Learning Translations from Monolingual Corpora

Hirokazu Suzuki
Research and Development Center
Toshiba Corporation
1 Komukai-Toshiba-cho, Saiwai-ku,
Kawasaki, Japan, 210-8582
hirokaz.suzuki@toshiba.co.jp

Akira Kumano
Research and Development Center
Toshiba Corporation
1 Komukai-Toshiba-cho, Saiwai-ku,
Kawasaki, Japan, 210-8582
akira.kumano@toshiba.co.jp

Abstract

This paper proposes a method for a machine translation (MT) system to automatically select and learn translation words, which suit the user's tastes or document fields by using a monolingual corpus manually compiled by the user, in order to achieve high-quality translation. We have constructed a system based on this method and carried out experiments to prove the validity of the proposed method. This learning system has been implemented in Toshiba's "The Honyaku" series.

1 Introduction

To realize high-quality machine translation, it is important not only to select from several candidates the target word that has the right meaning within a particular sentence but also to be appropriate with respect to such extra sentential factors as users' tastes, purposes, and document fields. If a machine-translated sentence includes an inappropriate word, users need to select the correct one from its several candidates and direct it to the system. Once it goes through this "**learning process,**" an MT system is able to select the prioritized word and thus output precise translation, which we call "**translation learning.**"

Previously, users were required to manually execute learning for each document. The fact that this takes many steps and imposes a burden on users calls for the development of efficient automatic translation learning.

Existing automatic translation learning methods can be roughly categorized into two types: One uses bilingual corpora [1,2,3] and the other, monolingual corpora [4].

The first type poses availability problems in that it is difficult to obtain bilingual corpora, which match the user's purposes and applied fields. While monolingual corpora are more available than bilingual ones, the latter type has much to improve in the accuracy of word selection needed for precise translation.

With this background, we first propose a new algorithm to enable highly accurate word selection with the use of monolingual corpora. Then, we implement this algorithm and prove the validity of this algorithm through experiments.

2 Learning Translations from Monolingual Corpora

2.1 The System

Prior to translation, users need to prepare several target-language documents in a specific domain as a monolingual corpus to ensure translation learning that is suitable for their tastes and purpose.

Here we limit our focus to noun and the direction to English-to-Japanese translation.

First, we will define "translation learning" as:

> If a noun in a sentence to be translated has more than one translation candidates, the one that satisfies the criteria based on statistical information in the prepared monolingual corpus is preferentially selected as a suitable translation.

Consider translating the following sentence (1) into Japanese using documents in a domain "space.":

(1) *In 1978 the United States launched the Pioneer Venus mission.*

Without learning, MT translation for (1) would be something like:

(2) *1978 年には、アメリカが初期のビーナスミッションを打ち上げた。*

The word "mission" in the original sentence has several translation candidates, such as "使命(an appointed task to be accomplished)," "代表団(a delegation sent to foreign countries to conduct negotiation)," "伝道団(a delegation sent by a religious community to propagate faith)," and "ミ

ッション(an appointed task for some purpose such as military and space exploring)." In this case, the last one is selected because its semantic rule is applicable. "Semantic rule" means the rule used in order to translate properly when some translation candidates exist.

The word “Venus” has also several translation candidates, such as "ビーナス(goddess of love)," "金星(second planet from sun in solar system)," "美人(a beautiful woman)." Although in (2) "ビーナス" is selected as a default translation(not applied semantic rules), it is suitable to select "金星" considering the domain of the document.

