Translation Memory Sharing Models in XMCAT

Yidong Chen, Xiaodong Shi, Changle Zhou, Tangqiu Li, Qingyang Hong
Institute of Artificial Intelligence, Xiamen University, Xiamen, Fujian, P.R. China
{ydchen; mandel; dozero; tqli; qyhong}@xmu.edu.cn

Abstract

In this paper, two Translation Memory (TM) sharing models adopted in XMCAT, a Computer Assisted Translation tool (CAT) supporting cooperated work in machine translation, was described in detail. One is Center-based TM sharing model, which is only fit for users in a local area network (LAN) and the other is a novel model called P2P-based TM sharing model, which could be used through Internet by geographically distributed users. With the two TM sharing models, a user may share data with other users through network, so that he/she may reduce the repeated work further and cooperate with others more easily. Besides, the methods used in XMCAT to deal with the problem of multi-translations arose in the cooperated memory sharing models, were also proposed in this paper. XMCAT system has been adopted and approved by some translation companies.

Keywords: CSCW, TM, Center-based, P2P-based.

1. Introduction

Translation memory (TM) is defined by the Expert Advisory Group on Language Engineering Standards (EAGLES) Evaluation Working Group's document on the evaluation of natural language processing systems as "a multilingual text archive containing (segmented, aligned, parsed and classified) multilingual texts, allowing storage and retrieval of aligned multilingual text segments against various search conditions." [1] In other words, translation memory (also known as sentence memory) consists of a database that stores the source and target language pairs of text segments in some index mechanism that can be retrieved for use in the translation of new texts to be translated.

Translation Memory (TM) is one of the key technologies used in today's Computer Assisted Translation (CAT) system, such as TRADOS [2] and YaXin CAT [3]. By using TM technology, a translator does not have to re-translate work he/she has already completed.

The advantages of TM could be furthered by using TM sharing technologies. In this paper, the TM sharing models in XMCAT, which is a CAT tools developed by us is described in detail. The TM sharing models will greatly reduce the effort paid by users for vast number of repeated works. At the same time, the TM sharing models will greatly improve the coherence of the translations performed by different translators hens enhance the cooperation between translators.

The rest of this paper is organized as follows. Section 2 gives a brief description of XMCAT. Section 3 describes in detail the two TM sharing models introduced in XMCAT. Section 4 proposes some discussions about how to deal with the problem of multi-translations. Section 5 is conclusions.

2. XMCAT

XMCAT is an English-Chinese and Chinese-English Computer Assisted Translation tool developed by the Institute of Artificial Intelligence, Xiamen University, P.R. China. The system has a strong functionality to assist users to promote productivity. Another feature of XMCAT is that it can be integrated into MS Word, so it can take the full advantage of a friendly user interface and the additional editing functionality provided by Word.

In XMCAT, when a user tries to translate a sentence, three types of translation candidates are provided to him/her for selection and further processing. The three types of translation candidates are: the outputs from an automatic rule-based MT system, the outputs from the template-based MT system and the ones from TM-based translations (See Figure 1).

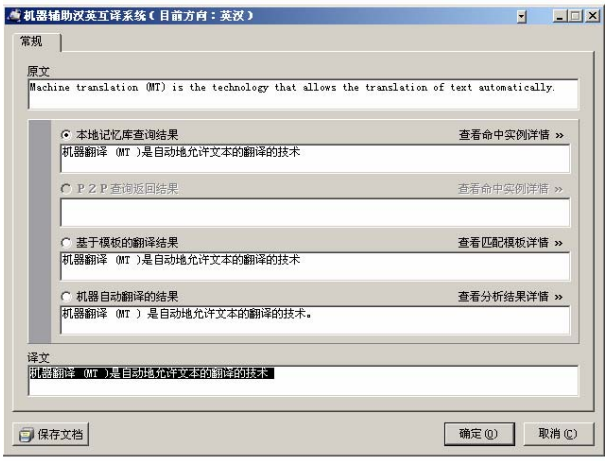

Figure 1. **Multi-types of Translation Candidates**

With multi-types of translation candidates presented, the users have more information about the translation

and then more possibility to produce a desired translation result quickly and conveniently.

XMCAT supports various kinds of user adjusting means, with which a user could modify the translation candidates and generate the correct translation easily. One novel mean provided by XMCAT is the syntax tree adjusting model. Using this model a user could view and adjust the parse results of the rule-based sub-system in a visual way (See Figure 2).

