1.*html*.