BUILDING AN MT DICTIONARY FROM PARALLEL TEXTS BASED ON LINGUISTIC AND STATISTICAL INFORMATION

Akira Kumano Hideki Hirakawa

R & D Center, Toshiba Corporation
1, Komukai Toshiba-cho, Saiwai-ku, Kawasaki, 210, JAPAN
{kmn,hirakawa}@isl.rdc.toshiba.co.jp

Abstract

A method for generating a machine translation (MT) dictionary from parallel texts is described. This method utilizes both statistical information and linguistic information to obtain corresponding words or phrases in parallel texts. By combining these two types of information, translation pairs which cannot be obtained by a linguistic-based method can be extracted. Over 70% accurate translations of compound nouns and over 50% of unknown words are obtained as the first candidate from small Japanese/English parallel texts containing severe distortions.

1 INTRODUCTION

Parallel texts (corpora) are useful resources for acquiring a variety of linguistic knowledge (Dangan, 1991; Matsumoto, 1993), especially for machine translation systems which inherently require customizations. Translation dictionaries are, needless to say, the most basic and powerful knowledge source for improving and customizing translation systems. Our research interest lies in automatic generation of translation dictionaries from parallel texts. In this perspective, finding corresponding words or phrases in bilingual texts will be the fundamental factor for accurate translation.

Statistics-based processing has proven to be very powerful for aligning sentences and words in parallel corpora (Brown, 1991; Gale, 1993; Chen, 1993). Kupiec proposes an algorithm for finding noun phrases in bilingual corpora (Kupiec, 1993). In this algorithm, noun-phrase candidates are extracted from tagged and aligned parallel texts using a noun phrase recognizer and the correspondences of these noun phrases are calculated based on the EM algorithm. Accuracy of around 90% has been attained for the hundred highest ranking correspondences. Statistics-based processing is effective when a relatively large amount of parallel texts is available, i.e. when high frequencies are obtained.

On the other hand, existing linguistic knowledge can be used for finding corresponding words or phrases in parallel texts. For example, possible target expressions for a source expression provided by a translation system (linguistic knowledge source) can be a key in searching the corresponding expressions in a corpus (Nogami, 1991; Katoh, 1993). Yamamoto (1993) proposes a method for generating a translation dictionary from Japanese/English parallel texts. In this method, English and Japanese compound noun phrases are extracted from parallel texts and their correspondences are searched by matching their possible translations generated by the existing translation dictionary. However, acquirable noun phrases are limited by the linguistic generative power of the translation dictionary. Furthermore, this method utilizes no sentence alignment information which can reduce errors in finding noun phrase correspondences.

This paper proposes a new method for generating an MT dictionary from parallel texts. It utilizes both statistical and linguistic information to obtain corresponding words or phrases in parallel texts. By combining these two types of information, translation pairs which cannot be obtained by the above linguistic-based method can be extracted, and a highly accurate translation dictionary is generated from relatively small parallel texts.

2 APPROACH TO BUILDING AN MT DICTIONARY

Our goal in building an MT dictionary from parallel texts is to develop a robust method which enables highly accurate extraction of translation pairs from a relatively small amount of parallel texts as well as from parallel texts containing severe distortions.

In real-world applications, generally it is extremely difficult especially for MT users to obtain a large amount of high quality parallel texts of one specific domain. If source and target languages do not belong to the same linguistic family, like Japanese and English, the situation becomes grave.

