Machine Translation Using Bilingual Term Entries Extracted from Parallel Texts

Tatsuya Izuha

Corporate Research & Development Center, Toshiba Corporation
1, Komukai Toshiba-cho, Saiwai-ku, Kawasaki, 210-8582
JAPAN
tatsuya.izuha@toshiba.co.jp

Abstract

Patent summaries are machine-translated using bilingual term entries extracted from parallel texts for evaluation. The result shows that bilingual term entries extracted from 2,000 pairs of parallel texts which share a specific domain with the input texts introduce more improvements than a technical term dictionary with 38,000 entries which covers a broader domain. The result also shows that only 10 pairs of parallel texts found by similar document retrieval have comparable effects to the technical term dictionary, suggesting that parallel texts to be used do not need to be classified into fields prior to term extraction.

Keywords

machine translation, parallel text, bilingual lexicon, technical term, similar document retrieval

1. Introduction

Bilingual technical term lexicons are indispensable for better translation of technical documents. Since technical term dictionaries developed for machine translation systems generally cover broad domains such as "business" and "medicine," they often fail to contain those terms used only in some specific domains, and sometimes incorrectly translate those words according to more specific domains. However, technical term dictionaries for more specific domains are difficult to build and maintain because there exist so many technical domains and the number is constantly increasing.

In recent years, various methods have been proposed to build bilingual lexicons (semi-)automatically from parallel texts (Kumano & Hirakawa 1992; Melamed 1996; Smaja et al. 1996; Haruno et al. 1996; Kitamura & Matsumoto 1996). Many of these methods, however, require humans with expertise in the domain to check the final output entailing considerable cost on the part of humans. In this paper, bilingual term entries automatically extracted from parallel texts are directly used by a machine translation system to reduce such cost dramatically.

This paper is organized as follows: In the next section, the overview of our approach is presented. In Sections 3, 4 and 5, each component is described in detail. In Sections 6 and 7, evaluations of patent summaries and discussions on the results are given. Section 8 concludes the paper with final remarks.

2. Overview of our approach

Figure 1 shows the flow of Japanese-English translation using bilingual term entries extracted from parallel texts.

When a Japanese text to be translated into English is given to the system, the system retrieves parallel texts similar to the input text from the parallel text database. Then, it extracts bilingual term entries consulting the Japanese-English bilingual dictionary and chooses those entries that should be registered into the user dictionary used for machine translation. The system then translates the input text with the newly registered entries in the user dictionary. The resulting translation is compared with manual translation for evaluation.

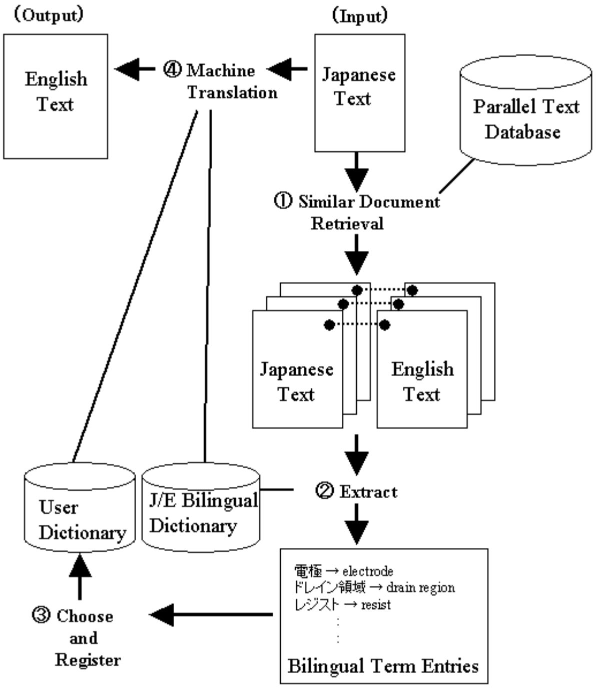

Figure 1 : Flow of machine translation using bilingual entries extracted from parallel texts

3. Similar document retrieval

Ideally, bilingual term entries should be extracted from texts which belong to the same domain as the input text. However, parallel texts classified into specific domains are not always available to the user. To deal with this problem, a similar document retrieval component is supplied in our approach. With the component, parallel texts similar to the input text can be found among the texts across various different domains.

