Implementation of Collaborative Translation Environment `Yakushite Net'

Toshiki Murata, Mihoko Kitamura, Tsuyoshi Fukui, and Tatsuya Sukehiro

Oki Electric Industry Co., Ltd.
2-5-7 Hon-machi, Chuo-ku, Osaka 541-0053, JAPAN
{murata656, kitamura655, fukui556, sukehiro564}@oki.com

Abstract

This paper describes an implementation of Collaborative Translation Environment 'Yakushite Net'. In 'Yakushite Net', Internet users collaborate in enhancing the dictionaries of their specialty fields, and the system thus improves and expands its accuracy and areas of translations. In the course of realization of this system, we encountered several technical challenges. We would like to first explain those challenges, and then the solutions to them. Our future plan will also be explained at the end.

1 Introduction

There are many machine translation systems on the Web, and carefully compiled domain dictionaries of MT systems are indispensable in order to achieve high-quality translation of various web documents. However, for companies and individuals who offer this kind of service, maintaining their dictionaries is always a pain in the neck. There are also many human translations on the Web, which are done voluntarily by groups of interested people in cooperation on the Internet.

We focused on the collaborative work among users on the Internet, and invented a Web-based collaborative translation environment named 'Yakushite Net.' 'Yakushite Net' enables people with deep knowledge of particular subjects to collaborate in enhancing the specialized dictionaries for online machine translation, and thus acquires dictionaries with higher accuracy. (Shimohata et al. 2001)

The collaborative translation environment has 3 basic functions; Translation, Post-Editing, and Dictionary Management, and 3 community functions; Community Management, Bulletin Board System, and Q&A. All of these functions are integrated with machine translation system.

Although it is not easy to realize this system because of some technical difficulties, we have found solutions and implemented them in an experimental Web site. We will first describe these difficulties, and then our solutions. Our future plan will also be explained.

2 Technical Challenges in Realization of Collaborative Translation Environment

First of all, it is necessary that collaborative translation environment has capability to handle vast amount of community dictionary entries because the goal of the environment is to collect as many specialized dictionary entries as possible, and to improve and expand its accuracy and areas of translations. In general, the more user dictionary entries are registered, the slower the MT system becomes. A translation environment with great many community dictionary entries is practically useless because of the slowness. Therefore, system that maintains appropriate translation speed regardless of number of entries is needed.

Secondly, immediate reflection of dictionary registration should be realized in order to enable numbers of users to collaborate on the Web in improving the dictionaries. Users should always be able to use updated data to obtain better translations and to avoid redundant registration. In addition, users want to see translation right after their registration in order to check if the registered entries are correctly reflected.

Thirdly, since the environment is available on the Internet, it has to have scalability to keep up with the growing number of users. It is desirable to

simply add machines in accordance with increased number of users, not to modify the system.

Fourthly, variety of components (i.e. BSS, Q&A, etc.) should be integrated with the translation system and community system. For example, BBS contents in a particular language in a certain community should be also available in other languages by using the dictionary of that community.

Finally, security issue must be considered especially because the environment displays the translation of any pages produced by anyone. This means that the environment faces with cross-site scripting security vulnerability.

3 Solutions and Implementation

3.1 Architecture

We devised an architecture as shown in Figure 1 in order to enable numbers of users to collaborate on the Web in improving the dictionaries. Although usual MT systems store entries registered by users only in the special dictionary files (user dictionaries of the systems), we store the entries both in the database management system (DBMS) and the MT's special dictionary files.

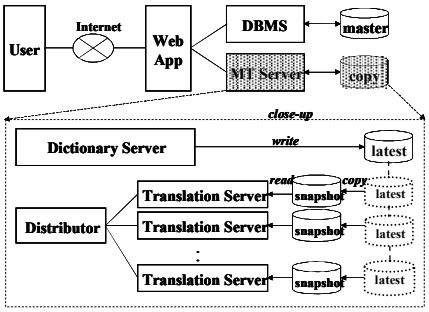

Figure 1 System Architecture of Collaborative Translation Environment

When users register new words, WebApp stores the data in the DBMS and then requests the MT Server to store the data. The data in the DBMS should be the master data since DBMS is more WebApp-friendly, and the data are searched, displayed, and modified by WebApp. MT's special dictionary files are generated from the data in the DBMS, and these files are used only by machine translation system. Data used by users are always of the DBMS.