As one typical example of MT dictionary compilation, we have selected Japanese and English patent documents which contain many state-of-the-art technical terms. Although these documents are not cul-

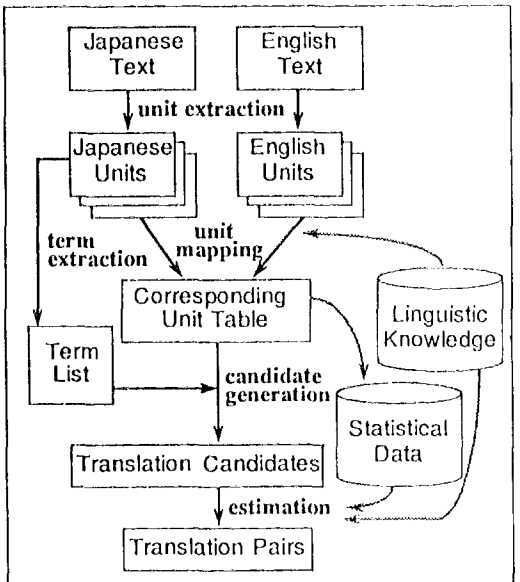

Fig. 1: Flow of building an MT dictionary from parallel texts

turally biased, in many cases, the organization between Japanese and English greatly differs and extensive changes are made in translating from Japanese to English text and vice versa. Hence, the difficulty of word extraction from patents.

To solve this problem, we explored the appropriate integration method considering the use of linguistic information and statistical information to this end. Linguistic information is useful in making an intelligent judgment about correspondence between two languages even from partial texts because of its lexical, syntactic, and semantic knowledge; statistical information is characterized by its robustness against noise because it can transform many actual examples into an abstract form.

Below is the flow of our method illustrated in Fig. 1:

(1) Unit Extraction:
Parts of documents ("units") are extracted from both Japanese and English texts.

(2) Unit Mapping:
Each Japanese unit is mapped into English units.

(3) Term Extraction:
Japanese term candidates are extracted by the NP recognizer.

(4) Translation Candidate Generation:
English translation candidates for Japanese terms are extracted from English units.

(5) English Translation Estimation:
The translation candidates are evaluated to obtain the best one.

The subsequent sections show the details of each processing.

3 FORMING UNIT CORRESPONDENCES

The plausible hypothesis that parallel sentences contain corresponding linguistic expressions is the major premise in Kupiec (1993). This type of information should be widely used. The problem is that the alignment method based on the sentence bead model (Brown, 1991) is not applicable to patent documents due to their severe distortions in document structures and sentence correspondences. Consequently, we have introduced a concept called "unit" which corresponds to a part of sentence and adopted a new method to extract corresponding units by using linguistic knowledge as a primary source of information.

3.1 Extraction of Units

First, units are extracted from parallel texts. The unit corresponds to sentences or phrases in the text. Terms which should be extracted can be found within a unit. The rest of words in the unit is called contextual information for the extracted term. The size of units determines the effectiveness of the succeeding unit mapping process. For example, if we set noun phrases (entry words in a dictionary) as a unit, no contextual information is available, and thus the probability that corresponding relations hold decreases. In our present implementation, we set sentences as a unit for the first approximation.

3.2 Mapping of Units

Next, the unit mapping process creates a corresponding unit table from Japanese and English units. This table stores the correspondence relationship between units and its likelihood. The likelihood is calculated based on the linguistic information in an MT bilingual dictionary.

Our unit mapping algorithm is given below:

(1) Let J be a set of all content words in the Japanese unit JU. (m is the number of words)
$$J = \{ J_1, J_2, \ldots J_m \}$$

(2) Let E be a set of all content words in the English unit EU. (n is the number of words)
$$E = \{ E_1, E_2, \ldots E_n \}$$

(3) x is the number of J_i's whose translation candi-

date list includes some E_j in E.

(4) y is the number of E_j's which is included in the translation candidate list of some J_i in J.

(5) The correspondence likelihood CL is given by $CL(JU, EU) = \frac{x+y}{m+n}$

For each JU, M (currently 3) English units with the highest CL(JU, EU) are stored in the corresponding unit table.