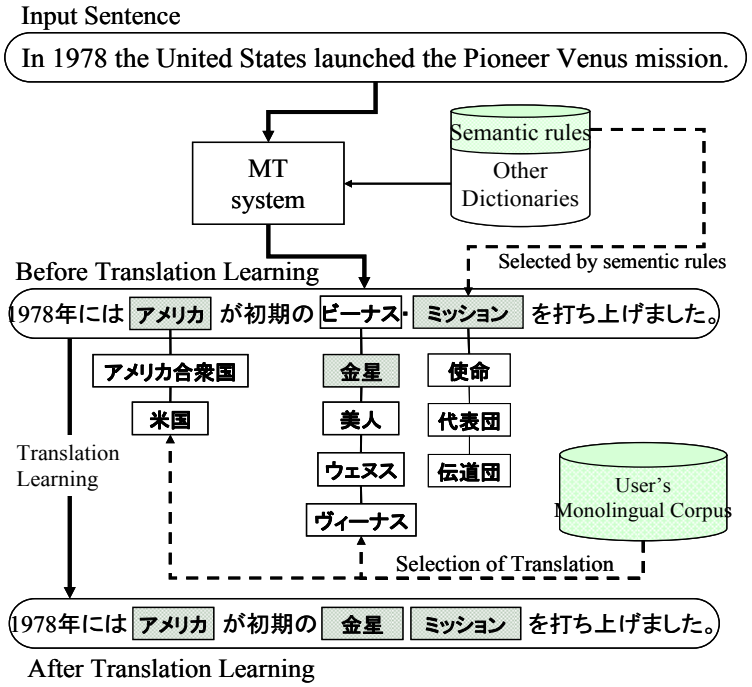

Fig. 1: Translation Learning from Monolingual Corpora

While the above example is domain-dependent, some translations are determined by users’ tastes or purpose.

Now the next question is which translation word is the most suitable for the 'user'?

Following the above definition of “translation learning,” our system selects one translation word as the most suitable for the user. For example, if the word "金星" appears far more frequently than other candidates, or co-occurs with other translation words, the system assumes "金星" as the best translation for “Venus.” Thus, the output of the system would be like (3):

(3) *1978 年には、アメリカが初期の金星ミッションを打ち上げた。*

This automatic translation learning reduces the burden on the part of users. Fig.1 depicts these progresses.

In (3) we regard that the translation word "ミッション" as more accurate than "ビーナス," because unlike the latter, the former is derived from its semantic rule. By contrast, the default translation word "ビーナス" should be reconsidered.

In this paper, those words whose translations are strictly determined by semantic and other rules, in this case 'mission,' are called '**fixed words,**' and its translations, '**fixed translation,**' while those words whose translation word should be reselected from its candidates, in this case 'Venus,' are called ‘**learning targets.**'

In 2.2, we consider how the system selects the best translation from the candidates of a 'learning target' referring to the statistical information in the corpus.

2.2 Main Algorithm

Our translation learning algorithm, which works upon the completion of translation, or as post-processing of machine translation, can be broken down into the following five steps:

i. The system extracts all the nouns in a sentence to be translated.
ii. If a word whose translation words are derived from its semantic rules exists, then those translation words are pushed into a '**fixed translation list.**'
iii. For each word not covered by step ii., the system regards its translation as a 'key' and all its translation candidates as 'values', and pushes a set of a key and values into a '**learning target array.**'
iv. For each key the best translation words are selected from the stored values using a '**selection algorithm.**'
v. The word selected in step iv. replaces the default translation as a new translation in the target sentence, regarded as the translation word has been changed from the key in the learning target array. (also, there is the case that the translation word does 'NOT' change.)

The selection algorithm in step iv. is the algorithm that selects the best suitable translation word, given the fixed translation list and the learning target array.

In next chapter, we explain the selection criteria used in the selection algorithm.

3 Translation Selection Criteria

In this paper, we make use of the co-occurrence information as one of the statistical information from corpus. Frist, we explain the way to count the number of the sentences in which the co-occurrence words exist. And next, we explain the criteria of the co-occurrence intensity.

3.1 Definition of The Weighted Co-occurrence Sentence

To begin with, we define co-occurring words as all words appearing simultaneously within a sentence. The reasons we have adopted this definition are:

- It is pratical for implementation because of easy design of the algorithm
- It can exhastively extract co-occurring words.

