Collaborative Translation Environment on the Web

Sayori Shimohata, Mihoko Kitamura, Tatsuya Sukehiro, and Toshiki Murata

Oki Electric Industry Co., Ltd.
1-2-27 Shiromi, Chuo-ku, Osaka 540-6025
Japan
{shimohata245,kitamura655,sukehiro564,murata656}@oki.co.jp

Abstract

This paper describes a comprehensive translation environment build on the Internet. This environment is designed not only to translate web pages but also to support translation work on the web. We first introduce a basic idea and implementation of this environment and then compare it to conventional machine translation (MT) systems available on the web and translation memories.

Keywords

Translation on the web, dictionary building tools, management by communities

1 Introduction

There are many translation systems on the web (JEIDA, 01). Once you designate a URL of a web page, you will get the same page translated into your favourite language. Since these systems are mainly used to understand the gist of the web pages, their engines and dictionaries are rather light and small compared with general-purpose MT systems.

Recently, however, demand for the translation quality is growing even in such MT systems. This is because of a diversity of web documents. Small MT dictionaries cannot cover the vocabulary dependent on the various domains. To achieve high-quality translation, carefully compiled domain dictionaries are indispensable. But most of these MT systems have no alternative dictionaries on the web.

Another trend in web translation is localization through network community. A group of interested people translates web pages and technical reports to get up-to-date information before it appears in print (LINUX, JPNIC, GNU). This type of translation is often carried out on a voluntary basis. We call this "collaborative translation" hereafter. For such group work, an online translation workbench is necessary. However, few MT systems support such translation work on the web.

To meet demands of both types of translation, we build an innovative translation environment on the web. This aims to provide high-quality translation for light translation and a user-friendly, efficient translation environment for collaborative translation. In the following sections, we will discuss a basic idea of collaborative translation environment and implementation of necessary functions. Then, we will compare the proposed system with MTs available on the web and translation memories and conclude with future directions.

2 Collaborative Translation Environment

Collaborative translation is often carried out by a group of people who are not professional translators but have a good knowledge of the specific subject. Since the group members have different backgrounds and degrees of translation ability, how to share translation knowledge is the key factor in success of the translation project.

Sharing translation knowledge means to:

- Standardize terms and styles of translation
- Store correct translations and reuse them effectively
- Make a lively exchange between the members.

These are essential to keep translation quality and improve translation productivity. Once you start collaborative translation, you realize the need of these. Under the present situation, many collaborative translation projects utilize their homepages and mailing lists for that purpose.

We propose a collaborative translation environment using MT on the web. Figure 1 shows system architecture of this environment. Users access this environment through the Internet.

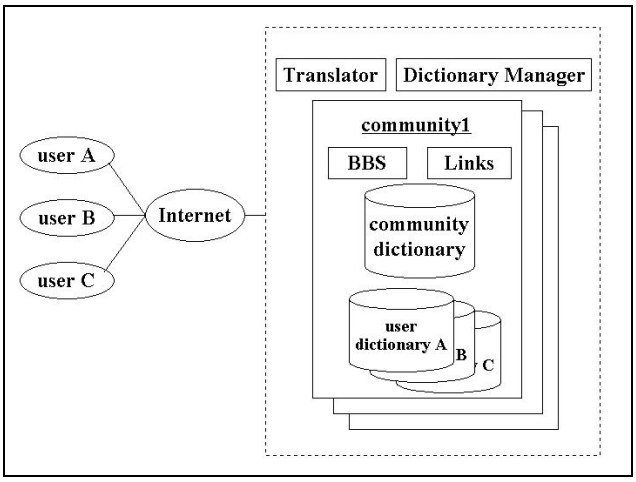

Figure 1: System Architecture of Collaborative Translation Environment

Management by Communities

In this translation environment, we introduce "a community" as a unit of translation domain. A community is a particular group of people who are all interested in a

specific subject or participate in a specific translation project. Community members have a lively exchange of opinions on the related subject.

The management of a community is entrusted to a system operator. For example, a community dictionary is developed in cooperation with community members, but a system operator has authority to decide the entries. A new community is established in response to a user's request. When establishing a community, community members can decide to make the community open or closed.

Figure 2: Top Page of the Translation Environment

Figure 2 is an image of the top page. This page is a portal heading down the community path. Communities are categorized according to the subject into the tree structure.

3 Implementation

The translation environment provides 2 basic functions, Translator and Dictionary Manager, and 3 community functions, Community Dictionary, Bilingual Bulletin Board System (BBS) and Links to the related web pages. Users (not only community members) can access to the basic functions from a top page and each community site. In addition to basic functions, community members can utilize the community functions to communicate with each other. In this section, we will explain the implementation of each function.

Translator

Translator has 3 translation modes according to the input: Web translation translates a designated URL, text translation translates sentences in a text box as a translation button is clicked, and file translation translates a designated file in a text or html format.

Users go on to the translation page from a top page and from a community site. When users make translation from a top page, only a user dictionary is selected in translation. As far as users make translation from a top page without any options, it is just like conventional web translation systems. When users make translation from a community site, a user dictionary and community dictionaries concerned are automatically selected in translation. (Details about a user dictionary and a community dictionary are mentioned in the following paragraph.) Other community dictionaries are also available if users assign them manually.