4 GENERATING TRANSLATION CANDIDATES

4.1 Extraction of Japanese Terms

Errors in the extraction of terms and phrases from parallel texts eventually lead to a failure in acquiring the correct term/phrase correspondences. In Kupiec (1993) and Yamamoto (1993), term and phrase extraction is applied to both of parallel texts. In contrast, we extract from units only Japanese terms, thereby reducing the errors caused by term/phrase recognizer. Japanese NP's can be recognized more accurately than English NP's because Japanese has considerably less multi-category words.

In the current implementation, the following two types of term candidates are extracted by the NP recognizer:

(A) Compound nouns (including verbal nouns)
Examples: "オープンビット線方式"
(=*open bit line configuration*)
"最小加工寸法"
(=*minimum featuring size*)

(B) Unknown words (nouns, verbal nouns)
Examples: "積層する" (=*to laminate, to form*)
"ポリッシング" (=*polishing*)

Our NP recognizer utilizes the sentence analyzer of a practical MT system. The word dictionary includes approximately 70,000 Japanese entries.

4.2 Finding Translation Candidates

Generation of English translation candidates for a Japanese term is essentially based on the following hypothesis:

Hypothesis 1

The English translation of an extracted term in a Japanese unit is contained in the English corresponding unit.

Now an arbitrary word sequence in corresponding units can be a translation candidate of the Japanese term. We extract English translation candidates in two steps:

Step 1: Select English corresponding units.
Step 2: Extract n-gram data from the units.

Step 1:

When the extracted term appears in N Japanese units, $N \times M$ English units will be stored in the corresponding unit table with their correspondence likelihood. The N highest corresponding units within $N \times M$ combinations are extracted. When N is less than M, the M highest combinations are selected.

Step 2:

Suppose that the correct English translation of the Japanese term JW is EW, and that the number of Japanese units in which JW appears is FJU(JW) (= N). From Hypothesis 1 that the translation is contained in the corresponding units EU_1, EU_2,..., $EU_{FJU(JW)}$, EW would be a word sequence which often appears in corresponding units. In order to get such EW, we use n-gram data.

The frequency of each n-gram ($1 \leq n \leq 2 \times$ (the number of component words in JW)) data in FJU(JW) English units is calculated and then EW candidates are ranked by the frequency as EWC_1, EWC_2,... EWC_j. Because EWC with a low frequency in the corresponding units is unlikely to be the correct translation, the data with a frequency less than $\frac{FJU(JW)}{4}$ are heuristically excluded from the candidates. The data containing *be* verb and the data which starts or ends with a preposition or an article are also excluded from the candidates.

5 ESTIMATING ENGLISH TRANSLATIONS

The translation likelihood (TL) of one translation candidate EWC_i for the term JW is defined as:

$$TL(JW, EWC_i) = F(TLS(JW, EWC_i), TLL(JW, EWC_i))$$

where TLS(JW, EWC_i) is "Translation Likelihood based on Statistical information," and TLL(JW, EWC_i) "Translation Likelihood based on Linguistic information."

5.1 Statistical Information

TLS(JW, EWC_i) is the frequency score based on the statistical information from Hypothesis 1 that a word which appears as often in the corresponding units as JW in Japanese units is more likely to be EW. It is quantitatively defined as the probability in which the translation candidate appears in the corresponding units. That is,

$$TLS(JW, EWC_i) = \frac{FEU(EWC_i)}{FJU(JW)}$$

where $FEU(EWC_i)$ is the number of corresponding units in which EWC_i appears.

5.2 Linguistic Information

$TLL(JW, EWC_i)$ is the word similarity score based on the accuracy of the correspondence term JW and the translation candidate EWC_i obtained by using linguistic information in the MT bilingual dictionary. Suppose one translation candidate of term $JW=wj_1, wj_2,... wj_k$ is $EWC_i=we_1, we_2,... we_l$. Then we use the following hypothesis.

Hypothesis 2

(a) If the length of EWC_i is close to the length of JW, JW and EWC_i are likely to correspond each other.

(b) JW and EWC_i with more word translation correspondences are likely to correspond each other.