The second reason is especially important here since we deal with only those learning targets whose part of speech is noun and hence they may not be adjacent to each other within a sentence.

Next we describe how to count sentences which include co-occurring words. We denote a group of learning targets in the target sentence T_S as $\{W_{S_1}, W_{S_2}, \cdots, W_{S_r}\}$. Let the number of translation candidates of $W_{S_i}(1 \le i \le r)$ be n_i, each translation candidate be $w_{ij}(1 \le j \le n_i)$, and the most suitable translation word of W_{S_i} be W_{T_i}. A fixed translation within one sentence, namely a translation word which is determined by semantic rules or is selected by other evaluation criteria than co-occurring criteria, is denoted as $W_C = \{\tilde{w}_1, \tilde{w}_2, \cdots, \tilde{w}_L\}$. We pick out the most suitable translation word from the translation candidates of a learning target according to the intensity of co-occurrences between those candidates and fixed translations.

And then, a sentence in the corpus is T_m, and the corpus is $D_M = \{T_1, T_2, \cdots, T_M\}$, where M is the number of all the sentences in the corpus.

The number of sentences in which the translation candidate w_{ij} and the fixed translation $\tilde{w}_l(1 \le l \le L)$ co-occur, is $C_{w_{ij},\tilde{w}_l}$. The number of the word pair $(w_{ij}, \tilde{w}_l)$ in the sentence T_m is $P_{T_m,w_{ij},\tilde{w}_l}$, and the number in the corpus D_M is $P_{w_{ij},\tilde{w}_l}$. Thus:

$$P_{T_m,w_{ij},\tilde{w}_l} \equiv \min\left\{P_{T_m,w_{ij}}, P_{T_m,\tilde{w}_l}\right\} \quad (1)$$

$$P_{w_{ij},\tilde{w}_l} \equiv \sum_{m=1}^{M} P_{T_m,w_{ij},\tilde{w}_l} \quad (2)$$

$P_{T_m,w_{ij}}$ and $P_{T_m,\tilde{w}_l}$ show the frequency of w_{ij} and $\tilde{w}_l$ in the sentence T_m respectively.

Because of the definition, when two noun words are in one sentence, we approve that pair as the co-occurrence. But also, there is a case that, in one sentence, the same co-occurrence pair is more than two. We can say that the sentence, has some same pairs, has 'strong' co-occurrence, compared to the sentence has one pair. Therefore, in this paper, we count the number of the sentence that has the co-occurrence pair, weighting according to the number of the same pairs in the sentence.

When the sentence that has the co-occurrence pair $(w_{ij}, \tilde{w}_l)$ is extracted, its sentence is expressed by $T_{k,w_{ij},\tilde{w}_l}\left(1 \le k \le C_{w_{ij},\tilde{w}_l}\right)$. We defined '**Weighted number of the sentence having the co-occurrence**' $S_{w_{ij},\tilde{w}_l}$ as follows:

$$S_{w_{ij},\tilde{w}_l} \equiv \sum_{k=1}^{C_{w_{ij},\tilde{w}_l}} \Gamma_{T_{k,w_{ij},\tilde{w}_l}} \quad (3)$$

where we use the weight $\Gamma_{T_{k,w_{ij},\tilde{w}_l}}$ related to the frequency of $(w_{ij}, \tilde{w}_l)$ in the sentence $T_{k,w_{ij},\tilde{w}_l}$.

