BILINGUAL CORPUS CLEANING FOCUSING ON TRANSLATION LITERALITY

Kenji Imamura and Eiichiro Sumita

ATR Spoken Language Translation Research Laboratories
2-2 Hikaridai, Seika-cho, Soraku-Gun, Kyoto, Japan

ABSTRACT

When we automatically acquire translation knowledge from a bilingual corpus, redundant rules are generated due to translation variety. To overcome this problem, we propose bilingual corpus cleaning based on translation literality. Word-level correspondence and phrase-level correspondence are applied as the criteria of literality. Using these criteria, a bilingual corpus was cleaned, and translation knowledge for a pattern-based MT system was acquired from the cleaned corpus. As a result, the translation quality of the MT was improved despite reductions in the the corpus size to about 81% and 87% by using word-level and phrase-level literality scores, respectively.

1. INTRODUCTION

Many machine translation (MT) systems based on enriching bilingual or multi-lingual corpora have been proposed. Such systems automatically learn their knowledge from the corpora [1, 2]. We have also proposed a pattern-based MT system that assumes automatic acquisition of translation knowledge from a large corpus. However, as the system learns a larger amount of knowledge, some deterioration was observed, caused either by bilingual sentences or the corpus itself.

Generally, when we discuss corpus-based natural language processing, the main focus is on the quantities of the corpora. On the other hand, the quality of corpora has not been given much attention. In previous research efforts on controlled languages [3], the quality of the language was discussed, but their targets were monolingual. However, they reported advantages, such as semantic understanding disambiguation, by restricting the vocabulary or grammar. In the case of a bilingual corpus, we can also expect ambiguities to be resolved or the capability of machine learning to be increased (i.e., needing a smaller corpus) by restricting translations.

In this paper, we propose using translation literality to evaluate the quality of bilingual sentences and cleaning the corpus based on literality. In order to demonstrate the effect of this approach, a partial corpus is created as a collection of high-literality sentences. Then, translation knowledge is automatically acquired from the cleaned corpus, and the translation quality of an MT system is improved as a result.

2. TRANSLATION VARIETY IN CORPORA

First, we describe the problems in acquiring translation knowledge by focusing on the variety of translations in corpora.

2.1. Multiple expressions

Generally speaking, a single source expression can be translated into multiple target expressions. For example, the Japanese sentence “Madogawa no seki wo onegai-shimasu” can be translated into English sentences such as “A window table, please,” “May I have a window table?,” or “I would like a table by the window.” These translations are all correct. However, if translation knowledge is acquired from a corpus that contains multiple expressions, redundant knowledge will be generated. Actually, in the case of the pattern-based MT system described in [4], multiple transfer rules are generated from multiple target sentences, although only one rule is necessary for translating a sentence. Unfortunately, redundant rules increase ambiguity or decrease translation speed.[5]

2.2. Context-dependent translation

Corpora usually contain translations that are not always correct. The most common phenomenon is context-dependent translation.

For instance, the determiner ‘the’ is not generally translated when English is translated into Japanese. However, when a human translator cannot semantically identify the following noun, a determinant modifier such as ‘watashi-no’ or ‘sono’ is supplied. These rules depend on the context, so if they are used in the wrong context, the translation will be wrong.

Note that the context-dependent translations are different from non-literal translations. Non-literal translations do
not depend on the context, and source expressions are almost always translated into specific target expressions. Thus, they are features of language pairs rather than problems of corpora. Knowledge acquisition from non-literal sentences has been studied in [6] as partially possible.

3. CORPUS CLEANING BASED ON TRANSLATION LITERALITY

In order to solve the problems described above, we propose corpus cleaning based on translation literality.

3.1. Translation literality

‘Literal translation’ is also commonly called ‘word-to-word translation’ to indicate word-level translation. Moreover, the transfer units in most MTs are words or phrases, and their minimal units are words. Therefore, translations that have high literality are more suitable for machine translation. However, words do not actually correspond one-by-one between different languages, especially between such different language families as English and Japanese. Accordingly, we expand the meaning of literality and define the following two literality scores.