The engine of the Translator is Japanese-English MT system PENSEE which is implemented with JAVA (Shimohata et al. 99). PENSEE employs a pattern-based transfer approach that uses a set of bilingual patterns shown in Table 1.

English	Japanese
S -> 1:NP 2:VP	S -> 1:NP が(subj marker) 2:VP
NP -> a 1:N	NP -> 1:N
VP -> 1:V 2:NP	VP -> 2:VP を(obj marker) 2:V
VP -> take 1:NP	VP -> 1:NP を(obj marker) する(do)
VP -> take a bath	VP -> 風呂(bath) に(in/at) 入る(enter)
V -> take	V -> とる(take)
N -> bath	N -> 風呂(bath)

Table 1: Examples of bilingual patterns

Pattern-based MT is fit to customize translations easily and effectively. It can represent complicated linguistic phenomena and even correspondences between quite different structures in the languages. This feature enables users to add various kinds of expressions into the dictionaries.

Dictionary Manager

Dictionary Manager handles interface between users and dictionaries. Users can search, add, and modify entries with the Dictionary Manager.

A unique feature for the collaborative translation is that a user can only add entries and leave translation fields blank if he/she is not sure of appropriate translations. Then, other community members fill the translations if they know a correct translation.

To build a dictionary efficiently, the dictionary manager provides 2 dictionary-building tools: Term Extractor and Example Learner. We will explain them in the next section.

Dictionaries

In this translation environment, a user dictionary and a community dictionary are open to the users. Both dictionaries are same in scheme but different in usage.

A user dictionary is used as a user's personal dictionary. Each user has his/her own dictionary. Users add words and phrases to the user dictionaries without restraint. Another usage is that users check the effect of entries before they add them to a community dictionary. A community dictionary is used as a domain dictionary open to the public. Only the community members can add entries but everyone can use that community dictionary.

Bilingual Bulletin Board System (BBS)

Bilingual BBS is available to the community members. This BBS aims to strengthen cooperative spirit and promote smooth translation work through lively discussion on the community related subject. The BBS translates English message into Japanese and vice versa (with our MT system). Therefore, members can post messages both in English and Japanese.

Links to the Related Web Pages

A collection of links is registered in a community main page. All links are closely related to the community and selected by the community members. These links are useful for people who are interested in the subject. The registered links are always translated with the community dictionary.

4 Dictionary Building tools

Every MT user recognizes that building a sufficient domain dictionary is essential to a high-quality translation. But it is an astoundingly time-consuming task for MT users. To support this, the system provides 2 dictionary-building tools: Term Extractor and Example Learner.

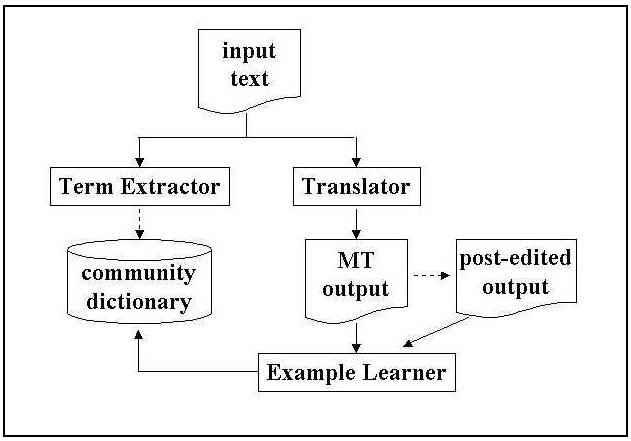

Figure 3: Workflow of Dictionary Building

Figure 3 shows the workflow of the dictionary building process. Broken lines indicate that human work is needed for the process. The output of these tools is stored in the community dictionary.

Term Extractor

Term Extractor extracts technical terms from an input text and makes an entry list for a domain dictionary. The process is composed of the following steps:

- Extracts words and phrases frequently used in the text (Shimohata et. al, 97)
- Analyses their parts of speech (Shimohata, 00)
- Gives correspondent translations by a human translator
- Adds entries to a community dictionary.

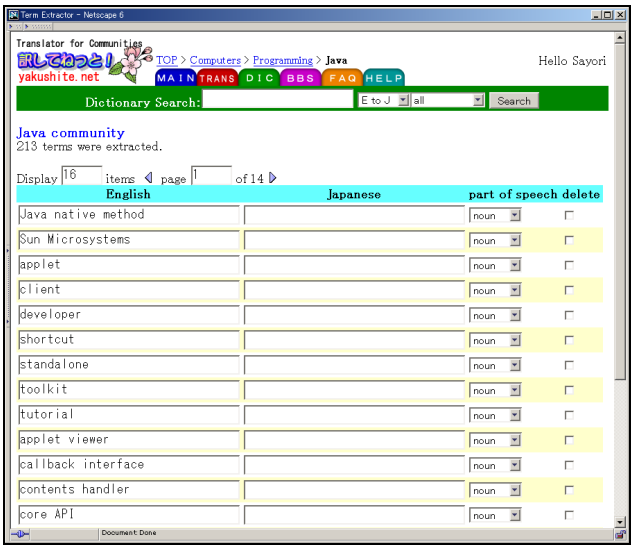

Figure 4: Image of Term Extractor

Figure 4 is a result of the term extraction. Since the input text is monolingual, an entry list is also monolingual. Therefore, users need to give translations to the list. Although we haven't provided alignment tools in the translation environment, it is possible to acquire bilingual entry list if a bilingual corpus is available, (Shimohata et. al, 99). By using Term Extractor, we can pick out entries frequently used in the domain and build a domain dictionary efficiently.