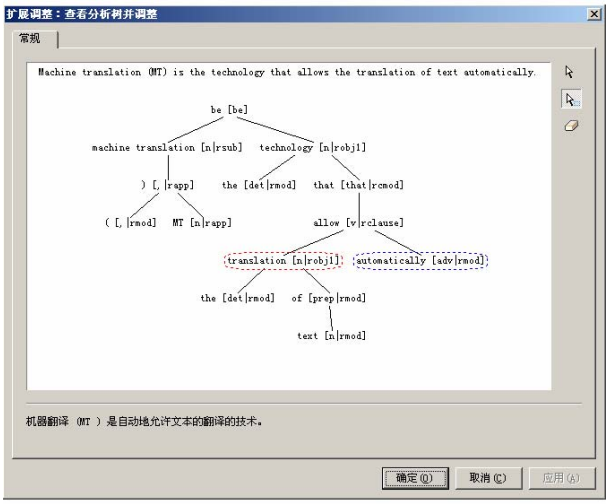

Figure 2. **Syntax Tree Adjusting Model**

As mentioned in Section 1, TM sharing technologies were another powerful means integrated in XMCAT. In addition to the traditional TM technology, two TM sharing models were introduced, one is center-based TM sharing models and the other is P2P-based TM sharing model. The details of these sharing models will be described in Section 3.

3. TM Sharing Models

3.1. Center-based TM Sharing Model

In the Center-based TM Sharing Model, a central server was built to store the shared TM. Before a user start to do his/her translation, he/she must login into the server. Besides the shared TM, each client will maintain its own local TM. The model is shown in Figure 3.

In the center-based TM model, when a user wants to translate a sentence, XMCAT will search both the local TM and the shared TM for translations. The results form difference sources are merged and presented to the user. Then, after the user chooses the correct translation, the translation will be used to update not only the local TM but also the shared TM in the server.

Some famous CAT products [2, 3] also incorporate the center-based TM sharing model in their enterprise versions to further the advantages of TM and enhance the cooperation between translators.

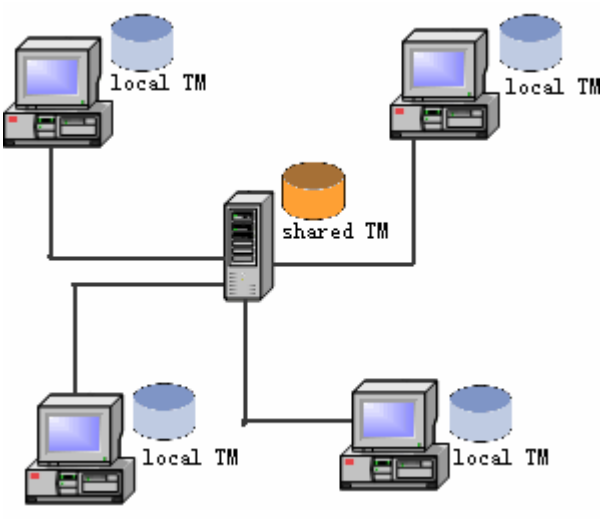

Figure 3. **Center-based TM Model**

3.2. P2P-based TM Sharing Model

The center-based TM sharing model is efficient and can be easily implemented. However, it is only suitable for the local area network (LAN) environment, but not for Internet environment. To fulfill the TM sharing in the Internet environment, we used the P2P model.

P2P is a special distributed system on the application layer, where each pair of peers can communicate each other through the routing protocol in P2P layers. In P2P model, each node takes the role of both client and server. As a client, it can query and download its wanted objects from other nodes (peers). As a server, it can provide objects to other nodes at the same time [4].

In the P2P-based TM sharing model, there are no central servers for storing shared TM. The TMs are distributed in all the P2P nodes (see Figure 4), and are shared among them via message mechanism.

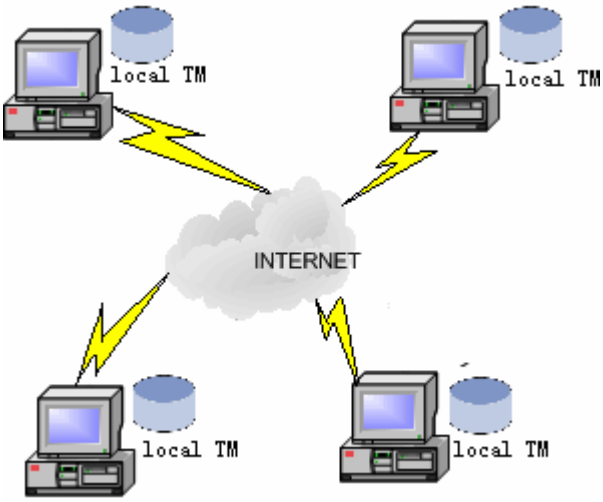

Figure 4. **P2P-based TM Model**

The P2P-based TM sharing model could be realized in the following steps:

Firstly, when a node wants to translate a sentence, a corresponding translation request will be broadcast to all the other nodes in the P2P network.

Then, the sending node will wait for the responses from other nodes.

When a node receives a translation request from other nodes, it will try to search its local TM for translations and response to the sender if it successfully gets some results.

Finally, after the waiting time expired, the sending node will collect all the responses from other nodes and merge them, then present them to the user together with its local translation results.

In the process described above, a node may only wait for 3-4 seconds for the responses, so the responses that do not arrive timely should be dropped. We use a simple timestamp mechanism to do with this problem: every translation request is assigned with a timestamp before it is sent, and a respondent should assign the same timestamp of the request to the corresponding response. Since no timestamps are equal as far as a node is concerned, a node could drop the responses with timestamps unequal to the current timestamp.

