Collection of Dictionary Data through Internet Translation Service

Keisuke Nakayama
R&D Center, Toshiba Corporation
Japan

Akira Kumano
R&D Center, Toshiba Corporation
Japan

Abstract

We have developed an Internet translation service, which we began to provide in 1997 for English to Japanese translation and in 1998 for Japanese to English. In this service, users send a translation request from a web page and receive by e-mail the result of the translation outputted by Toshiba's machine translation system. As in other similar services, users can specify English-Japanese word pairs(dictionary data) when making a translation request. What distinguishes our service from others is that our service system constructs users' own dictionaries on the server and helps them with this work by extracting words which the system expects to improve the system's translation quality if included in the dictionaries. With this function, users can efficiently add new word pairs so as to upgrade their own dictionaries when requesting re-translation. The dictionary data thus obtained from users can be utilized to improve the system dictionary on the server also.

1 Introduction

Through the spread of the Internet, currently a wide variety of services, most typically, homepage information retrieval, is available on the Web. Here, people can easily access foreign sites just like domestic sites and thus have many more opportunities to encounter foreign languages, predominantly English. As such, the need of machine translation which helps users understand Web sites written in other languages is rapidly increasing.

In Japan, numerous software products which translate English documents on the Internet directly into Japanese have been developed and are being sold on the market. This translation software is now used by a great many users on their personal computers. At the same time, several Internet translation services, in which users send their documents to the translation server through the Internet and obtain the translation result from the server, are also available. Our translation service named "MT Avenue," which we will describe in this paper, is one example of such machine translation services.

In the meantime, in terms of translation quality, the current machine translation systems are still in their development stage; therefore, expansion of dictionary knowledge, including that of newly coined words, is indispensable to achieve higher accuracy. We have devised a method to collect and accumulate dictionary knowledge, which forms the core knowledge of the whole translation system, and began an Internet translation service for English to Japanese and Japanese to English translations.

One of the innovative features of this service is that it stores user dictionaries on the server, which are a compilation of translation pairs obtained by users when they send a translation request to the server. Furthermore, the system reduces users' task by extracting candidate words from users' text and presenting them along with the translation result. By referring to this word list, uses can add new translation pairs and request re-translation.

This paper is organized as follows: Chapter 2 reviews the merits of translation service in general. Chapters 3 to 5 explain two important functions of our service system "MT Avenue", namely terms extraction and automatic field classification; Chapter 6 presents the result of our analysis on some 5,800 English-to-Japanese translation pairs we have collected from users.

2 Merits of translation service

2.1 Translation software package vs. translation service

In Japan, over twenty different translation software packages are being sold for English to Japanese and/or Japanese to English translation. Despite their popularity on the market, these packages have the following two shortcomings. First, not all the users have the ability to use a machine translation system with ease. As a rule, users need to install machine translation software on their PCs before use. The operation needed for installment may be simple for intermediate and advanced users, but difficult for beginners.

Second, there are a number of requirements for ensuring high-quality translation. To begin with, the machine translation system each user has installed is not automatically upgraded. Users wishing to upgrade their system need to install a newly upgraded version on their PCs on their own for every release. Users also need to maintain their user dictionaries and store dictionary data on their PCs on an individual basis. Furthermore, the PCs for installing the system should have sufficiently high performance because translation involves complex processing.

In contrast, Internet translation service, which is based on a client-server type system, is free from all these problems. To begin with, it enables users to easily access a translation server through the Internet using a web browser and send their translation request to the server from a web browser, to obtain the translation result outputted by the server, via e-mail and other media. (Fig. 1)

