Evaluation of Generality of Inductive Learning for Preprocessing in Machine Translation

Yasuto Nagashima, Kenji Araki and Koji Tochinai

Division of Electronics and Information Engineering, Graduate School of Engineering, Hokkaido University
Kita-13, Nishi-8, Kita-ku, Sapporo, 060-8628, JAPAN
E-mail: {naga, araki, tochinai}@media.eng.hokudai.ac.jp

Abstract

There are many machine translation systems recently. However, the results of these machine translation systems include various errors on a selection of translated word, a dependency relation and so on. The purpose of our research is to correct these errors automatically and improve the translation accuracy by preprocessing.

This paper presents a method for preprocessing in machine translation system using inductive learning and results of evaluation experiment.

Keywords
preprocessing, machine translation, inductive learning, generality

1. Introduction

Recently, we have many occasions to see sentences written in other language by spreading of worldwide network. A machine translation system is very effective in such situation, and there are many machine translation systems actually. However, the translation results of these machine translation systems include various errors on a selection of translated word, a dependency relation and so on [1] [2].

Preprocessing is one of methods to improve translation accuracy, and in a sense, a kind of paraphrasing. There are few preceding researches on paraphrasing for the purpose of improving translation accuracy, and researches of paraphrasing to improve translation accuracy are specialized in specific machine translation systems [3] [4] [5] [6]. Therefore we cannot apply these methods to various machine translation systems we can use now. To solve these problems, we propose a method of preprocessing in machine translation system that can be applied to any machine translation systems.

Our method uses the inductive learning. The inductive learning is a method to obtain rules from some examples [7] [8]. Most of traditional methods for preprocessing in machine translation are Rule-based methods. However the Rule-based method cannot deal with various situations because of its limited rules. By using the inductive learning, our system can adapt to any machine translation systems because our system does not use any information of a specific machine translation system.

In this paper, we illustrate an outline of our system at first, and then show results of evaluation experiment.

2. Outline of Processing

Figure 1 shows a procedure of our system. First, morphological analysis is carried out on input sentence, and the result is input into the searching process. Our method uses only morphological information, because we can consider that errors of sentence structure analysis cause errors of translation. Then, in the searching process, the system checks if there are rules that can be applied to the input sentence in a transformation rule dictionary. If there are rules that can be applied, the rules are used on input sentence in the application process. We use two words "translate" and "transform" in this paper. We define "translate" as a change from a language into another language and "transform" as a change within a specific language. Next, the system translates the result of transformation using a machine translation system.

A user judges this translation result by comparing translation result of an input sentence, and updates positive translation degrees and negative translation degrees (see 2.3.1) of used rules in the feedback process. At the last, the system obtains translation...
rules in the learning process from the input sentence and its correct transformed sentence made by a person.

2.1 Searching Process

In this process, the system checks if there are rules that can be applied to the input sentence in a transformation rule dictionary in which obtained rules are registered. Each rule has application conditions, and rules that satisfy its conditions are applied to input sentence. When there are plural rules that satisfy their conditions and applied to the same place, our system chooses a rule that has the highest Rule Correct Rate (see 2.3.2). If rule correct rates are equal, we give priority to the latest rule.

2.2 Application Process

In this process, input sentence is transformed by transformation rules selected on the searching process. Transformation process consists of addition and deletion of words, and division of a sentence. If no rules satisfy their conditions, the input sentence is output without any transformation.

2.3 Feedback Process

This process changes application conditions of used rules. A user judges a translation result of output sentence from the application process by comparing translation result of an input sentence, and updates positive translation degree and negative translation degrees of used rules.

2.3.1 Positive and Negative Translation Degrees

Each rule has positive and negative translation degree. We call them PTD and NTD respectively. In the feedback process, when a user makes positive judge on a used rule, the PTD is added one. When a user makes negative judge on a used rule, the NTD is added one. And if there are no changes on a result of translation although rules are applied, the PTD and NTD of the rules are kept the state.

2.3.2 Rule Correct Rate and Application Condition

The rule correct rate is calculated as follows:

$$RCR = \frac{PTD}{PTD + NTD} \times 100 \quad [\%]$$

The application condition changes depending on the rule correct rate as follows:

(1) In the case the rule correct rate reaches 80% or more:

Correspondence of the difference part.

(2) In the case the rule correct rate is more than 50% and less than 80%.

Correspondence of the difference part and 2 parts of speech before and after difference part respectively.

(3) In the case the rule correct rate does not reach 50%.

Correspondence of the difference part, 2 parts of speech and 2 words before and after difference part respectively.
Input Sentence: Therefore/RB ./, our/PRP\$ proposed/VBN method/NN consists/VBZ of/IN the/DT analysis/NN and/CC learning/NN ./.