It is rather expensive to establish our own P2P network, so we adopt an easier way, that is, to use a P2P network that previously existed. We finally decided to implement our P2P-based TM sharing model based on MSN, because it is very popular and has open and well-designed client protocol.

By using MSN network, TMs are only shared among the users who are MSN friends of one another. All the actions such as requests, responses and negotiations are implemented simply by sending and receiving MSN messages. This way, all the MSN friends can perform their cooperated translation through internet, no mater how they are geographically distributed.

4. Multi-Translations Problem

Given a source sentence, a TM system will usually return multi-translations, so the user often have to pick up the best one from a large number of the alternatives, which is a time-consuming process. After the TM sharing models were introduced, this problem seems more serious, since more translations will be found for a given source sentence.

In XMCAT, we used some ranking methods to deal with this problem. A smart ranking model could be helpful since it can always put into the front the better one that will usually be preferred by users. In the following sub-sections, we will address these methods respectively.

4.1. Ranking by the Meta-Information

The first ranking method is based on the meta-information. The meta-information of a given translation memory record (TMR) includes the information such as *who* uses the record, *when* the record is used or *how often* the record is used, and etc.

In XMCAT, a TMR is represented as a 5-tuple, i.e., (*Src*, *Tgt*, *UID*, *LT*, *Freq*), where *Src* is the source sentence or segment, *Tgt* is the corresponding target sentence or segment, *Freq* is the total times the record was selected, *UID* is the user id of the highest-ranked user who used the record, and *LT* is the timestamp for the latest use of the record.

Thus, we may address the rule for compare two TMRs as follows:

TMR_1=(Src_1, Tgt_1, UID_1, LT_1, $Freq_1$) is said to be more preferred than TMR_2=(Src_2, Tgt_2, UID_2, LT_2, $Freq_2$) if the user rank of UID_1 is higher than that of UID_2 or else if LT_1 is larger than LT_2 or else if $Freq_1$ is larger than $Freq_2$.

Given the rule above, we could rank the multi-translations easily.

The idea behind the rule is that the advice of more important persons in the translating group may be more reliable, and the newer TMR should be more preferred, and the more frequently used TMR should also be more preferred.

The following points about the rule should be noted:

Firstly, in the case of the P2P-based TM sharing model we only used *UID* information, because it is complicated to synchronize the *LT* and *Freq* information among nodes in this case.

Secondly, the ranks of the users should be predefined by users involved in the translating.

Thirdly, we adopt *LT*-first but not *Freq*-first because we believe that the latest used translation may also be the more preferred one for the current document.

4.2. Ranking by the Terminologies

The second ranking method is based on a pre-defined terminology list. The ranking process includes the following steps:

Firstly, the users involved in the same translation project should negotiate one another and define a list containing the key terminologies (in target language).

Then, the list is used to rank the multi-translations that matched the source sentence. The more the terminology contained in the list is found in a translation, the more the translation will be preferred.

This method is especially suitable for the situations that many translators work together to translate a large document.

4.3. Ranking by the Context

The third ranking method is based on the context. Here, a context is a list of key terminologies which are recently translated.

The basic ranking technique used in this method is the same as that of the method in Section 4.2. However, this method differs from the one in Section 4.2 in that it does not require a pre-defined terminology list. The

terminology list used in this method is constructed dynamically.

Please note that, only *N* sentences before a sentence are used to extract its context information. The context window size *N* is set to be 8 in XMCAT. And only several special classes of words are looked on as important. In XMCAT, nouns and verbs are reserved as key terminologies.

5. Conclusions

In this paper, we described in detail the two TM sharing models adopted in our CAT system, XMCAT. The sharing models could help users to reduce the repeated work and enhance their cooperation. We also proposed some methods we used in XMCAT to deal with the problem of multi-translations, which becomes more important when using sharing models. XMCAT has been used in some translation companies and received many approvals and praises. Moreover, it achieved the *2005 Sci-Tech Progress Award (III) of Fujian Province*.

Acknowledgement

This work was supported by the National Natural Science Foundation of China (Grant No. 60573189), National 863 High-tech Program (Grant No. 2006AA01Z139), Natural Science Foundation of Fujian Province (Grant No. 2006J0043) and the Fund of Key Research Project of Fujian Province (Grant No. 2006H0038).

References

[1] EAGLES (Evaluation of Natural Language Processing Systems), "Benchmarking translation memories", *Document EAG-EWG-PR.2 (http://issco-www.unige.ch/ewg95)*, Sept. 1995.

[2] TRADOS, *http://www.trados.com/*.

[3] YaXin CAT, *http://www.yiba.com/*.

[4] Chonggang Wang and Bo Li, "Peer-to-Peer Overlay Networks: A Survey", *Technical article of Department of Computer Science, Hong Kong University of Science and Technology*, April 2003.