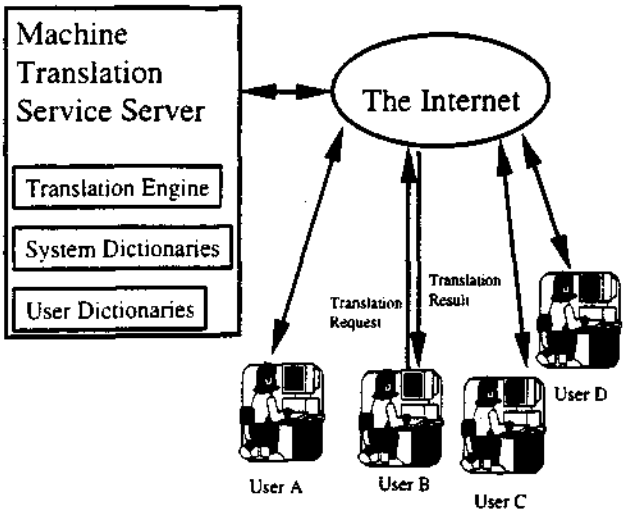

Figure 1: Internet translation service

Furthermore, users are only required to have a web browser for accessing Internet homepages and an e-mail tool for getting replies from the server. Basically, there are no other requirements. In other words, users do not need to install machine translation software on their own PCs. Moreover, they are freed from an onerous upgrading operation, since service providers are to upgrade the translation system on the server every time a newer version is released. This means that many users can easily use the up-to-date translation software at all times, which is perhaps the greatest benefit to users. Third, users' PCs may have relatively low performance, since translation processing is done on the server.

2.2 Automatic collection of dictionary data

Besides all these advantages outlined above, our translation service features a unique function to automatically collect dictionary data for the benefit of both users and service providers. Dictionary data is the most fundamental knowledge to offset the present quality of machine translation systems. On the Internet, new words and technical terms are coined every day while a massive amount of information is being exchanged. To handle these new terms, our dictionary data should be constantly updated. Whereas it is not easy for the server alone to collect such new terms, users trying to translate a certain text know which words should be included in their dictionary along with their equivalents in the target language. (Note that here we assume users of Internet translation services have some knowledge of English and Japanese so that they know, if not perfectly, how the original text should be translated.) Therefore, if each user's knowledge about terminology is passed on to the server, then it should be fairly easy for the server to automatically accumulate linguistic data covering broad fields.

With this background, we have explored how to make the data flow between the server and each user smooth for higher efficiency and created an option to allow users to specify the translations of nouns and noun phrases when making a translation request. We implemented this function in an Internet translation service system named "MT Avenue."

3 Collection of dictionary data through Toshiba's Internet translation service

In this section, we will describe the details of data collection.

3.1 General outline of the service

In this subsection, we will present the general outline of the service, by referring to the Figure 2, which shows the configuration of our translation system.

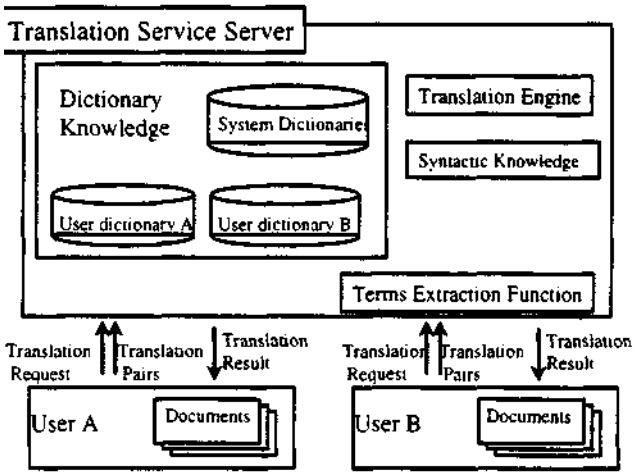

Figure 2: System Composition of MT Avenue

First, users access our homepage ([MT Avenue]) on the Internet and input the English or Japanese text they wish to translate. At the same time, they have an option to specify the translations of nouns and noun phrases by typing in the translation pairs. They can also choose one technical term dictionary for better translation. Currently, six dictionaries are available for English-to-Japanese(information, Internet, electricity, chemistry, machinery, and politics and economy), and four for Japanese-to-English (information, chemistry, machinery, and politics and economy).

Figure 3 shows the homepage for English to Japanese translation. Note that on our actual homepage instructions are given only in Japanese at this moment

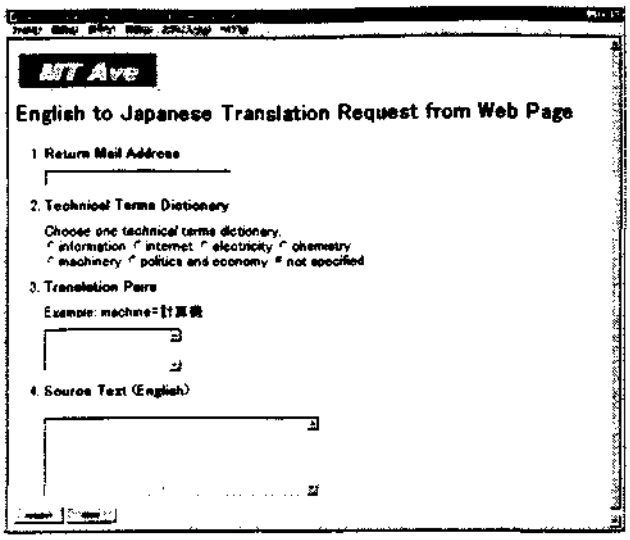

Figure 3: Homepage to request translations

When users click the "request" button at the bottom of the page, all the information entered by the users are transmitted to the translation server through a CGI decoder program on the http server. Then, our machine translation system translates the text under the given environment and extracts linguistic data wherever possible; the translation result is e-mailed to each user, together with the extracted data. If they are not satisfied with the result, they could ask for re-translation, along with additional list of translation pairs.