This retrieval component is based on the vector space model (Salton et al. 1975) where each text is represented by means of feature vector describing the text. Similarity scores between two texts are expressed by the cosine value of two corresponding feature vectors which are generated from source language (Japanese) texts. Each dimension of a feature vector corresponds to a word which appears in the text and its value is weighted with TF-IDF scores. When a text is given to the component, similarity scores to each text in the database are calculated and a list of texts ranked by similarity scores is generated.

If texts in the database have been classified into specific domains beforehand, similar document retrieval is not necessarily needed. Consider the case of patent specifications are classified according to the International Patent Classification (IPC)[1], a hierarchical classification system comprising sections, classes, subclasses, and groups. Its current (seventh) version consists of almost 69,000 groups.

4. Extracting bilingual term entries from parallel texts

To extract bilingual term entries, we have adopted the method proposed by Kumano & Hirakawa(1992) characterized by the use of both statistical information and linguistic information. Below are the steps of their method illustrated in Figure 2:

(1) Both Japanese and English texts are split into sentences.
(2) Each Japanese sentence is mapped into English sentences using an MT bilingual dictionary.
(3) A selected group of terms (JW) are extracted from Japanese sentences.
(4) Word n-grams are generated as translation candidates (EW_i) from English sentences corresponding to the Japanese sentences which contain JW.
(5) Translation likelihood (TL) of each EW_i is calculated.

In Step (3), single nouns, compound nouns and unknown words are extracted as terms. Kumano & Hirakawa extracted only the last two because these two serve their purpose to build a technical term dictionary. Here we have included single nouns for more precise translation in light of the text domain.

In Step (5), translation likelihood based on linguistic information (TLL) and translation likelihood based on statistical information (TLS) are calculated. TL is the weighted average of TLL and TLS.

TLL of EW_i is calculated based on the following two hypotheses:

(Hypothesis 1)
If the number of element words in EW_i is close to the number of element words in JW, EW_i is likely to be the translation of JW.

(Hypothesis 2)
EW_i with more element word translation correspondences with JW is likely to the translation of JW

Element word correspondences in (Hypothesis 2) are judged using an MT bilingual dictionary. When JW and EW_i contain k element words, l element words, respectively, and x words are lexically correspondent between JW and EW_i, TLL(JW, EW_i) is given as below:

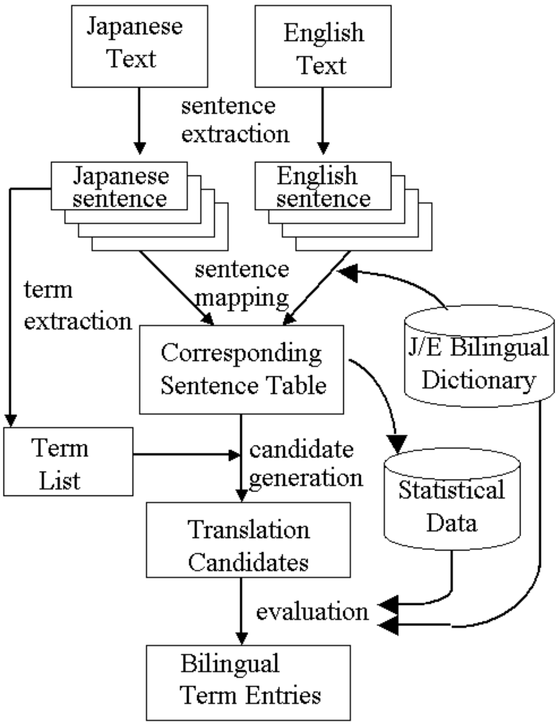

Figure 2 : Flow of extraction of bilingual entries from parallel texts

$$TLL(JW, EW_i) = \frac{P \times \min(k,l) + \alpha P \times x}{P \times k + \alpha P \times k}$$

where P is the unit score, and α (>0) is a coefficient.