Example Learner

Example Learner learns translation patterns from a difference between MT output and correct post-edited translation. The process is composed of the following steps:

- Translates a source sentence by MT system
- Post-edits MT output by a human translator
- Detects differences between MT output and post-edited sentences
- Generates new translation patterns
- Adds the new translation patterns to the dictionary.

Through the process, only the second process needs human intervention.

Figure 5: Image of Example Learner

Figure 5 shows an example of a source sentence, MT output, and a post-edited sentence. Example Learner parses MT output and post-edited sentences and detects a difference between them in a translation pattern level. In this example, "青い(blue) クジラ(whale)" and "シロナガスクジラ(blue whale)" are the difference.

(1) [en:NP a [1:Adj] [2:N]]
[ja:NP [1:Adj] [2:N]];

(2) [en:Adj blue]
[ja:Adj 青い];

(3) [en:N whale]
[ja:N クジラ];

(4) [en:NP blue whale]
[ja:NP シロナガスクジラ];

Figure 6: Example of Translation Patterns

Figure 6 shows an example of translation patterns used in the parsing process. Translation patterns (1), (2), and (3) are used in the parsing of "青い(blue) クジラ(whale)" while (4) is used in the parsing of "シロナガスクジラ (blue whale)." Example Learner generates a new entry from the translation patterns (4) and adds it to a user dictionary (or a community dictionary). Then, the system translates "a blue whale" into "シロナガスクジラ(blue whale)" correctly next time. Like this, Example Learner stores correct translations from the post-edited sentences.

5 Related Work

Then, we will compare the proposed system with conventional MTs available on the web and translation memories.

MT Systems on the Web

MT systems available on the web are designed to obtain general translation with simple operation. Therefore, few MT systems allow users to choose domain dictionaries or to add their vocabulary to the user dictionaries. As mentioned previously, the proposed system can be used just like these MT systems, but, at the same time, it can be used as a flexible translation tool if users want to control translation quality.

Translation Memories

Translation memories (TMs) such as TRADOS (TRADOS, 01) and STAR (STAR, 01) are widely used in the translation industry. TMs store a large number of translation examples in their database and provide them when a human translator translates the same or a similar sentence. Most TMs adopt TMX (Translation Memory eXchange) format (TMX, 01) to their database. It allows compatibility between TMs. If there is no similar translation in the database, a human translator translates the sentence and stores it for the future use.

TMs provide human translators with various facilities to increase productivity and improve translation quality. On the contrary, conventional MTs, including our former system, aim at fully automated process. Under the present situation, however, we realize that it is difficult to obtain sufficient translation without human aid. Then, the proposed system integrates merits of TMs into MT environment.

In the first stage, we provide a translation environment with dictionary building tools and communication tools on the web. In the next stage, we will adopt TMX format and improve the compatibility of our dictionaries. We believe the trend in this direction goes on in the MT community.

6 Conclusion

We have discussed a collaborative translation environment build on the web. In accordance with the rapid growth of the network community, users' demand toward MT systems has also changed. The proposed system is designed to meet diversified demands of network translation. This environment will be open to the public in the near future. We believe that our system will become widely accepted in the translation community on the web.

References

JEIDA (the Japan Electronic Industry Development Association). (2001). Machine Translation Service on the Internet, In Report on Human Interface technology (pp.87-98) (in Japanese).

Shimohata, S., Sugio, T., & Nagata, J. (1997). Retrieving Collocations by Co-occurrences and Word Order Constraints. In Proceedings of 35th Annual Meeting of the Association for Computational Linguistics (pp.476-481).

Shimohata, S., Murata, T., Ikeno, A., Fukui, T., & Yamamoto, H. (1999). Machine Translation System PENSEE: System Design and Implementation. In Proceedings of the MT Summit VII (pp. 380--384).

Shimohata, S. (2000). An Empirical Method for Identifying and Translating Technical Terminology. In Proceedings of the 18th International Conference on Computational Linguistics (pp. 782-788).

GNUjdoc, http://www.gnu.org/software/gnujdoc/gnujdoc.html

JPNIC (Japan Network Information Center), http://rfc-jp.nic.ad.jp/tojapanese/

LINUX, http://www.linux.or.jp/jman/

STAR, http://www.star-ag.ch/

TMX, http://www.lisa.org/tmx/

TRADOS, http://www.trados.com/