Below, we will take a closer look at two important functions of this service: registration of translation pairs and extraction of suggested translation pairs.

3.2 Registration of translation pairs

Let us explain how the translation pairs designated by users could be best utilized. While they are used for translating source language texts as in other similar translation services, our system accumulates the data on the server classified by each field and user, so that they could be used again later when users make another translation request. In this way, user dictionaries are automatically constructed on the server, without any burden on users. That is, users do not need to type in the same pairs of translations every time they make a translation request. Moreover, either by e-mail or an Internet browser, users can refer to their own user dictionaries at any time and delete part of the data when the need arises. Examples of translation pairs are shown below.

(Ex. 1)
machine = 計算機 (*keisanki*) (computer)
translation software =
機械翻訳ソフト (*kikai-honyaku-sofuto*)
(machine translation software)

In these examples, the Japanese word "計算機"(*keisanki*) is specified as the translation of an English noun "machine", and the Japanese word "機械翻訳ソフト"(*kikai-honyaku-sofuto*) as the translation of English compound noun "translation software".

In addition, out of this data, the service providers can extract the dictionary knowledge which would be useful for dictionary developers to improve translation quality. For example, those pairs of common terms could be incorporated into the system dictionaries of machine translation systems for both client-server types and software packages.[1]

3.3 Extraction of suggested pairs

In many cases, users do not have a clear idea as to which nouns and noun phrases should be registered to get higher-quality translation effectively. To support such users, we have created a function to extract during translation a list of nouns and noun phrases, which the system judges might be better to include in user dictionaries and send it to each user along with the translation result. This function can be viewed as a help system for building a user dictionary.

Users can obtain better translation results through the interaction with this help system. The data flow between a user and the server is shown in Figure 4.

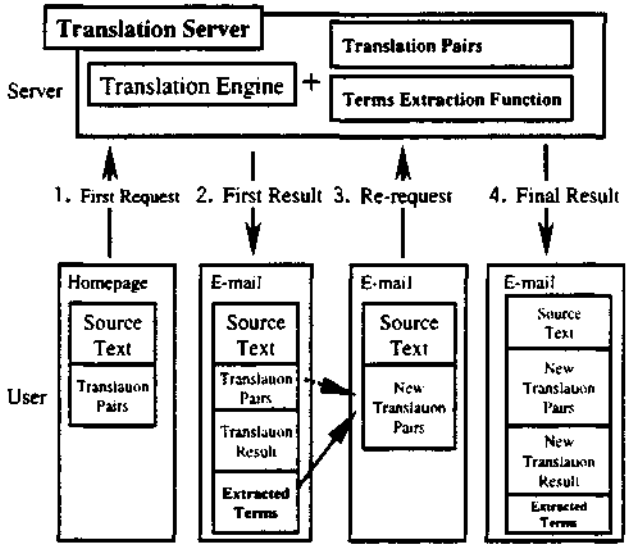

Figure 4: Interaction between a user and the server

Currently, we extract two types of nouns and noun phrases. Examples are given below.

(Ex. 2) Frequency 8: domain name = 領域名 (*ryouiki-mei*)
(Ex. 3) Frequency 5: JavaScript = JavaScript

(Ex.2) involves the case where the output translation is a mere combination of each noun composing the noun phrase, but the original noun phrases appear more than once, in this example, 8 times. The system dictionary gives "領域"(*ryouiki*) as a Japanese equivalent of the

[1] One clause was included in the conditions of use of our service so that we could use users' dictionary data. When users use our MT Avenue system, they have to surely consent to this clause.

English word "domain" and "名"(*mei*) for the word "name", but does not have an entry for the term "domain name." Because of its frequency, it is likely to be a fixed term. In fact, another term "ドメインネーム"(*domein-neemu*), which happens to be the phonetic equivalent of the English term "domain name" would be preferable over "領域名"(*ryouiki-mei*).

