Helping Volunteer Translators, Fostering Language Resources

Masao Utiyama
MASTAR Project
NICT
mutiyama@nict.go.jp

Takeshi Abekawa
National Institute
of Informatics
abekawa@nii.ac.jp

Eiichiro Sumita
MASTAR Project
NICT
eiichiro.sumita@nict.go.jp

Kyo Kageura
Tokyo University
kyo@p.u-tokyo.ac.jp

Abstract

This paper introduces a website called *Minna no Hon'yaku* (MNH, "Translation for All"), which hosts online volunteer translators. Its core features are (1) a set of translation aid tools, (2) high quality, comprehensive language resources, and (3) the legal sharing of translations. As of May 2010, there are about 1200 users and 4 groups registered to MNH. The groups using it include such major NGOs as Amnesty International Japan and Democracy Now! Japan.

Figure 1: Screenshot of "Minna no Hon'yaku" site (`http://trans-aid.jp`)

1 Introduction

This paper introduces a website called *Minna no Hon'yaku* (MNH, "Translation for All", Figure 1), which hosts online volunteer translators (Utiyama et al., 2009).[1] Its core features are (1) a set of translation aid tools, (2) high quality, comprehensive language resources, and (3) the legal sharing of translations.

First, the translation aid tools in MNH consist of the translation aid editor, QRedit, a bilingual concordancer, and a bilingual term extraction tool. These tools help volunteer translators to translate their documents easily as described in Section 3. These tools also produce language resources that are useful for natural language processing as the byproduct of their use as described in Section 4.

Second, MNH provides comprehensive language resources, which are easily looked up in QRedit. MNH, in cooperation with Sanseido, provides "*Grand Concise English Japanese Dictionary*" (Sanseido, 2001) and plans to provide "*Grand Concise Japanese English Dictionary*" (Sanseido, 2002) in fiscal year 2010. These dictionaries have about 360,000 and 320,000 entries, respectively, and are widely accepted as standard and comprehensive dictionaries among translators. MNH also provides seamless access to the web. For example, MNH provides a dictionary that was made from the English Wikipedia. This enable translators to reference Wikipedia articles during the translation process as if they are looking up dictionaries.

Third, MNH uses Creative Commons Licenses (CCLs) to help translators share their translations. CCLs are essential for sharing and opening translations.

[1] Currently, MNH hosts volunteer translators who translate Japanese (English) documents into English (Japanese). The English and Japanese interfaces are available at `http://trans-aid.jp/en` and `http://trans-aid.jp/ja`, respectively.

Beijing, August 2010
Proceedings of the 2nd Workshop on "Collaboratively Constructed Semantic Resources", Coling 2010, pages 63–66,

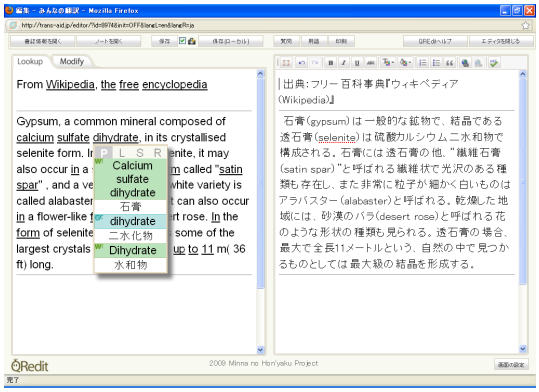

Figure 2: Screenshot of QRedit

Figure 3: Screenshot of bilingual concordancer

2 Related work

There are many translation support tools, such as Google Translator Toolkit, WikiBABEL (Kumaran et al., 2009), BEYtrans (Bey et al., 2008), Caitra (Koehn, 2009) and Idiom WorldServer system,[2] an online multilingual document management system with translation memory functions.

The functions that MNH provides are closer to those provided by Idiom WorldServer, but MNH provides a high-quality bilingual dictionaries and functions for seamless Wikipedia and web searches within the integrated translation aid editor QRedit. It also enables translators to share their translations, which are also used as language resources.