Correct Transformed Sentence: Therefore/RB ./, the/DT method/VBN we/PRP proposed/VBN consists/VBZ of/IN the/DT analysis/NN and/CC learning/NN ./.

Difference Part: “our proposed method” and “the method we proposed”

Obtained Rule: our proposed method ⇒ the method we proposed

Before Difference Part: Therefore/RB ./,
After Difference Part: consists/VBZ of/IN


Fig.2: Example of the Learning Process

2.4 Learning process

Figure 2 shows an example of the learning process. The system obtains translation rules from the pair of an input sentence and its correct transformed sentence that is made by a person. In this process, our system obtains translation rules by comparing a result of morphological analysis on input sentence with its correct transformed sentence and extracting difference parts between the sentences. We define a common part as a part where more than 3 words are the same consecutively, and a difference part as a part placed between common parts.

A rule consists of a difference parts, words and parts of speech before and behind a difference part, The PTD and the NTD. When a new rule was obtained, we set the PTD as 1 and the NTD as 1, not 1 and 0. It is because we consider a subordinate system must give priority from more correct answer the system can make to less wrong answer the system make. From this point of view, we chose figures mentioned above.

3. Correct Transformed Sentence

In this section, we present targets of preprocessing among some errors included in translation result, and a procedure for making a correct transformed sentence used in the inductive learning.

3.1 Targets of Preprocessing

There are some kinds of errors in a result of translation. In this paper, we limit the targets of preprocessing within 3 points as follow:

1. Errors on a selection of translated word
2. Errors on a dependency relation
3. Errors depend on a user

where, “3. Errors depend on a users” is an error that is occurred by a user when the user writes sentences1. About errors on naturalness of a translated sentence, it is out of target of this method. That is because about improving naturalness of a translated sentence, post processing is more effective than preprocessing. We are researching on it, too [9].

3.2 Procedure for Making Correct Transformed Sentence

A correct translated sentence is made through a procedure showed as follows:

Step 1: If there are any errors in translated

---

1Erroneous spelling such as “informations”, “knowledges” appears frequently. That is to say erroneous knowledge of a user.
sentence of an input sentence, make correct translated sentence by a person.

Step 2: Compare a translated sentence by a translation system with correct translated sentence made in step 1, and find wrong parts.

Step 3: Among errors on a selection of translated word, correct errors that are correct as concerns selection of part of speech.

   Step 3.1: In the correct sentence made in Step 1, translate the part correspond to the errors by MT system. And use the result as the first selection.

   Step 3.2: In the correct sentence made in Step 1, translate the part correspond to the errors by dictionary of MT system. If there are some words, use a less polysemical word.

Step 4: Correct the rest of errors on a selection of translated word.

   Step 4.1: The same as step 3.1.

   Step 4.2: The same as step 3.2.

Step 5: Correct errors on a dependency relation by dividing structure of an input sentence.

   Step 5.1: Add "," to each gaps of phrase around wrong part of input sentence one by one.

   Step 5.2: Divide an input sentence at each gaps of phrase around wrong part one by one.

   Step 5.3: In the case the structure of input sentence becomes wrong in the step 5.1, or step 5.2, modify it by supplementing some words around divided part.

Step 6: If there are any uncorrected parts on step 3, repeat from step 3.

Correct translated sentences are made by the above-mentioned process sentence by sentence.

4. Evaluation Experiment

We carried out an experiment to evaluate our system by using two English-Japanese machine translation systems that are developed by A and B companies. We call them A and B respectively.

4.1 Data and Procedure

In this experiment, we used 357 sentences taken from two theses on natural language field. We call them text A and text B respectively. And we used brill tagger for morphological analysis on input sentences [10]. The initial transformation rule dictionary was empty.
Table 1: The Change of State of Input Sentences (A)

<table>
<thead>
<tr>
<th>Item</th>
<th>Before Processing</th>
<th>After Processing</th>
<th>Change of State</th>
</tr>
</thead>
<tbody>
<tr>
<td>The Number of Correct Sentence</td>
<td>151</td>
<td>194</td>
<td>43</td>
</tr>
<tr>
<td>Translation Accuracy</td>
<td>42.3%</td>
<td>54.3%</td>
<td>12.0%</td>
</tr>
<tr>
<td>The Number of Untranslated Sentence</td>
<td>10</td>
<td>7</td>
<td>3</td>
</tr>
<tr>
<td>The Number of Wrong Transformed Part</td>
<td>-</td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td>The Number of Wrong Part</td>
<td>246</td>
<td>161</td>
<td>85</td>
</tr>
<tr>
<td>Selection of Translated Word</td>
<td>114</td>
<td>69</td>
<td>45</td>
</tr>
<tr>
<td>Dependency Relation</td>
<td>93</td>
<td>71</td>
<td>22</td>
</tr>
<tr>
<td>Depend on User</td>
<td>39</td>
<td>21</td>
<td>18</td>
</tr>
</tbody>
</table>