Under this hypothesis, the following correspondence relation (1) is the best. Term JW and translation candidate EWC_i have the same length $k(=l)$, and all of their component words correspond in the dictionary. $wj_i \Rightarrow we_i$ indicates that we_i is included in wj_i's translation candidates in the MT bilingual dictionary.

(1) $wj_1 \Rightarrow we_1, wj_2 \Rightarrow we_2, ..., wj_k \Rightarrow we_k$

More generally, the relation of each word (wj) in term JW and each word (we) in translation candidate EWC_i is classified into the following four classes:

i) $wj \Rightarrow we$
ii) $wj \rightarrow we$
iii) $wj \rightarrow \phi$
iv) $\phi \rightarrow we$ (ϕ indicates *no word*)

ii) shows a pair whose correspondence is not described in the bilingual dictionary. iii) and iv) indicate that the corresponding word for wj or we is missing. In iii), JW is longer than EWC_i; and vice versa in iv).

In order to estimate correspondence between JW and EWC_i, i) and ii) are scored by similarity to the virtual translation which holds the relation (1). When the number of words is the same, score Q (constant) is given. αQ ($\alpha>0$) is added to Q when there is a translation relation to reflect higher reliability of i). Therefore, $Q+\alpha Q=(1+\alpha)Q$ is given to the word pair of i), and Q to the word pair of ii).

Now since we disregard the word order of a term, JW and EWC_i are represented as sets of words:

$$JW = wj_1, wj_2,... wj_k = \{wj_1, wj_2,... wj_k\}$$
$$EWC_i = we_1, we_2,... we_l = \{we_1, we_2,... we_l\}$$

The number of words with a lexical correspondence relation in wj and we, the number of words in wj without a relation and the number of words in we without a relation are counted as x, y, z respectively. That is, $x + y = k$ and $x + z = l$.

$TLL(JW, EWC_i)$ is given as the ratio of the score of the virtual translation to the score of EWC_i.

When $y \geq z$,

$$TLL(JW, EWC_i) = \frac{x(1+\alpha)Q + zQ}{(x+y)(1+\alpha)Q}$$

Otherwise,

$$TLL(JW, EWC_i) = \frac{x(1+\alpha)Q + yQ - (z-y)Q}{(x+y)(1+\alpha)Q}$$

Thus,

$$TLL(JW, EWC_i) = \begin{cases} \dfrac{x(1+\alpha)+z}{(x+y)(1+\alpha)} & (y \geq z) \\ \dfrac{x(1+\alpha)+2y-z}{(x+y)(1+\alpha)} & (\text{otherwise}) \end{cases}$$

By definition, $TLL(JW, EWC_i) < 1$. The value of α is determined as 2 by evaluating sample translation pairs.

Followings are the TLL's of three EWC's for JW:オープンビット線方式 which consists of four component words (k=4); "オープン(=*open*)," "ビット(=*bit*)," "線(=*line*)," and "方式(=*method, process*)."

bit line configuration
x=2, y=2, z=1 $\quad \therefore TLL = (2\times3+1)/4\times3 = 0.58$
open bit line
x=3, y=1, z=0 $\quad \therefore TLL = (3\times3)/4\times3 = 0.75$
open bit line configuration
x=3, y=1, z=1 $\quad \therefore TLL = (3\times3+1)/4\times3 = 0.83$

5.3 Combination of Statistical and Linguistic Information

We define the translation likelihood $TL(JW, EWC_i)$ as below:

$$TL(JW, EWC_i) = \frac{m\, TLS(JW, EWCi) + n\, TLL(JW, EWCi)}{m+n}$$

Examining the value with the ratio n/m constant, a low value of $TLS(JW, EWC_i)$ ill affects the total score, especially when the frequency

FJU(JW) is 5 or less. This shows that TLS(JW, EWC_i) should be much weighed for JW's which appear often, but not for JW's with a low frequency. Therefore we tentatively define $\beta = n/m$ as a function of frequency FJU(JW), because β should be higher when FJU(JW) is low.