(Ex. 3) is the case the system dictionary does not have an entry for the word "JavaScript" but the word appeared five times in the text, which implies that the word is a key term within the text. Since no translation is available, the system outputted the original English. If users know the suitable translation for such unregistered words, they could improve the quality of translation by registering those pairs when requesting translation.

3.4 Re-translation based on the extracted data

Users not satisfied with the obtained result sent by e-mail may request re-translation by directly replying to the e-mail they received. Here, they can enter additional translation pairs by reviewing the suggested list of words extracted by the system. In our example above, the following translation pairs could be added at the second request.[2]

domain name = ドメインネーム (*domein-neemu*)
JavaScript = ジャバスクリプト (*jaba-sukuriputo*)

This re-translation request can be repeated for a number of times according to users' need.

4 Evaluations of the term extraction function

In order to see how the term extraction function contributes to the construction of user dictionaries, we examined 1) the amount of translation pairs (dictionary data) examined; 2) the percentage of extracted terms in a total of terms registered by users; and 3) the percentage of extracted terms found in a list of terms registered by users.

4.1 The amount of translation pairs (dictionary data)

In order to collect as much data as possible from users, we offered our Internet service for free for the first three months in the case of English-to-Japanese translation whereas the service for Japanese to English has been available for free from its very start. During the first three months, we collected 5,860 translation pairs as raw data for English; for Japanese to English translation, we collected 1,728 pairs over the first seven months. Table 1 presents the distribution of the total number of translation pairs by each user.

	E-to-J	J-to-E
100 or more words	7 users	0 user
50-99 words	20 users	1 user
20-49 words	39 users	15 users
10-19 words	55 users	27 users
5-9 words	118 users	57 users
2-4 words	232 users	141 users
1 word	218 users	145 users

Table 1: The distribution of the number of translation pairs by users

4.2 The amount of extracted terms and registered terms

We calculated the amount of extracted terms and registered terms only for English to Japanese translation.

For 2) the percentage of extracted terms in a total of terms registered by users, we got:

N_{RE}/N_{TR} = 664/5,860 = 11.3(%)
N_{RE}: number of registered words which are also found in the result of extraction
N_{TR}: total of terms registered by users

Though specifying translation pairs referring to the output of terms extraction function is only available in re-translation requests, i.e., the reference is not available in users' first request and terms registration, the percentage is relatively high, which implies that many users requested re-translations referring to the extracted terms. We cannot conclude that most users referred to the results of terms extraction in specifying translation pairs and registering them in their dictionaries on the server; however, we can say that many users used the results of terms extraction function in spite of the fact that all of the terms shown to the users had been extracted by mechanical processing.

As for 3) percentage of extracted terms found in a list of terms registered by users, we got:

N_{ER}/N_{TE} = 474 / 23,328 = 2.0(%)
N_{ER}: number of words extracted by server and registered by users
N_{TE}: the numbers of extracted words presented to users (excluding those extracted words presented to users who registered no words[3])

The percentage of what were actually registered to users' dictionaries among the extracted terms shown to the users who registered at least one word was only 2%. We consider that further improvement of the knowledge used in terms extraction contributes to increase this rate.

[2] Notice, in Japanese, English computer terms are often translated as the phonetic equivalents of each English word, as shown in the parentheses.

[3] This is because it is highly likely that those who registered no words did not use the terms extraction function.

5 Automatic field classification of the translation pairs

5.1 The method

As mentioned in 3.1, users can choose one technical term dictionary to be used when translating the requested text. With this information, our translation service system automatically categorizes translation pairs by technical field and accumulates the result in the server. Since we have six technical term dictionaries for English to Japanese and four for Japanese to English, translation pairs fall under one of the following seven categories for the former and five for the latter, except those for which technical term dictionaries were unspecified. The data can be tabulated as below:

	E-to-J	J-to-E
information	1,629 words	483 words
internet	950 words	-
electricity and electronics	752 words	-
chemistry	483 words	205 words
machinery	458 words	235 words
politics and economy	264 words	55 words
unspecified	1,556 words	811 words
total[4]	5,860 words	1,728 words

Table 2: The number of translation pairs in each field

5.2 Evaluation

We surveyed translation pairs by technical field and found that those for information and Internet account for the most of the data (See Table 2 again.) as was expected because many of the users of this system are familiar with those fields. For the rest of the fields, there were no aberrant data which would lower the system's translation quality. That is, there were only a few cases where information terms were classified under politics and economy or where chemistry terms were classified as Internet terms.