And now, if we assume $1 \le \Gamma_{T_k,w_{ij},\tilde{w}_l} \le P_{T_k,w_{ij},\tilde{w}_l}$ then $C_{w_{ij},\tilde{w}_l} \le S_{w_{ij},\tilde{w}_l} \le P_{w_{ij},\tilde{w}_l}$. In this paper, we set the value as follows:

$$\Gamma_{T_k,w_{ij},\tilde{w}_l} = \begin{cases} 1 & if\ P_{T_k,w_{ij},\tilde{w}_l} = 1 \\ 1.2 & if\ P_{T_k,w_{ij},\tilde{w}_l} = 2 \\ 1.5 & if\ P_{T_k,w_{ij},\tilde{w}_l} \ge 3 \end{cases} \quad (4)$$

3.2 Intensity of Co-occurrence

In this section, we describe the evaluation criteria of the intensity of the co-occurrence. This uses the linear combination function composed of '**total weighted mutual information**' and '**average mutual information**'.

3.2.1 Total Weighted Mutual Information

At first, we define the evaluation criterion of the intensity of the mutual information by use of $I(w_{ij} : \tilde{w}_l)$:

$$I(w_{ij} : \tilde{w}_l) \equiv \log_2 \frac{P(w_{ij}, \tilde{w}_l)}{P(w_{ij}) \cdot P(\tilde{w}_l)} \quad (5)$$

In the equation (5), $P(w_{ij}, \tilde{w}_l)$, $P(w_{ij})$ and $P(\tilde{w}_l)$ are the probability of the co-occurrence $(w_{ij}, \tilde{w}_l)$, the appearance probability of w_{ij} and $\tilde{w}_l$ respectively. Since these probabilities are unknown generally, we use the following probable value by means of the weighted number of the sentence having the co-occurrence:

$$\hat{P}(w_{ij}, \tilde{w}_l) = \frac{S_{w_{ij}, \tilde{w}_l}}{M} \quad (6)$$

$$\hat{P}(w_{ij}) = \frac{\sum_{m=1}^{M} P_{T_m, w_{ij}}}{M} \quad (7)$$

$$\hat{P}(\tilde{w}_l) = \frac{\sum_{m=1}^{M} P_{T_m, \tilde{w}_l}}{M} \quad (8)$$

And then, the probable value $\hat{I}(w_{ij} : \tilde{w}_l)$ of $I(w_{ij} : \tilde{w}_l)$ equals this:

$$\hat{I}(w_{ij} : \tilde{w}_l) = \log_2 \frac{S_{w_{ij}, \tilde{w}_l} \cdot M}{\sum_{m=1}^{M} P_{T_m, w_{ij}} \cdot \sum_{m=1}^{M} P_{T_m, \tilde{w}_l}} \quad (9)$$

The above value is called '**Weighted mutual information**'. Moreover the sum of the weighted mutual information about W_C is called '**Total weighted mutual information**', and is described by the following equation:

$$WI(w_{ij} : W_C) = \sum_{l=1}^{L} \tilde{I}(w_{ij} : \tilde{w}_l) \quad (10)$$

This shows the dependency between w_{ij} and W_C.

3.2.2 Average Mutual Information

In this section, we talk about the evaluation criterion of the mutual information's intensity by use of the entropy. If the fixed translation W_C has been already known, the average information content(**conditional entropy**) $H(w_{ij} | W_C)$ of w_{ij} is represented by this :

$$H(w_{ij} | W_C) = -\sum_{l=1}^{L} P(w_{ij}, \tilde{w}_l) \log_2 P(w_{ij} | \tilde{w}_l) \quad (11)$$

The probable value $\hat{P}(w_{ij}, \tilde{w}_l)$ of $P(w_{ij}, \tilde{w}_l)$ is given by the equation (6) and the probable value $\hat{P}(w_{ij} | \tilde{w}_l)$ of the conditional probability $P(w_{ij} | \tilde{w}_l)$ is given by the following equation:

$$\hat{P}(w_{ij} | \tilde{w}_l) = \frac{\hat{P}(w_{ij}, \tilde{w}_l)}{\hat{P}(\tilde{w}_l)} = \frac{S_{w_{ij}, \tilde{w}_l}}{\sum_{m=1}^{M} P_{T_m, \tilde{w}_l}} \quad (12)$$

wherein the equation (6) and (8) are used.