$$\beta = G(\mathrm{FJU(JW)}) = \frac{p}{\{\mathrm{FJU(JW)}\}^q - r} + s$$

where r is a possible minimum frequency, and s is limit of β as the word frequency is high enough. Values p=4, q=1, r=1, and s=0.5 are used in the following experiments. By introducing β, F is rewritten as:

$$F(\mathrm{TLS(JW, EWC}_i), \mathrm{TLL(JW, EWC}_i)) = \frac{\mathrm{TLS(JW, EWC}_i) + \beta\ \mathrm{TLL(JW, EWC}_i)}{1 + \beta}$$

In case $\{\mathrm{FJU(JW)}\}^q$ is equal to or less than r, β is meaningless. For such JW's, TL(JW, EWC_i) is redefined as simply:

$$\mathrm{TL(JW, EWC}_i) = \mathrm{TLL(JW, EWC}_i).$$

Finally the translation candidate EWC_i with the largest value of TL(JW, EWC_i) is assumed to be the correct English translation.

Table 1 shows the translation candidates for JW: オーブンビット線方式 with the best three TL's. Its frequency in Japanese text is FJU(JW) = 19 ($\beta = \frac{4}{19-1} + 0.5 = 0.72$). Consequently, the correct translation EWC_3, *open bit line configuration*, is obtained.

Table 1: Estimation of English translation

EWC_i	FEU	TLS	TLL	TL
bit line configuration	19	1.00	0.58	0.82
open bit line	18	0.95	0.75	0.86
open bit line configuration	18	0.95	0.83	0.90

6 EVALUATION AND DISCUSSION

To evaluate this method, we have estimated English translations of Japanese terms in seven parallel texts (Japanese specifications of patents on semiconductors and their English translations by human translators) and compared the translations with the correct data given by experts in building an MT dictionary. The size of a Japanese text is 7,508 to 26, 927 characters in 127 to 616 sentences; 99,286 characters in 2,148 sentences in total. Examples of correct translation pairs estimated with the highest TL are listed in Fig. 2.

Compound nouns:

最小加工寸法	*minimum featuring size*
素子分離領域	*element separation region*
オープンビット線方式	*open bit line configuration*
カラムアドレスストローブ	*column address strobe*
セルアレイ	*cell array*

Unknown words:

ポリッシング	*polishing*
コレクタ	*collector*
積層する	*to form*

Fig. 2: Correct translation pairs

Table 2 shows the ranking of the correctly estimated translation pairs in seven sample texts. The upper row shows the average of seven individual texts; the lower shows the result using all seven texts in one time. The translation of over 70% of compound nouns is obtained as the first candidate, and over 80% in the top three. The result for unknown words is 54.0% and 65.0%. Though the accuracy for the unknown words is relatively low, the estimation has been impossible for Yamamoto (1993). Here, the terms whose correct translations are not found in English texts are excepted from evaluation. Such data occur when human experts give a noun translation for Japanese verbal noun term which is translated as a verb in the actual text. The ratio of this kind of translation pairs is about 3%. The rate of the correct data is calculated by the ratio of the total occurrences.

The accuracy for the average of unknown words is 52.4% in the top three. The result using all texts is significantly better than the average because the statistical information is the major factor in the current implementation. Use of more linguistic information such as in Dangan (1991) and Matsumoto (1993) would improve the total performance.

Linguistic information has proven effective to estimate translations of low-frequency terms. Of terms which appeared only once in a Japanese text, 215 translations are obtained correctly as the first candidate from 327 terms (65.7%) in seven texts.