3.1.1. Word-level literality

Translations that have sufficient word correspondences are judged to have high literality. We call this score ‘the word correspondence score’ (WCS) and represent it by the following equation.

\[ WCS = \frac{CW_s + CW_t}{W_s + W_t}, \] (1)

where \( CW_s \) and \( CW_t \) denote corresponding word numbers of source and target sentences, respectively. \( W_s \) and \( W_t \) denote total word numbers of source and target sentences, respectively. Thus, WCS indicates coverage of corresponding words.

Word correspondences can be automatically obtained from bilingual corpora. For example, [7, 8] propose statistical word alignment.

3.1.2. Phrase-level literality

In the case of translations between different language families, a word is often translated into a word of a different part-of-speech, e.g., the English verb ‘book’ is sometimes translated into the Japanese noun ‘yoyaku.’ This means that the syntactic structure is transformed between English and Japanese. Therefore, phrase correspondence is necessary for judgment of literality.

We utilize ‘the phrase score’ (PS) to judge the phrase-level literality, which is a structural similarity measure used in Hierarchical Phrase Alignment [6]. PS becomes higher when syntactic structures become similar between bilingual sentences, and it is represent by the following equation.

\[ PS = WL + PC, \] (2)

where \( WL \) denotes the corresponding number of words, and \( PC \) is calculated by counting phrases that satisfy the following two conditions.

1. Words in the phrase pair correspond with no deficiency and no excess.
2. The phrases are of the same syntactic category.

3.2. Corpus cleaning

Basically, bilingual sentences that have high literality scores as described in Section 3.1 are selected from the entire corpus, and only such sentences are used in the learning of translation knowledge. However, the literality scores become low in the case of non-literal translations. If we remove all low-score translations, we preclude transfer rules corresponding to non-literal translations, and the coverage of an MT system becomes narrower.

To avoid this problem, we first unify the source sentences. When a source sentence has multiple target sentences, we calculate the literality score and select the target sentence that has the maximum score. If multiple sentences have this maximum score, they are all selected.

This means that the coverage of transfer rules is maintained because all source sentences are utilized in the knowledge acquisition. Furthermore, only translations that are suitable for machine translation will remain from the entire human-acceptable translation set.

4. EXPERIMENT

4.1. Experimental setting

The target is English-to-Japanese translation in this experiment. The effect of corpus cleaning is evaluated by the translation quality of an MT system when translation knowledge is acquired from the cleaned corpus. The experimental setting was prepared as follows.

4.1.1. Bilingual corpus

We built a collection of Japanese sentences and their English translations based on expressions that are usually found in phrasebooks for foreign tourists (PB corpus; refer to [9].) We utilized a subset whose source and target sentence pairs were unified in advance. Table 1 shows the statistics of the PB corpus used in this experiment.
Table 1. Basic statistics of PB corpus

<table>
<thead>
<tr>
<th></th>
<th>English</th>
<th>Japanese</th>
</tr>
</thead>
<tbody>
<tr>
<td># Sentence Pairs</td>
<td>114,528</td>
<td></td>
</tr>
<tr>
<td># Unique Sentences</td>
<td>88,002</td>
<td>91,453</td>
</tr>
<tr>
<td># Total Words</td>
<td>714,012</td>
<td>803,519</td>
</tr>
<tr>
<td>Vocabulary</td>
<td>11,235</td>
<td>17,566</td>
</tr>
</tbody>
</table>

Fig. 1. Translation variety in PB corpus

Because this corpus is a collection of expressions that are usually found in phrasebooks, there are many translations where one source sentence has multiple target sentences (Figure 1). About 13% of the English sentences have multiple Japanese translations, and the maximum number of translation variety is 62.