Users were especially good at making a distinction between information and Internet fields, although we had expected some confusion among users because these two fields are very closely related. See the examples in (Exs. 4-5).

(Ex. 4)Examples of translation pairs in internet field
Cool Site = クールサイト (*kuuru-saito*)
data encryption = データ暗号化 (*deeta-angouka*)
domain name = ドメインネーム (*domein-neemu*)
frame = フレーム (*fureemu*)
gateway = ゲートウェイ (*geeto-wei*)
Java platform = ジャバプラットフォーム (*jaba-purattofoomu*)
JavaScript = ジャバスクリプト (*jaba-sukuriputo*)
push technology = プッシュ技術 (*pusshu-gijutsu*

(Ex. 5)Examples of translation pairs in information field
GUI application = GUIアプリケーション (*GUI-apurikeeshon*)
instance = インスタンス (*insutansu*)
bus reset state = バスリセット状態 (*basu-risetto-joutai*)
daemon = デーモン (*deemon*)
device driver = デバイスドライバ (*debaisu-doraiba*)
method = メソッド (*mesoddo*)
registry = レジストリ (*rejisutori*)

All of this suggests that the data classified by users could be used as a feedback to the system's technical term dictionaries. In contrast, unclassified data, that is, those data for which users did not choose any technical term dictionaries, are found to be a mixture of words in a wide variety of fields. As such, further detailed analysis would be necessary if we were to enter these data in the system's non-technical term dictionary.

However, the classified data had one problem in which users tried to assign one of the fields given, even when the appropriate choice was not found. This is partly because the number of fields offered in our service is quite limited and partly because taxonomically, they were not systematic. Hence, translation pairs classified as chemistry terms included a substantial number of biology terms and medical terms. Similarly, those data classified as machinery terms included physics terms. The field "politics and economy" seemed to be too general because it has many subfields such as finance and demography. The examples are given in (Exs. 6-8), showing more appropriate field in the parentheses.

(Ex. 6)Examples of translation pairs in chemistry field
cornea = 角膜 (*kakumaku*) (biology)
cultured tissue = 培養組織 (*baiyou-soshiki*) (biology)
epidemiology = 伝染病学 (*densenbyou-gaku*) (medicine)

(Ex. 7)Examples of translation pairs in machinery field
aerodynamics = 気体力学 (*kitai-rikigaku*)(physics

(Ex. 8)Examples of translation pairs in politics-and-economy field
checking account = 当座預金 (*touza-yokin*) (finance)
infant mortality = 乳児死亡率 (*nyuuji-shibou-ritsu*) (demography

In sum, this problem could be solved by increasing the number of fields and their subfields. Furthermore, since a wider choice of fields is very likely to decrease the amount of unclassified data, we would be able to

[4] Some translation pairs were entered in multiple fields; therefore, the numbers in the above columns do not add up to the total at the bottom.

collect more detailed translation knowledge effectively.

6 Analysis of the collected dictionary data

We analyzed all the 5,860 data (translation pairs) for English to Japanese translation given in Table 2 from a linguistic view to see if they are effective in upgrading the system's translation quality. The following three points merit mention here.

First, the data contained a significant number of proper nouns and acronyms. Although the frequency of words beginning with a capital is not so high in users' source English texts, they accounted for about one third of the registered words. The primary reason for this is that the translation pairs we analyzed contained quite a number of proper nouns like (Ex. 9)[5] and acronyms like (Ex. 10). In many cases, they are peculiar to the document of a specific user, and therefore are usually not included in the system dictionary. In fact, there are many possible translations for acronyms; for example, other than "Clinical Practice Committee," CPC could be the acronyms for: card programming control, coated paper copier, chronic passive congestion, Communist Party of China, Community Patent Convention, concrete-plastic composite, concrete-polymer composite, and continuous path control.

(Ex. 9)
Damon Hill = デーモンヒル (*deemon-hiru*)
MOUNTAIN VIEW = マウンテンビュー (*maunten-byuu*)
Hoover = フーバ大統領 (*fuuba-daitouryou*) (President Hoover)
Rochester region = ロチェスター地域 (*rochesutaa-chiiki*)

(Ex. 10)
FDIC = 連邦保険預金機構 (*renpou-hoken-yokin-kikou*) (Federal Deposit Insurance Corporation)
CPC = 臨床実践委員会 (*rinsho-jissen-iinkai*) (Clinical Practice Committee)

Second, some of the data involved the case where the original English word and its Japanese equivalent are identical, i.e., words which should not be translated, such as (Ex. 11).