The fourth example of compound nouns in Fig. 2 shows the advantage of statistical information because the correct translation was obtained in spite of the wrong word segmentation. The Japanese term really consists of three words (カラム, アドレス, ストローブ), each of which corresponds to "*column*," "*address*" and "*strobe*" respectively. But word segmentation output four words (カラム, アドレス, スト, ローブ) because "ストローブ" is unknown and "ス

Table 2: Accuracy of translation estimates

	Compound nouns (occurrences)			Unknown words (occurrences)		
	total	first estimate	top 3 estimates	total	first estimate	top 3 estimates
1 text (average)	460.6	71.7% (330.3)	82.5% (380.1)	55.6	30.1% (16.7)	52.4% (29.1)
7 texts	3,224	72.9% (2,349)	83.3% (2,680)	389	54.0% (210)	65.0% (253)

ト" is known as "*strike*."

The cases where no correct translation has been obtained needs to be examined. The major reasons for failures are:

1. Errors in mapping corresponding units.
2. Errors in word segmentation of unknown compound words.

Mapping unit errors occur when the one-to-one unit correspondence does not exist. The experiment using one text shows that 12 out of 98 Japanese sentences have no one-to-one corresponding English sentence. For better unit correspondence, the units should be smaller, for example, a clause or a verb phrase, so as to make the corresponding accuracy and frequency in text higher and statistical information more effective. It would improve the unit mapping when one Japanese sentence is translated into several English sentences or vice versa.

The segmentation errors of unknown words arise often in case of *Katakana* compound word. *Katakana* is the phonetic alphabet in Japanese for spelling foreign words. Since many compound nouns in a technical field consist of *Katakana*'s with no space between component words, much larger lexicon will contribute to more accurate segmentation.

7 CONCLUSION

An MT dictionary has been generated from Japanese and English parallel texts. The method proposed in this paper assumes unit correspondence and utilizes linguistic information in an MT bilingual dictionary as well as statistical information, namely, word frequency, to estimate the English translation. Over 70% accurate translations for compound nouns are obtained as the first candidate from small (about 300 sentences) Japanese/English parallel texts (patent specifications) containing severe distortions. The accuracy of the first translation candidates for unknown words, which cannot be obtained by a linguistic-based method, is over 50%.

The current implementation shows promising results for a difficult target (patent texts) despite relatively simple linguistic knowledge. The overall performance will be improved by using more linguistic knowledge and optimizing parameters calculated by statistical information.

References

Brown, P. F.; Lai, J. C.; and Mercer, R. L. (1991). "Aligning sentences in parallel corpora." In *Proc. of the 29th Annual Meeting of the ACL*, 169-176.

Chen, S. F. (1993). "Aligning sentences in bilingual corpora using lexical information." In *Proc. of the 31st Annual Meeting of the ACL*, 9-16.

Dagan, I.; Itai, A.; and Schwall, U. (1991). "Two languages are more informative than one." In *Proc. of the 29th Annual Meeting of the ACL*, 130-137.

Gale, W. A., and Church, K. W. (1993). "A program for aligning sentences in bilingual corpora." *Computational Linguistics*, 19(1), 75-90.

Katoh, N. (1993). "Word selection by searching the translation candidates on monolingual texts in target language." *Technical Report of IEICE*, NLC93-32. (in Japanese)

Kupiec, J. (1993). "An algorithm for finding noun phrase correspondences in bilingual corpora." In *Proc. of the 31st Annual Meeting of the ACL*, 17-22.

Matsumoto, Y.; Ishimoto, H.; and Utsuro, T. (1993). "Structural Matching of Parallel Texts." In *Proc. of the 31st Annual Meeting of the ACL*, 23-30.

Nogami, H.; Kumano, A.; Tanaka, K.; and Amano, S. (1991). "Learning of translation words using target-language documents." In *Proc. of 42nd Annual Meeting of IPSJ*, 2C-6. (in Japanese)

Yamamoto, Y., and Sakamoto, M. (1993). "Extraction of technical term bilingual dictionary from bilingual corpus." *IPSJ SIG Notes*, NL94-12. (in Japanese)