Table 2: The Change of State of Input Sentences (B)

<table>
<thead>
<tr>
<th>Item</th>
<th>Before Processing</th>
<th>After Processing</th>
<th>Change of State</th>
</tr>
</thead>
<tbody>
<tr>
<td>The Number of Correct Sentence</td>
<td>104</td>
<td>163</td>
<td>59</td>
</tr>
<tr>
<td>Translation Accuracy</td>
<td>29.0%</td>
<td>45.7%</td>
<td>16.5%</td>
</tr>
<tr>
<td>The Number of Untranslated Sentence</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>The Number of Wrong Transformed Part</td>
<td>-</td>
<td>11</td>
<td>11</td>
</tr>
<tr>
<td>The Number of Wrong Part</td>
<td>279</td>
<td>190</td>
<td>89</td>
</tr>
<tr>
<td>Selection of Translated Word</td>
<td>131</td>
<td>79</td>
<td>52</td>
</tr>
<tr>
<td>Dependency Relation</td>
<td>121</td>
<td>98</td>
<td>23</td>
</tr>
<tr>
<td>Depend on User</td>
<td>27</td>
<td>13</td>
<td>14</td>
</tr>
</tbody>
</table>

Table 3: Average Length of Difference Part

<table>
<thead>
<tr>
<th>Item</th>
<th>Input Sentence</th>
<th>Correct Transformed Sentence</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>A</td>
<td>B</td>
</tr>
<tr>
<td>Selection of Translated Word</td>
<td>2.28</td>
<td>1.92</td>
</tr>
<tr>
<td>Dependency Relation</td>
<td>4.66</td>
<td>6.72</td>
</tr>
<tr>
<td>Depend on User</td>
<td>1.22</td>
<td>1.25</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Item</th>
<th>A</th>
<th>B</th>
</tr>
</thead>
<tbody>
<tr>
<td>Selection of Translated Word</td>
<td>2.76</td>
<td>2.81</td>
</tr>
<tr>
<td>Dependency Relation</td>
<td>8.21</td>
<td>9.11</td>
</tr>
<tr>
<td>Depend on User</td>
<td>1.31</td>
<td>1.39</td>
</tr>
</tbody>
</table>

4.2 Result and Consideration

Figure 3 and Figure 4 show the changes of recall rates and precision rates of A and B respectively. The recall rate is the ratio of the number of an improved part to the number of part to be improved and the precision rate is the ratio of the number of an improved part to the number of a processed part. Since the translation rule dictionary is empty, all rates are low. However, the more the systems learn, the higher these rates rise. It means that our system is adapting to each systems and each text. Around 150 wrong parts in Figure 3 and 170 wrong parts in Figure 4, the recall rates and the precision rates decreased. It is caused by change of text. However, these rates rise again by adapting to new texts.

Total recall rates of A and B are 34.6% and 31.9% respectively. And the total precision rates of A and B, it is important on subordinate system, are 91.4% and 89.0% respectively. Therefore, we can confirm the generality of our system.

Table 1 and Table 2 show the change of state of input sentences between before processing and after processing of A and B respectively. According to these results, the errors on a selection of translated word and the errors depend on a user are improved relatively well, about 40% of the errors. However the errors on dependency relation are improved only 20% of the errors. The reason of these results is the frequency of application of rules.
Table 3 shows average length of difference part of each kind of errors. On the errors on a selection of word and errors depend on a user, the transformation rules are applied frequently since the lengths of difference parts of transformation rules are short. On the other hand, since the lengths of the difference parts of the transformation rules on the errors on dependency relation are long, the transformation rules are not applied very much. And even if the rules are applied, sometimes it is wrong application. We can consider we need more information from not only around difference parts but also whole a sentence about dependency relation.

5. Conclusions

In this paper, we proposed a method using the inductive learning for preprocessing in machine translation, and have showed results of experiment to evaluate the generality of our method. Our method does not use any information of specific machine translation system. And by using inductive learning, our system can deal with various situations of sentences. According to the results of experiment, we could improve 85 errors among 246 errors for A, and 89 errors among 279 errors for B in the translation result. And we could improve the translation accuracies from 42.3% to 54.3% for A, and from 29.1% to 45.7% for B. Moreover, the precision rates 91.4% for A and 89.0% for B were high. It means that our method can improve the translation accuracy of various translation systems without harmful side effect.

Future problems of our method are the introduction of more wide information to transformation rules and more experiments to show the adaptability of our method to various sentences. We have performed experiment using only 337 sentences taken from two theses. We need more sentences for more stable results.

References