In consideration of the above, using the equation (11), (6) and (12), the probable value $\hat{H}(w_{ij} | W_C)$ of $H(w_{ij} | W_C)$ is presented by:

$$\hat{H}(w_{ij} | W_C) = -\sum_{l=1}^{L} \frac{S_{w_{ij}, \tilde{w}_l}}{M} \log_2 \frac{S_{w_{ij}, \tilde{w}_l}}{\sum_{m=1}^{M} P_{T_m, \tilde{w}_l}} \quad (13)$$

The average information content(entropy) $H(w_{ij})$ of w_{ij} is

$$H(w_{ij}) = -P(w_{ij}) \log_2 P(w_{ij}) \quad (14)$$

and the probable value $\hat{H}(w_{ij})$ of $H(w_{ij})$ is

$$\hat{H}(w_{ij}) = -\frac{\sum_{m=1}^{M} P_{T_m, w_{ij}}}{M} \log_2 \frac{\sum_{m=1}^{M} P_{T_m, w_{ij}}}{M} \quad (15)$$

wherein the equation (7) is used.

Then, '**Average mutual information**' $EI(w_{ij} : W_c)$ is described by the following equation :

$$EI(w_{ij} : W_C) = \hat{H}(w_{ij}) - \hat{H}(w_{ij} | W_C) \quad (16)$$

by use of the equation (13) and (15).

The average mutual information shows the reduced amount of the ambiguity about w_{ij}, when W_C has been known already.

3.2.3 Selection Criterion by Co-occurrence Intensity

We define the evaluation formula $\varepsilon[w_{ij}]$ by use of the equation (10) and (16) as follows:

$$\varepsilon[w_{ij}, W_C] \equiv \alpha WI(w_{ij} : W_C) + \beta EI(w_{ij} : W_C) \quad (17)$$

We decide the best translation word by the following criterion:

$$W_{T_i} = \arg \max_{w_{ij}} \varepsilon[w_{ij}, W_C] \quad (18)$$

In the above evaluation criterion $\varepsilon[w_{ij}]$, the first term is concerned with the total weighted mutual information and the second is concerned with the average mutual information. And further, we can decide which criterion we make much of by means of changeing the coefficient α and β. The coefficient α and β can be decided by experience respectively, and in this paper, we decide $\alpha = 0.5$ and $\beta = 600$.

3.3 Selection Criterion by Appearance Frequency

Also, we adopt the method of '**Interval**

estimation' as another criterion. We assume that for the translation candidate $w_{ij}\ (1 \le j \le n_i)$ of W_{S_i} in the corpus, the noun that has most frequently appeared is w' and its frequency is n', the noun that has secondary frequently appeared is w'' and its frequency is n''. If the following equation :

$$\ln \frac{n'+0.5}{n''+0.5} \ge \theta_{conf} + Z_{1-\alpha}\sqrt{\frac{1}{n'+0.5}+\frac{1}{n''+0.5}} \quad (19)$$

has realized, we decide the translation word W_{T_i} of W_{S_i} as w' [5]. In this paper, we use the threshold $\theta_{conf} = 0.2$ and $Z_{1-\alpha} = 1.04$ (the confidence is 85.083%).

4 Selection Algorithm

The algorithm of section 2.2 iv. is as follows:

Step 1: [**Definition**] For the sentence T_S in the original document, we assume that, the group of the key in the learning target array is $W_U = \{W_{S_1}, \cdots, W_{S_r}\}$, the value of W_{S_i} is $\{w_{i1}, \cdots, w_{in_i}\}$, fixed translation is $W_C = \{\tilde{w}_1, \cdots, \tilde{w}_L\}$, the word that is selected as the translation of W_{S_i} is W_{T_i}.