The reasons why we use both the DBMS and the MT's special dictionary files are:

- The DBMS is not adequate for direct use by the MT system
- The MT's special dictionary files are not adequate for dictionary management.
- During translation, other users may register new entries.

To explain the first reason, the speed of translation becomes much faster when the MT uses its user dictionaries instead of using the DBMS directly. This is because MT's special dictionary files have structure (e.g. Trie) especially suitable for machine translation.

The second reason means that MT's special dictionary files in general do not necessarily have full-text search function, nor can store related information such as user information, creation dates, last-modified dates, and history information, all of which are unnecessary in machine translation process.

The third reason means that inconsistent requirements are inevitably imposed to the system. While user's actions cannot be stopped, all dictionaries the MT uses should not be overwritten during translation.

3.2 Translation and Dictionary Management

Translation and dictionary management contains two mechanisms:

- pattern-based machine translation
- integration with post-editing

3.2.1 Pattern-Based Machine Translation

Pattern-based machine translation method is our key technology (Shimohata et al. 1999) (Kitamura et al. 2003). In this technology, all grammars, word dictionaries, idioms, expressions, and sentences are treated as translation patterns shown in Table 1.

```
[en:Sentence [1:NP]  [2:VP]]
[ja:Sentence [1:NP]は[2:VP]];

[en:N  piano]
[ja:N  ピアノ];

[en:VP  play [1:NP]]
[ja:VP  [1:NP]を弾く];
```

Table 1: Examples of translation patterns

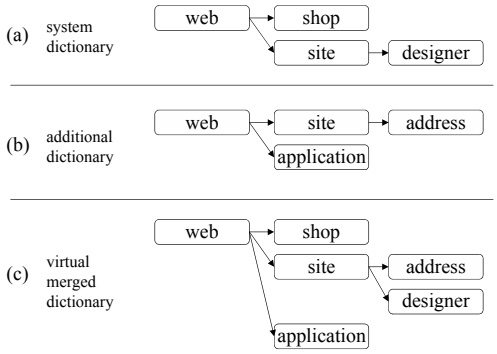

Figure 2 Examples of Trie Structures

The proprietary dictionary of the system has trie structure. This feature is crucial in handling vast amount of community dictionary entries.

Let us assume that two phrases ("web shop", "web site designer") are being registered to a system dictionary. The system stores them using trie structure as shown in Figure 2 (a). Then two other phrases ("web site address", "web application") are being registered to a community dictionary, and the system stores them as shown in Figure 2 (b) as well. When the system uses the system dictionary and the community dictionary, trie structure of selected dictionaries are virtually merged. The system requires only those trie trees shown in the input sentences, therefore the system performance is not affected by amount of community dictionaries or community dictionary entries. Thus we successfully maintain translation speed while handling vast amount of community dictionary entries.

3.2.2 Integration with Post-Editing

The collaborative translation environment has three sources of translation; text translation, web page translation, and file translation, and users can post-edit translation results of any of these sources. The pair of original sentence and the correct post-edited translation is stored as a sentence pair to a community dictionary in the same way shown in Figure 2. Therefore when users translate the same sentence again, the environment can handle any amount of correct post-edited translation while maintaining sufficient translation speed by using the method described in the previous paragraph.

3.3 Scalability

To keep up with the growing number of users, the environment needs to have scalability. In order to acquire scalability, the environment has following two mechanisms:

- distributed machine translation
- dictionary sharing

3.3.1 Distributed Machine Translation

Distributed machine translation mechanism is a technology to translate one document on multiple servers. As shown in Figure 3, multiple translation servers translate a sentence divided from the document (text, web page, or file) simultaneously. When a translation server finishes translating a sentence, it receives a next sentence which no server has translated yet.

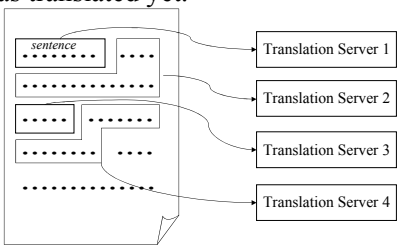

Figure 3 Distributed Machine Translation

When a server machine has multiple CPUs, multi-thread translation system has higher translation speed. In single-thread model, caches of dictionary entries are needed in every process as shown in Figure 4 (a), and the each cache uses significant amount of memory resources. On the other hand, in multi-thread model, multiple threads can share the caches of dictionary entries as shown in Figure 4 (b). Moreover, hit ration of cache increases in this model, which also contributes to translation speed.

