Management and use of terminological resources for distributed users in the translation hosting site Minna no Hon'yaku[1]

Takeshi Abekawa, National Institute of Informatics
Masao Utiyama and Eiichiro Sumita, National Institute of Information and Communication Technology
Kyo Kageura, University of Tokyo, Japan

In this demonstration, we show the terminology management module of Minna no Hon'yaku (MNH: http://trans-aid.jp/), a translation hosting site with integrated translation-aid mechanisms, which was made publicly available in April 2009. As of February 8th, 2010, 1062 users have registered with MNH and more than 3400 documents have been translated, of which more than 1600 translations have been published on the site. On MNH, users can translate documents individually or can define groups and share the translation task. It provides users with functions such as lookup of high-quality dictionaries and terminologies, seamless access to Wikipedia and Google search, and reference to TM. There are two types of terminological resources on MNH, i.e. those provided by the system and those registered by users. The demonstration shows how terms are registered, shared and used.

1. Introduction

This demonstration shows the translation hosting site Minna no Hon'yaku (MNH: translation of/for/by all) accessible at http://trans-aid.jp/, placing special emphasis on the mechanisms that enable users to manage and make use of terminological resources. MNH was made public on April 7, 2009. As of February 8th, 2010, 1062 users have registered with MNH, 3430 documents have been translated using the translation-aid functions provided by MNH, and of these 1630 have been published on the MNH site. Currently MNH accommodates the English-to-Japanese and Japanese-to-English language pairs, but the coverage is to be extended to English-to-Chinese, Chinese-to-English, Japanese-to-Chinese and Chinese-to-Japanese in near future.

MNH provides functions to aid individual translators, including an automatic flexible lookup of high-quality dictionaries and terminology resources, seamless access to Wikipedia and Google search, and access to translation memory (TM). Registered users can also constitute groups, within which members can share translation tasks and user-defined resources. Terminologies are a critical resource for translation and their management is especially important because, together with TM, many translators or translator groups keep their own lists, and update and maintain them in the course of translation. This poses a challenge to any open translation-aid platform on which translation groups may be defined less systematically than in a well established translation environment.

In the following, we give a brief overview of MNH in section 2, and summarise the nature and status of terminological resources in a distributed user environment in section 3. Section 4 outlines the terminology management functions of MNH, first at the registration phase and then at the use phase. Section 5 outlines the prospects for MNH.

[1] This work is partly supported by the Japan Society for the Promotion of Sciences (JSPS) grant-in-aid (A) 21240021 'Developing an integrated translation-aid site which provides comprehensive reference sources for translators' and by the 2009 National Institute of Informatics (NII) research cooperation project 'Construction and use of practical terminological resources from a variety of information sources.'

2. MNH: A brief overview

The basic functions provided by MNH are as follows (Utiyama et. al. 2009):

1. anybody can register with MNH anonymously, and is provided with her/his personal space;
2. users can publish their translations on the MNH site, if copyright permits;
3. a variety of social networking functions are provided, including social tagging, message exchange, question and answer, translation request, etc.;
4. users can define a group on MNH, in which they can co-edit translations, share registered terms, share translation memories;
5. register terms, upload and manage terminologies, register translation memory database;
6. search translation texts, translated sentence pairs (TM), translators, tags, and registered terms.

Translators who register with MNH can produce translations by using QRedit. QRedit is a two-pane translation-aid editor incorporated in MNH (Figure 1), which provides the following functions for online translators (Abekawa & Kageura 2007; Takeuchi et. al. 2007):

1. flexible (idiom variations can be matched to dictionary entries), stratified (important or difficult multi-word elements are emphasised) lookup in and copy-and-paste from a high-quality dictionary (三省堂 2001), some free dictionaries, and terminologies;
2. seamless connection to Wikipedia monolingual and bilingual entries;
3. seamless connection to Google search;
4. function to register terms in the process of translating and immediately enable their lookup;
5. an easy-to-use and effective interface which enables users to concentrate on translation.

Note that users can lookup terms not only within QRedit but also from the search box provided by MNH. From the point of view of process, lookup in QRedit constitutes an integral part of translation, while lookup from MNH is one step away from the translation process itself.