Step 2: [**Addition of the definitely determined translation to the fixed translation**] If the number of the translation candidate is just one, its translation is added to the fixed translation W_C, and it is regarded as W'_C instead of W_C.

Step 3: This process is terminated, if $W_U = \phi$.

Step 4: [**Addition of the candidates of top 2 high entropy to the fixed translation**] If at least one conditions of the following are satisfied:

- $W'_C = \phi$
- $W'_C \ne \phi$ and $^{\forall}\tilde{w}_l \in W'_C, {}^{\forall}T_m \in D_M, P_{T_m,\tilde{w}_l} = 0$
- $W'_C \ne \phi$ and $^{\exists}\tilde{w}_l \in W'_C, {}^{\exists}T_m \in D_M, P_{T_m,\tilde{w}_l} = 0$ and $|W'_C| < 6$

, among the words w_{ij} of which entropy $\hat{H}(w_{ij})$ satisfies:

$$\hat{H}(w_{ij}) > \theta_{entropy}$$

, the word $\tilde{w}_{L+1}$ that maximizes $\hat{H}(w_{ij})$ and the word $\tilde{w}_{L+2}$ that secondarily maximizes $\hat{H}(w_{ij})$ is added to W'_C. Then, the entry of the key that relates to the above is got rid of from the learning target array, wherein $\theta_{entropy} = 0.002$. The modified fixed translation is W''_C.

Step 5: [**Translation selection based on the co-occurrence with the fixed translation including the word by Step 2, 4**] When the group of the key in the modified learning target array is $W'_U = \{W'_{S_1}, \cdots, W'_{S_{r'}}\}$, and the value of W'_{S_h} is $\{w'_{h_1}, \cdots, w'_{h_{n'_h}}\}$, the best translation word is determined by the following formula:

$$W'_{T_h} = \arg\max_{w'_{h_k}} \varepsilon\left[w'_{h_k}, W''_C\right]$$

wherein W'_{T_h} is the translation of W'_{S_h}.

Step 6: [**Translation selection based on the interval estimation**] We assume the learning target that has not been identified yet by Step 5 is $\{W''_{S_1}, \cdots, W''_{S_{r''}}\}$. Among $W''_{S_u}\ (1 \le u \le r'')$, the word of the top 2 high frequency in the corpus is regarded as w' and w'', their number is regarded as n' and n'' respectively. The translation word is determined based on the following formula:

$$W''_{T_u} = w' \qquad if\ (19)\ is\ realized.$$

wherein W''_{T_u} is the translation of W''_{S_u}.

Step 7: [**Translation selection based on the co-occurrence with all fixed translation**] We assume W'''_C includes W''_C and the selected translation by Step 5 and 6, the learning target that has not determined yet is $\{W'''_{S_1}, \cdots, W'''_{S_{r'''}}\}$, the translation candidate of W'''_{S_v} is $\{w'''_{v_1}, \cdots, w'''_{v_{n'''_v}}\}$. The translation word is determined based on the following formula:

$$W'''_{T_v} = \arg\max_{w'''_{v_y}} \varepsilon\left[w'''_{v_y}, W'''_C\right]$$

wherein W'''_{Tv} is the translation of W'''_{Sv}.

Step 8: [**Output**] The system outputs the selected translation. But if the translation has not determined by the above steps, the system outputs the default translation.

5 Experiments and Evaluations

We use two kinds of the monolingual corpora in Japanese for the experiments. The one is part-of-speech-annotated corpus created from the documents(number of sentences is 8,224) that has been collected mainly from Web page of NASDA(National Space Development Agency of Japan)[8] [1], the other is from the

[1] On October 1, 2003, ISAS(Institute of Space and Astronautical Science), NAL(National Aerospace Laboratory of Japan) and NASDA had been merged into one independent administrative institution: the Japan Aerospace Exploration Agency (JAXA).

documents(number of sentences is 10,018) that has been collected from Web page about F1(Formula One).