In the equation above, the first term of the numerator, P×min(k,l), is EW_i 's score based on (Hypothesis 1). The second term, $\alpha P \times x$, is EW_i 's score based on (Hypothesis 2). The denominator is the score of a virtual translation which best satisfies the conditions in both (Hypothesis 1) and (Hypothesis 2).

When JW appears in m sentences, and EW_i appears in n corresponding sentences, TLS(JW, EW_i) is given as below:

$$TLS(JW, EW_i) = \frac{n}{m}$$

5. Choosing bilingual term entries to be used in machine translation

Bilingual entries extracted from parallel texts are registered into the user dictionary of the machine translation system and henceforth reflected in the translation outputs.

In the procedure described in Section 4, several translation candidates (EW_i) with TL values are obtained for each Japanese term (JW). Out of these candidates only one is chosen and registered into the user dictionary, or all

[1] http://www.wipo.int/classifications/en/index.html

candidates are discarded if the conditions are not met. The simplest way is to choose the candidate with the largest TL above a prespecified threshold. However, this may allow inappropriate entries to be registered when the difference between the largest TL and the second largest TL is small.

To deal with this problem, we register the candidate with the largest TL which meets the following condition.

$$(Condition\,1) \cap$$
$$((Condition\,2-1) \cup (Condition\,2-2))$$

(Condition 1) :

$$TL \geq \beta$$

(Condition 2-1) :

(TLL of the candidate with the largest TL) >
(TLL of the candidate with the second largest TL)

(Condition 2-2) :

$$\ln\frac{n_1}{n_2} - Z_{1-\alpha}\sqrt{\frac{1}{n_1}+\frac{1}{n_2}} \geq \theta$$

The value of β in (Condition 1) is determined as 0.5 from the results of preliminary experiments.

Since TLL is more reliable than TLS, the candidate with larger TLL is chosen without taking into account the difference in (Condition 2-1).

(Condition 2-2) is the criterion proposed by Dagan et al. (1994). Let p_i be the probability based on statistical information that EW_i is the translation of JW, and n_i ($1 \leq i \leq k$) be the number of the sentences in which EW_i appears. The estimator p_i for p_i is

$$\hat{p}_i = \frac{n_i}{\sum_{j=1}^{k} n_j}$$

In the discussion below, n_i is numbered in decreasing order. EW_i, p_i and p_i are numbered in correspondence with n_i. To avoid choosing EW_1 when the difference between p_1 and p_2 is small, the odds ratio $p_1/p_2\,(=n_1/n_2)$ must exceed a prespecified threshold. Furthermore, the threshold should be large when the counts (n_1, n_2) are small and decreases as the counts increase. By use of confidence intervals and the log odds ratio $\ln(p_1/p_2)$, (Condition 2-2) is obtained, where $Z_{1-\alpha}$ is the size of the 100(1-α)% confidence interval of standard normal distribution, and θ is a constant to be determined empirically. In this paper, α =0.1 ($Z_{0.9}$=1.282) and the value of θ is determined as 0.2.

Two examples are presented in Figure 3, where the choice depends on (Condition 2-2).

In the first example in Figure 3, the candidate "field strength" does not meet (Condition 2-2) and nothing is registered into the user dictionary.