3 Helping Volunteer translators

This section describes a set of translation aid tools installed in MNH.

3.1 QRedit

QRedit is a translation aid system which is designed for volunteer translators working mainly online (Abekawa and Kageura, 2007). When a URL of a source language (SL) text is given to QRedit, it loads the corresponding text into the left panel, as shown in Figure 2. Then, QRedit automatically looks up all words in the SL text. When a user clicks an SL word, its translation candidates are displayed in a pop-up window.

[2] `http://www.idiominc.com/en/`

3.2 Bilingual concordancer

The translations published on MNH are used to make a parallel corpus by using a sentence alignment method (Utiyama and Isahara, 2003). MNH also has parallel texts from the Amnesty International Japan, Democracy Now! Japan, and open source software manuals (Ishisaka et al., 2009). These parallel texts are searched by using a simple bilingual concordancer as shown in Figure 3.

3.3 Bilingual term extraction tool

MNH has a bilingual term extraction tool that is composed of a translation estimation tool (Tonoike et al., 2006) and a term extraction tool (Nakagawa and Mori, 2003).

First, we apply the translation estimation tool to extract Japanese term candidates and their English translation candidates. Next, we apply the term extraction tool to extract English term candidates. If these English term candidates are found in the English translation candidates, then, we accept these term candidates as the translations of those Japanese term candidates.

4 Fostering language resources

Being a “one stop” translation aid tool for online translators, MNH incorporates mechanisms which enable users to naturally foster important translation resources, i.e. terminological resources and translation logs.

4.1 Terminological resources

As with most translation-aid systems, MNH provides functions that enable users to register their own terminologies. Users can assign the status of availability to the registered terms. They can keep the registered terms for private use, make them available for a specified group of people, or make them publicly available. Several NGO groups are using MNH for their translation activities. For instance, Amnesty International, which uses MNH, maintains a list of term translations in the field of human rights by which translators should abide. Thus groups such as Amnesty upload a pre-compiled list of terms and make them available among volunteers. It is our assumption and aim that these groups make their terminological resources not only available among the group but also publicly available, which will create win-win situation: NGOs and other groups which make their lists of terms available will have more chance of recruiting volunteer translators, while MNH has more chance of attracting further users.

At the time of writing this paper (May 2010), 56,319 terms are registered, of which 45,843 are made publicly available. More than 80 per cent of the registered terms are made public. Currently, MNH does not identify duplicated terms registered by different users, but when the number of registered terms become larger, this and other aspects of quality control of registered terms will become an important issue.

4.2 Translation corpus

Another important language resources accumulated on MNH is the translation corpus. As mentioned in the introduction, being a hosting site, MNH naturally accumulates source and target documents with a clear copyright status. Of particular importance in MNH, however, is that it can accumulate a corpus that contains draft and final translations made by human together with their source texts (henceforth SDF corpus for succinctness). This type of corpus is important and useful, because it can be used for the training of inexperienced translators (for instance, the MeLLANGE corpus, which contains different versions of translation, is well known for its usefulness in translator training (MeLLANGE, 2009)) and also because it provides a useful information for improving the performance of machine translation and translation-aid systems. While the importance of such corpora has been widely recognized, the construction of such a corpus is not easy because the data are not readily available due to the reluctance on the side of translators of releasing the draft translation data.

The basic mechanisms of accumulating SDF corpus is simple. Translators using MNH save their translations to keep the data when they finish the translation. MNH keeps the log of up to 10 versions of translation for each document. MNH introduced two saving modes, i.e. snapshot mode and normal mode. The translation version saved in the normal mode is overwritten when the next version is saved. Translation versions saved in snapshot mode are retained, up to 10 versions. Translators can thus consciously keep the versions of their translations.

MNH can collect not only draft and final translations made by a single translator, but also those made by different translators. MNH has a function that enables users to give permission for other translators registered with MNH to edit their original translations, thus facilitating the collaborative translations. Such permission can be open-ended, or restricted to a particular group of users.