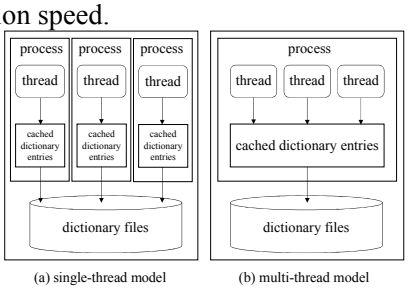

Figure 4 Two thread models

3.3.2 Dictionary Sharing

In pervious section, we mentioned that the collaborative translation environment needs multiple translation servers. This means that a dictionary sharing mechanism is also needed. We invented an architecture as shown in lower part of Figure 1.

First, we equipped the architecture with a dictionary server which is designed solely for dictionary management such as registering, deleting, and updating entries. WebApp calls the server for dictionary management. The server has MT's special dictionary files which are the latest version.

Each translation server, which may be running on other machines, creates snapshots of the latest dictionary files by copying them before using them in translation, then the translation server translates using the snapshots. Thus concurrent processing of registration to a certain community dictionary and translation using the same dictionary is realized. Although these snapshots are older than the latest dictionary, the condition has to be consistent throughout a translating process in one document.

4 Framework

We highly expect that the environment has the expandability to various information systems. For example, let us assume an environment integrated with corporate information systems. For developers of such systems, it is desirable that they can use many different components of the collaborative translation environment without having any knowledge about the mechanisms of the environment. Therefore, as shown in Figure 5, we built the Collaborative Translation Framework. All the components are connected with each other. These components can use information of communities, users and etc., and can translate using community dictionaries. Also, integrators can customize these components for various purposes.

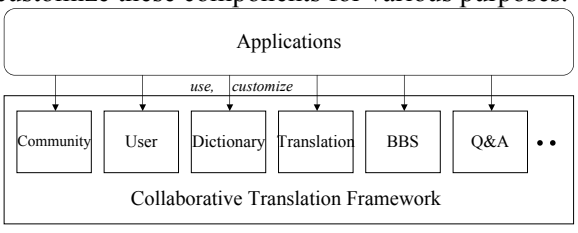

Figure 5 Collaborative Translation Framework

5 Conclusion

We have discussed implementation of the collaborative translation environment. The proposed solutions are implemented as ' Yakushite Net' as shown in Figure 6. This environment will be open to the public in the near future.

In addition, we plan to apply the environment as Web service, develop more rich clients than web browser, and integrate other various kinds of systems. We believe that our system will become widely accepted in the translation community on the web.

Acknowledgement

This research is supported by a grant from Telecommunication Advancement Organization of Japan(TAO).

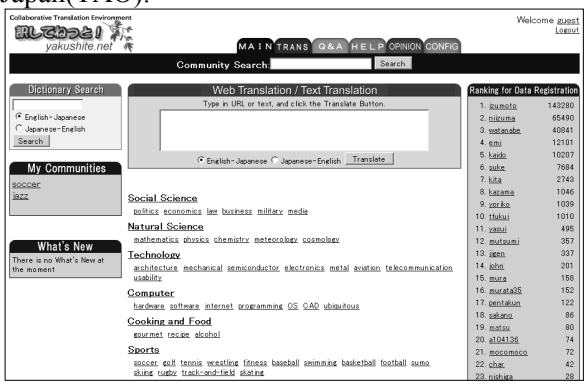

Figure 6 Collaborative Translation Environment ' Yakushite Net'

6 Bibliographical References

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., Kitamura, M., Sukehiro, T., Murata, T. 2001. Collaborative Translation Environment on the Web. In *Proceedings of the MT Summit VIII*, pp. 331-334.

Yakushite Net: http://www.yakushite.net/

Kitamura, M., Murata, T., Sukehiro, T., Shimohata, S., Sasaki, M., Matsunaga, T., Nakagawa, T. 2003. Technology and Development on Collaborative Translation Environment ``Yakushite Net''. In *Proceedings of the 65th Annual Conference of Information Processing Society of Japan(IPSJ)*, pp. (5)319-322.