【電界/強度】 [m=6]

[1] "field strength"
TL=0.67 (*TLL*=0.67 [x=1], *TLS*=0.67 [n=4])
[2] "field intensity"
TL=0.58 (*TLL*=0.67 [x=1], *TLS*=0.33 [n=2])

【酸化/膜】 [m=26]

[1] "oxide film"
TL=0.76 (*TLL*=0.67 [x=1], *TLS* =0.96 [n=25])
[2] "nitride film"
TL=0.60 (*TLL* =0.67 [x=1], *TLS* =0.46 [n=12])

Figure 3 : Examples of bilingual term entries extracted from parallel texts

Japanese terms	**Original translations**	**Improved translations**
ゲート電極 領域	gate 電極 domain	gate electrode region
ソース オーミック コンタクト層	sauce オーミック contact layer	source ohmic contact layer
電子線 レジスト	electronic line register strike	electron beam resist
窒化水素	nitriding hydrogen	hydrogen nitride
最上層	best layer	uppermost layer

Figure 4 : Examples of improved translations of baseNPs

$$\ln\frac{n_1}{n_2} - Z_{0.9}\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$$

$$= \ln\frac{4}{2} - 1.282\sqrt{\frac{1}{4}+\frac{1}{2}} = -0.268 < \theta\left(=0.2\right)$$

In contrast, in the second example in Figure 3, the translation candidate "oxide film" meets (Condition 2-2) and hence is registered into the user dictionary.

$$\ln\frac{25}{12}-1.282\sqrt{\frac{1}{25}+\frac{1}{12}}=0.383>\theta\,(=0.2)$$

6. Evaluation

To evaluate our proposed method, 20 patent summaries in Japanese are translated into English by a Japanese-English MT system, whose standard dictionary contains about 250,000 entries. The parallel text database contains 2,000 pairs of bilingual patent summary texts. All texts in both sets belong to the IPC main group H01L 21/00, which is related to semiconductor devices.

We have limited the targets of evaluation to baseNPs (Ramshaw & Marcus; 1995) in the translated texts. The occurrences of baseNPs in the 20 texts totalling 666 are judged either improved or degraded in light of manual translation results. The baseline for evaluation is the translation results without bilingual entries extraction. Note that there may be cases where translations which are different from manual ones but can be considered correct are judged as degraded.

Results of the evaluation are given in Table. 1. The columns labelled "Top 10 Texts" and "Top 100 Texts" show the results where bilingual entries are extracted from top 10 and top 100 pairs of parallel texts found by similar document retrieval respectively. "All Texts" means that bilingual entries are extracted from all the parallel texts contained in the database without using similar document retrieval. The column labelled "Tech Dic" gives the results where a technical term dictionary is used instead of bilingual entries extracted from parallel texts. The technical term dictionary covers electronics and electric engineering, and contains about 38,000 entries. The column labelled "Tech Dic + All Texts" shows the results where both the technical term dictionary and bilingual entries extracted from all the parallel texts are used.

Examples of improved translations are shown in Figure 4.

7. Discussion

More bilingual term entries with higher utility and reliability could be extracted from more pairs of parallel texts which share the specific domain. "All Texts" case is under ideal conditions from this point of view. Hence discussions below will focus on this case.

First, comparison between "All Texts" and "Tech Dic" demonstrates that bilingual term entries extracted from parallel texts introduce more improvements than the technical term dictionary. Although the number of degraded translations is larger with the former, those terms used only in some specific domains are extracted from parallel texts. The last two examples of improved translations in Figure 4 cannot be obtained by the technical term dictionary. Additionally, those entries whish lead to domain dependent translations are extracted from parallel texts. For example, a Japanese word “RYOUIKI” is translated into “domain” by the MT system without extra information though it should be translated into “region” in the domain of semiconductor devices. Such information can be obtained from parallel texts. Using both resources ((a)+(b)) introduces more improvements.

	Top 10 Texts	Top 100 Texts	All Texts (a)	Tech Dic (b)	(a) + (b)
Improved (1)	133	157	177	138	192
Degraded (2)	7	18	25	8	22
(1)-(2)	126	139	152	130	170

Table 1 : Improvement with bilingual term entries extracted from parallel texts

	1,500 pairs	2,000 pairs	2,500 pairs
Improved (a)	157	177	181
Degraded (b)	17	25	17
(a) – (b)	140	152	164

Table 2 : Relation between database size and improvement

Comparison between "Top 10 Texts" and "Tech Dic" suggests that comparable improvements to the technical term dictionary can be brought about with a small number of parallel texts found by similar document retrieval. This is because a few high-frequency words representing key concepts of the domain are extracted from a small number of parallel texts. For example, 75 out of 177 improved baseNPs with "All Texts" contain either "electrode", "source", "drain", or "region" as element words, while 67 out of 133 improved baseNPs with "Top 10 Texts" contain one of the four.