One kind of the test documents are about the space, and have been collected from the web page of NASA[9], Chandra X-ray observatory[10] and SEC(Space Environment Center)[11], number of sentences is 201. Another kind of the test documents are about F1, and have been collected from the web page[12,13] about news and technical report, the number of sentences is 198.

In this section, we evaluate the translation word by 6 grade(Table 1) for all noun before and after the translation learning. In Table 1, "Right" means 'correct or proper translation selected' and "Wrong" means 'improper translation selected'.

We introduce the following criteria to evaluate objectively :

$$\text{(Quality Progression Ratio)} = \frac{(number\ of\ correct\ noun) - (number\ of\ incorrect\ noun)}{number\ of\ noun}$$

$$\text{(Improvement Ratio)} = \frac{number\ of\ improved\ noun}{number\ of\ noun}$$

$$\text{(Applicability)} = \frac{number\ of\ correct\ noun}{number\ of\ noun}$$

$$\text{(Precision)} = \frac{number\ of\ improved\ noun}{number\ of\ modified\ noun}$$

Evaluation	Before learning translation	After learning translation	Translation changed ?
Unchanged_Right	Right	Right	No
Wrong→Right	Wrong	Right	Yes
Right→Right	Right	Right	Yes
Right→Wrong	Right	Wrong	Yes
Wrong→Wrong	Wrong	Wrong	Yes
Unchanged_Wrong	Wrong	Wrong	No

Table 1 : Evaluation Categories

Domain	**Space**		F1	
Mode	**Normal**	**Collo**	**Normal**	**Collo**
Num. of Sentence	**201**		**198**	
Number of noun	**2047**		**1678**	
Num. of modified	**385**	**387**	**485**	**488**
Unchanged_Right	**1,580**	**1,586**	**1,145**	**1,142**
Wrong→Right	**118**	**125**	**173**	**189**
Right→Right	**181**	**180**	**209**	**216**
Right→Wrong	**66**	**62**	**76**	**64**
Wrong→Wrong	**20**	**20**	**27**	**19**
Unchanged_Wrong	**82**	**74**	**48**	**48**

Table 2 : Evaluation Results

Domain	**Space**		F1	
Mode	**Normal**	**Collo**	**Normal**	**Collo**
Quality Progression Ratio	**0.835**	**0.847**	**0.820**	**0.843**
Improvement Ratio	**0.146**	**0.149**	**0.199**	**0.241**
Applicability	**0.918**	**0.924**	**0.910**	**0.922**
Precision	**0.777**	**0.788**	**0.788**	**0.830**
P-Value	**$7.75E^{-5}$**	**$2.38E^{-6}$**	**$3.61E^{-10}$**	**$8.29E^{-16}$**

Table 3 : Evaluation Results of Criteria

Moreover, we try another algorithm, Collocation Mode, too. In advance, we extract the collocation estimated from the prepared corpus and create '**Collocation candidate list**' for each domain. In advance, We extract the collocation estimated from the prepared corpus and create 'Collocation candidate list' for each domain. In Step2 of the above mentioned algorithm(Normal Mode), if the nouns more than two are continuously located in a target sentence, among these words, the word that has been registered in the collocation candidate list is selected preferentially and added to the fixed translation.

The results are shown in Table 2 and 3.

Because the precision of each algorithm is more than 77%, we can say that high accuracy translation selection has been realized. Especially, since Collocation Mode improves the precision greatly, we can also say that the adoption of the information from the extracted collocation in the corpus works efficiently.

P-value of each mode is shown in Table3.

In both modes, the translation quality has been changed, wherein its level of significance is more than 99.99%, and finally, we can see this algorithm is effective.

6 Conclusion

In this paper, we have proposed the translation learning method that selects the suitable translation for user's taste or purpose when several candidates exist.

This method determines the translation on the basis of the co-occurrence intensity between each candidate and the fixed translation in the monolingual corpus, that is evaluated by use of the criterion ε consists of 'total weighted mutual information' and 'average mutual information'.