4.1.2. Extraction of word correspondence

For high-frequency word correspondence that occurs more than ten times in the corpus, statistical word alignment is carried out by a similar method to [7]. For low-frequency words, a semantic dictionary is used, and correspondences are based on having the same semantic codes.

4.1.3. MT system

HPAT [4] is utilized for the experiment. This is a pattern-based MT system based on the syntactic transfer method, and its transfer rules (patterns) are acquired automatically by using Hierarchical Phrase Alignment [6] from a bilingual corpus.

4.1.4. Evaluation procedure

The corpus is cleaned by the three methods described below. The test set contains 510 sentences randomly selected and excluded from the PB corpus in advance. Subjective evaluation is carried out by a Japanese native speaker. The evaluation uses paired comparison. In other words, the source sentence, the MT result by method A, and the MT result by method B are shown to the evaluator at the same time. The evaluator judges which MT result is better or that they have the same quality.

1. Base
   All bilingual sentences are used for knowledge acquisition.

2. Word-level
   The corpus is cleaned by using WCS as described in Section 3.1.1.

3. Phrase-level
   The corpus is cleaned by using PS as described in Section 3.1.2.

4.2. Results of experiments

Table 2 shows the numbers of sentences contained in the cleaned corpora and acquired transfer rules. Table 3 compares the quality of translations among the three cleaning methods.

First, focusing on the sentence number after cleaning, corpus sizes were reduced to about 81% in the case of word-level cleaning and to about 87% in the case of phrase-level cleaning. The phrase-level reduction rate was worse than the word-level rate because there were many translations that have the same best PSs. On the other hand, WCS is normalized by the total word number, so it is possible to increase the reduction rate by normalizing the phrase score in a similar way.

Along with the reduction in corpus size, the transfer rule number was also reduced. However, as seen in Table 3, the translation qualities were improved in both cases (Base vs. Word-level and Base vs. Phrase-level). When PS is utilized for corpus cleaning, about 4.5% of the MT results were improved in comparison with the Base.

In comparing quality between Word-level and Phrase-level, the quality of Phrase-level was slightly better. Since the phrase score includes the number of word correspondences, the difference indicates the actual effect of the phrase-level literality. In this experiment, the gain rate shows that
<table>
<thead>
<tr>
<th>Method</th>
<th>(A) Won</th>
<th>(B) Won</th>
<th>Same Quality (Same Translation)</th>
<th>(B) Gain Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>(A) Base vs. (B) Word-level</td>
<td>29</td>
<td>48</td>
<td>433 (400)</td>
<td>+3.7%</td>
</tr>
<tr>
<td>(A) Base vs. (B) Phrase-level</td>
<td>28</td>
<td>51</td>
<td>431 (396)</td>
<td>+4.5%</td>
</tr>
<tr>
<td>(A) Word-level vs. (B) Phrase-level</td>
<td>17</td>
<td>22</td>
<td>471 (450)</td>
<td>+1.0%</td>
</tr>
</tbody>
</table>

5. CONCLUSIONS

In this paper, we proposed a corpus cleaning method based on translation literality. Word-level and phrase-level literality scores were defined and used to clean a corpus, and about 19% and 13%, respectively, of sentences were removed. In addition, the transfer rule sets for a pattern-based MT system were automatically acquired from the cleaned corpora. As a result, the translation quality was improved by applying this.

The corpus used here is high-density, in which many source sentences have multiple target sentences. We did not compare it with other corpora, but it is assumed that other corpora, such as newspaper articles, do not contain such a large number of multiple translations. However, if we enrich corpora without restriction, the problems discussed in this paper would inevitably become relevant. Therefore, this is an important problem for machine translation using large corpora.

Since the cleaning method described in this paper aims to increase the quality of corpora, it can be applied not only to pattern-based MTs but also to all corpus-based MTs such as statistical machine translation (SMT; [2, 10].) We are planning to apply this method to SMT and confirm its effect.

6. REFERENCES