3. The nature and role of terminological resources

Terminological resources tend to be varied, compared to ordinary dictionaries (ordinary dictionaries are numerous, but they are categorised into a few general groups). There are some standard terminological lexicons for some domains, while other domains have no coherent terminological lexicons. Translators, subject specialists and in-house terminologists tend to maintain terminologies of their own or of the group they work for. For instance, Amnesty International Japan, which uses MNH, maintains its own list of terms to be used in translation. The position of these terminological resources from the point of view of usage may vary as well. Sometimes the use of translations of terms given in certain lists is obligatory. Sometimes the lists are used to provide translators with possible candidates, which are not obligatory.

Once a translation of a term is chosen, it should be used consistently, to maintain the quality of translations throughout a text, a group of texts, or a domain. The importance of terminology management in translation-aid systems has long been recognised (Hutchins 1998), and most major TM or translation-aid tools provide terminology management modules. Unlike commercial TM or terminology management tools, most of which assume well

organised and managed users, MNH is open to anybody and thus cannot make such assumptions. This creates challenges for the functional design of terminological management.

Figure 1. An image of translation-aid editor QRedit

Considering these factors, it is necessary for the terminological management module on MNH to allow the stratification of terminological resources, possibly according to the type of resource and to users' requirements. We can roughly classify four types of terminological resources on MNH: (i) coherent and edited terminologies, which have more or less the same status as general dictionaries, and are provided by the system itself; (ii) an open-ended depository of bilingual terms registered by MNH users (which include (iii) and (iv)); (iii) terminological lists managed by specific groups such as Amnesty International; and (iv) personal lists of terms registered by individual MNH users. Many group users need organisation and coordination of terminologies, while individual users should preferably be able to choose what to use and how among these resources. The end result is a loosely connected network for sharing user-registered information.

4. Management of terminological resources on MNH

4.1. Terminological resources provided by MNH

Well established terminologies of various domains can be provided by MNH administrators, just like general high-quality dictionaries. Currently, bilingual entries from Wikipedia are provided. A terminology of the legal domain, a semi-automatically constructed proper name bilingual lexicon (Sato 2009), and a large scale, semi-automatically compiled terminological lexicon (Abekawa & Kageura 2009) are to be provided as well. Users can freely lookup these terminological lexicons from QRedit. If they do not wish to consult some of these resources, they can disable their look-up from the page 'selecting target dictionaries/terminologies for QRedit lookup' on MNH.

4.2. Terminological resources registered by users

Users can register bilingual terms at two different phases, (a) in the process of translation when they are using QRedit, and (b) at the MNH terminology management page (see section 2). In QRedit, users can register terms only one by one. The MNH terminology management page provides users with three ways of registering terms: (i) register terms one by one, in the same way as registering terms on QRedit; (ii) upload a pre-compiled list of terms; and (iii) edit, select and register automatically extracted bilingual term candidates from translation texts specified by users (users can select translation texts accumulated on MNH by translator, by a variety of tags attached to translations, or by pre-defined groups of translations). At the time of registration or at a later stage, users can designate the status of terms as 'personal use', 'open to a limited range of users', and 'open to all MNH users'.

The terms registered by users can be looked up either from the MNH search box or directly from QRedit in the process of translating (section 2). For the purpose of explanation, let AP be the set of terms registered by the user A with 'personal use' status, AL the set of terms registered by A with 'open to a limited range of users' status, and AA the set of terms registered by A with 'open to all the MNH users' status. The user A can look up all AP, AL and AA immediately both from MNH and from QRedit. For users other than A, the accessibility of AP, AL and AA is as follows:

AP: not accessible, either on MNH or on QRedit;

AL: becomes accessible on MNH and on QRedit only when user A gives explicit permission for use by specific users. To make the terms available, therefore, A must take two steps: giving terms the status of 'limited use' and then giving permission for use to specific users.

AA: becomes immediately accessible by all users on MNH but becomes accessible in QRedit only when A gives permission to specific users (which means that the accessibility within QRedit is essentially the same as AL).