Furthermore, we have done the experiments using two kinds of test documents that have difference of domain, Space and F1. When the domain is the space and the collocation candidates aren't used, applicability is 91.8% and precision is 77.7%. Since the noun improvement ratio is 14.6%, we see that this method improves approximately 15% of all nouns in the test documents. Using the collocation candidates, applicability is 92.4% and precision is 78.8%, nearly 0.6 and 1.1 points higher respectively. When the domain is F1 and the collocation candidates aren't used, applicability is 91.0% and precision is 78.8%. Since the noun improvement ratio is 19.9%, we see that this method improves approximately one-fifth of all nouns in the test documents.

We have verified the validity of this method by sign test.

7 Discussion

In this study all the nouns are equally treated as learning targets. However, occasionally some translations should not be changed from the original default translation. For example, in case of the commonly used word "system," which has a great many translations, including "系統(a group of interacting, interrelated, or interdependent elements forming a complex whole)", "系 (a naturally occurring group of objects or phenomena, like solar system)", and "学説(an organized set of interrelated ideas or principles)," is used in a variety of domains, users might wish to use the default translation word "システム" which is the phonetic equivalent of the English word "system."

So we have excluded the words whose original translations should not be modified, which we call '**Unnecessary learning target**'. If unnecessary learning target is set, the system removes these words from translation target.

This additional function contributes to far more accurate translation learning.

8 Future Works

Future themes are as follows:

- **Co-occurrence Judgement by N-gram**

If two word appear in one sentence simultaneously, we regard them as co-occurrence. This method doesn't have high accuracy because of the tendency of getting many pairs in which a strong dependency doesn't exist. Therefor we think that more accurate translation learning can be realized by use of the co-occurrence based on N-gram.

- **Co-occurrence between Learning targets**

In this method, we have considered only the co-occurrence between 'fixed translation' and 'candidate of learning target'. But in case of translation selection, it is suitable to take the forward/backward words into consideration. We think that this is very effective to translate the collocation words. Therefore, we use the co-occurrence not only between 'fixed translation' and 'candidate of learning target', but also between 'candidates of learning target'.

References

[1] S. Niessen, S. Vogel, H. Ney and C. Tillman, "A dp based search algorithm for statistical machine translation", Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics, pp.960-966, 1998.

[2] Ye-Yi Wang and Alex Waibel, "Decoding algorithm in statistical machine translation", Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics, pp.366-372, 1997.

[3] Gale William and Kenneth Church, "A Program of Aligning Sentences in Bilingual Corpora", Computational Linguistics, 19(1), pp.75-102, 1993.

[4] Kumiko TANAKA and Hideya IWASAKI, "Extraction of Lexical Translations from Non-Aligned Corpora", Proceedings of the 16th International Conference on Computational Linguistics(COLING), pp.580-585, 1996.

[5] Ido Dagan and Alon Itai, "Word Sense Disambiguation Using a Second Language Monolingual Corpus", Association for Computational Linguistics, pp.563-596, 1994.

[6] Fung Pascale, "A pattern matching method for finding noun and proper noun translations from noisy parallel corpora", Proceedings of the 33rd Annual Conference of the Association for Computational Linguistics, 1995.
[7] Shinichi Doi and Kazunori Muraki, "Evaluation of DMAX Criteria for Selecting Equivalent Translation based on Dual Corpora Statistics", TMI'93, pp.302-311, 1993.
[8] National Space Development Agency of Japan (NASDA), http://www.nasda.go.jp/.
[9] NASA, http://www.nasa.gov/.
[10] Chandra X-ray observatory Center, http://chandra.harvard.edu/.
[11] Space Environment Center, http://www.sec.noaa.gov/.
[12] F1MECH.com, http://www.f1mech.com/.
[13] Planet-F1, http://www.planet-f1.com/.