This function is of particular importance for NGOs, NPOs, university classes and other groups involved in group-based translation. In these groups, it is a common process in translation that a draft translation is first made by inexperienced translators, which is then revised and finalized by experienced translators. If an inexperienced translator gives permission of editing his/her draft translations to experienced translators, the logs of revisions, including the draft and final versions, will be kept on MNH database.

This is particularly important and useful for the self-training of inexperienced translators and thus potentially extremely effective for NGOs and other groups that rely heavily on volunteer

Figure 4: Comparative view of different translation versions

translators. Many NGOs face chronically the problem of a paucity of good volunteer translators. The retention rate of volunteer translators is low, which increase the burden of a small number of experienced translators, leaving them no time to give advice to inexperienced translators, which further reduce the retention rate of volunteers. To overcome this vicious cycle, mechanisms to enable inexperienced volunteer translators to train themselves in the cycle of actual translation activities is urgently needed and expected to be highly effective. MNH provides a comparative view function of any pairwise translation versions of the same document, as shown in Figure 4. Translators can check which parts are modified very easily through the comparative view screen, which can effectively works as a transfer of translation knowledge from experienced translators to inexperienced translators.

At the time of writing this paper, MNH contains 1850 documents that have more than one translation versions, of which 764 are published. The number of documents translated by a group (more than one translator) is 110, of which 48 are published. Although the number of translations made by more than one translators is relatively small, they are steadily increasing both in number and in ratio.

5 Conclusion

We have developed a website called *Minna no Hon'yaku* (MNH, "Translation for All"), which hosts online volunteer translators. We plan to extend MNH to other language pairs in our future work.

References

Abekawa, Takeshi and Kyo Kageura. 2007. QRedit: An integrated editor system to support online volunteer translators. In *Digital humanities*, pages 3–5.

Bey, Y., K. Kageura, and C. Boitet. 2008. BEYTrans: A Wiki-based environment for helping online volunteer translators. Yuste, E. ed. *Topics in Language Resources for Translation and Localisation*. Amsterdam: John Benjamins. p. 139–154.

Ishisaka, Tatsuya, Masao Utiyama, Eiichiro Sumita, and Kazuhide Yamamoto. 2009. Development of a Japanese-English software manual parallel corpus. In *MT summit*.

Koehn, Philipp. 2009. A web-based interactive computer aided translation tool. In *ACL-IJCNLP Software Demonstrations*.

Kumaran, A, K Saravanan, Naren Datha, B Ashok, and Vikram Dendi. 2009. Wikibabel: A wiki-style platform for creation of parallel data. In *ACL-IJCNLP Software Demonstrations*.

MeLLANGE. 2009. Mellange. `ttp://corpus.leeds.ac.uk/mellange/ltc.`tml.

Nakagawa, Hiroshi and Tatsunori Mori. 2003. Automaic term recognition based on statistics of compound nouns and their components. *Terminology*, 9(2):201–209.

Sanseido. 2001. *Grand Concise English Japanese Dictionary*. Tokyo, Sanseido.

Sanseido. 2002. *Grand Concise Japanese English Dictionary*. Tokyo, Sanseido.

Tonoike, Masatsugu, Mitsuhiro Kida, Toshihiro Takagi, Yasuhiro Sasaki, Takehito Utsuro, and Satoshi Sato. 2006. A comparative study on compositional translation estimation usign a domain/topic-specific corpus collected from the web. In *Proc. of the 2nd International Workshop on Web as Corpus*, pages 11–18.

Utiyama, Masao and Hitoshi Isahara. 2003. Reliable measures for aligning Japanese-English news articles and sentences. In *ACL*, pages 72–79.

Utiyama, Masao, Takeshi Abekawa, Eiichiro Sumita, and Kyo Kageura. 2009. Hosting volunteer translators. In *MT summit*.