Furthermore, to examine relation between the database size and improvement, the same set of 20 texts are translated with bilingual term entries extracted from all texts in the databases of various sizes, as in Table 2.

Although the number of improvements increases very slightly in line with the increase of the database size because of a few predominant words, it seems that a larger size of database leads to more improvements.

It is important to examine the cases where translations are degraded with bilingual term entries extraction. We considered the case of “All Texts”. Aside from the 6 occurrences which can be judged correct but on the surface different from the manual translations, degradations in the 19 occurrences are classified into following four types.

(1) The head word of a NP is not a noun. 7
(2) (Hypothesis 2) does not hold true. 5
(3) The correct candidate is not generated. 4
(4) Errors in word segmentation of JW. 3

First, some Japanese nouns with specific postpositions are translated into attributive participles in English. From such parallel texts, a participle can be chosen as a translation of a Japanese noun. Such an error occurs because word n-grams are used as translation candidates and can be prevented by analyzing the morphological structure of candidates.

Second, (Hypothesis 2) does not always hold true. Especially when the correct English candidate contains more element words than the Japanese term, not only linguistic information but also statistical information prefers candidates containing fewer element words. To avoid such an error, (Condition 2-1) should be applied more carefully considering candidates with the same frequency containing more element words.

Third, current implementation filters out word n-grams which begin with a preposition. For example, "OFF current" is not generated as a translation candidate. A minor modification of implementation can prevent this error.

Fourth, word segmentation errors of Japanese terms sometimes lead to extracting inappropriate entries. Although a certain number of word segmentation errors are inevitable, discarding candidates with low frequency would prevent many of such inappropriate entries from being extracted .

8. Conclusion

Patent summaries are machine-translated using bilingual term entries extracted from parallel texts for evaluation. The result shows that bilingual term entries extracted from 2,000 pairs of parallel texts which share a specific domain with the input text introduce more improvements than a technical term dictionary with 3,800 entries which covers a broader domain. The result also shows that only 10 pairs of parallel texts found by similar document retrieval have comparable effects to the technical term dictionary, suggesting that parallel texts to be used do not need to be classified into fields prior to term extraction.

Further evaluations with more texts from various domains are needed to examine possible causes of degradations. At the same time, analysis to enhance improvements is also needed.

References

Dagan, I. & Itai, A. (1994). Word Sense Disambiguation using a Second Language Monolingual Corpus. omputational Linguistics, 20(4), .563--596.

Haruno, M., Ikehara, S. & Yamazaki, T. (1996). Learning bilingual collocations by word-level sorting. COLING 96 (pp. 525--530)

Kitamura, M. & Matsumoto, Y. (1996). Automatic Extraction of Translation Patterns in Parallel Corpora. IPSJ journal, 38(4), 727—736.

Kumano, A. & Hirakawa, H. (1992). Building an MT Dictionary from Parallel Texts Based on Linguistic and Statistical Information. COLING 92 (pp. 76—81)

Melamed, I.D. (1996). Automatic construction of clean broad-coverage translation lexicons. In Proceedings of 2nd Conference of Association for Machine Translation in the Americas (pp. 125--134).

Ramshaw, L.A. & Marcus, M.P. (1995). Text chunking using transformation-based learning. In Proceedings of the 3rd Workshop on Very Large Corpora (pp. 88—94).

Salton, G., Wong, A. & Yang, C. (1975). A Vector Space Model for Information Retrieval. Communications of the ACM, 18(11), 613—620.

Smadja, F., McKeown, K.R. & Hatzivassiloglou, V. (1996). Translation collocations for bilingual lexicons: a statistical approach. Computational Linguistics, 22(1), 1—38.