(Ex. 11)
ping = ping
basic = basic

The standard or default translation for the word "basic" is "基礎"(*kiso*); however, it is often more appropriate to give the original English "basic" as its translation in documents on information processing. This type of translation pairs was also found in other fields of study. Hence, this kind of data obtained from users in each field can be useful translation knowledge.

Third, the following three types of data were undesirable, in that they are expected to lower translation quality.

1) words other than nouns

To simplify the user interface for specifying translation pairs, our system currently allows only noun pairs and hence all the translation pairs are interpreted as nouns by the server. However, some users just ignored this instruction and registered translation pairs of words other than nouns including verbs and adjectives, as in (Ex. 12).

(Ex. 12)
adapted = 適応した (*tekiou-shita*) - past particle
align = 整列させる (*seiretsu-saseru*) - transitive verb
cooperable = 協力的 (*kyouryoku-teki*) - adjective

These data suggest that the system needs to make it possible for users to register words other than nouns, while keeping the user interface simple enough. Especially verbs require in-depth examination because good translation may not be obtained without specifying the case frame which is lexically dependent.

2) nouns in plural forms

Although users are supposed to specify translation pairs in the singular form, some users did so in the plural form. (Ex. 13) shows the case in which a user specified translation pairs both in the singular and plural forms.

(Ex. 13)
plug-in = プラグイン (*puragu-in*)
plug-ins = プラグイン (*puragu-in*)
gateway = ゲートウェイ (*geeto-wei*)
gateways = ゲートウェイ (*geeto-wei*)

The problem with this kind of data is the term "plug-ins" as in "These plug-ins are new." may be interpreted as a singular form, such that the number agreement between the subject "these plug-ins" and the verb "are" is not recognized. This causes parsing failure and worsens translation quality. We estimate that we can resolve this problem by asking users on the screen to make sure that their translation pairs are in the singular form.

3) translation pairs with multiple translations

In some cases, different users gave different translations to the same English terms, like the examples in Table 3.

[5] Notice here the specified translation for "Hoover" is more limited in meaning.

English word	Specified translations
encoding	符号化(*fugou-ka*)[2 users] 暗号化(*angou-ka*)[1 user]
ground	グランド(*gurando*)[2 users], 土地(*tochi*)[1 user]
specification	仕様(*shiyou*)[3 users], スペック(*supekku*)[2 users]

Table 3: Translations by different users

This kind of data is problematic if the specified translations belong to the same field because they can be regarded as contradictory knowledge. If we were to utilize this type of translation pairs in improving the system dictionary, we would have to check them manually to decide which translation should be given priority over others.

7 Conclusion

We have started to provide our Internet machine translation service for both English to Japanese and Japanese to English, in which users can easily send a translation request from a web page through a CGI protocol. Any people can use our service if their personal computer is connected to the Internet and has an e-mail tool; otherwise, there is no special software that needs to be installed.

Our system has several advantages over other similar systems. First, the translation pairs which users specify when making a translation request are accumulated in the server, so that they can semi-automatically construct their user dictionaries on the server. At the same time, the system providers can collect these accumulated data to use them as translation knowledge. Second, the system automatically classify the translation pairs obtained from users by referring to the choice of technical term dictionaries users made when requesting translation. Third, the system helps users in compiling dictionary data by its extraction function, i.e., by providing a list of words which the system recommends to enter in user dictionaries.

With these novel features, we have succeeded in automatically collecting as many as 5,860 translation pairs for English to Japanese and 1,728 for Japanese to English, which can be used to refine and update the system dictionary. The analysis of the data shows that it contained a great amount of new words in the Internet and information science fields, which could be added to our technical term dictionaries.

References

MT Avenue translation service.
http://mtave.softpark.jplaza.com/MTave/

FLM network translation service.
http://trns.cab.infoweb.or.jp/

AltaVista Translation with SYSTRAN.
http://babelfish.altavista.digital.com/

JST Translation Network Service.
http://www-jmt.jst.go.jp/

Kumano, A. and Hirakawa, H. (1994). “Building an MT Dictionary from Parallel Texts Based on Linguistic and Statistical Information”. In Proceedings of COLING 94.

Nakayama, K. and Kumano, A. (1997). “Collecting Dictionary Data from Internet Translation Service Users”. In Proceedings of the fourth annual meeting of the Association for Natural Language Processing in Japan. (in Japanese)