Currently, the status control of registered terms is only given to the supply side. The demand side (users who are given permission for use by A) does not have a mechanism to block the lookup of terms which have been made accessible. This is because we originally designed the mechanism to deal with shared terminology lookup as required by such NGOs as Amnesty International; like professional translators working with companies, volunteer translators working with such NGOs as Amnesty International must follow the prescribed usage of certain terms. So we made the demand side open - when the person in charge of terminology gives permission for translators in the same group, they must be able to look them up. On the other hand, if, in the absence of the demand side control, AA were to become directly accessible by all users within QRedit in the process of translation, it would become overwhelming; hence the setting of AA as above. This setting is fragile for spam terms. For instance, if a user registers a large number of wrong bilingual terms and gives permission to other users, they will have to face wrong translations in the QRedit small pop-up window for reference lookup. Recognising that this risk may increase as the number of users grows, we are currently developing a control mechanism for terms from the demand side.

5. Prospects

In this paper, we have explained the management and use of terminological resources on

MNH and in QRedit. As of February 8th, 2010, the number of registered terms on MNH was 55,508. Of these, 45,723 have been made public by those who registered terms. The growth of registered terms are shown in Figure 2. It is observed that there are periods when a huge chunk of terms are registered at once, which corresponds to the registration of term lists maintained by users.

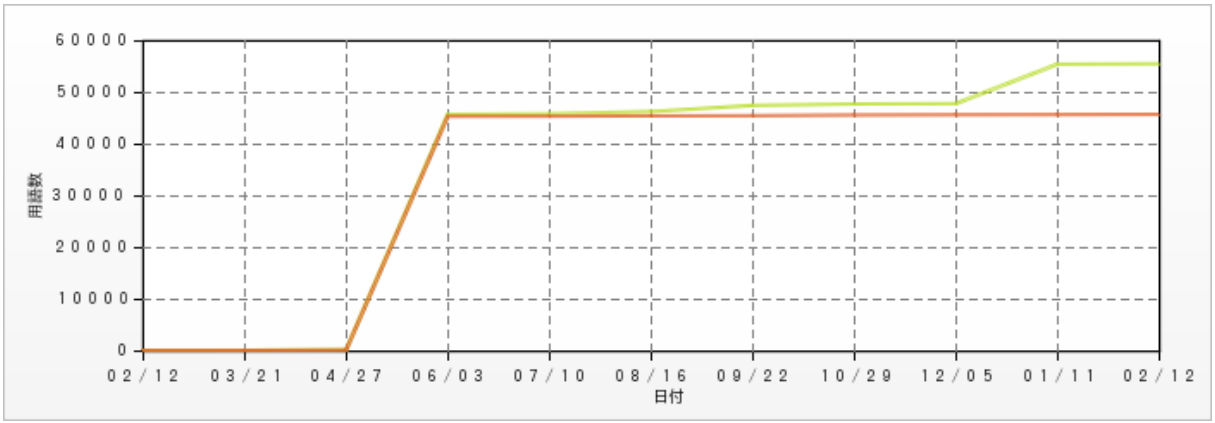

Figure 1. The growth of registered terms

As for their use, although the supply side control of the accessibility of terms currently adopted in MNH assumes the goodwill of users and can be risky, group users are currently making responsible use of this mechanism. Currently we have no problem reports from users with regard to terminology management.

Everything described in this paper is fully functional on the MNH site; the demonstration will thus be able to show the mechanisms of terminology management within the actual workflow of online translation by individual and group translators. Though the main interface is now in Japanese, an English interface will become available in 2009. Currently, MNH only supports English-to-Japanese and Japanese-to-English translation, but this is due to the limitated availability of high quality dictionaries. If good dictionaries become available, it can be made fully functional for English-to-anylanguage or Japanese-to-anylanguage, without the need for any modification to the current architecture.

Bibliography

Abekawa, T.; Kageura, K. (2007). 'A translation aid system with a stratified lookup interface'. In *ACL 2007 Demo and Poster Sessions*. 5-8.

Abekawa, T.; Kageura, K. (2009). 'QRpotato: A system that exaustivelly collects bilingual technical term pairs from the Web'. In *The 3rd International Universal Communication Symposium*. 115-119.

Hutchins, J. (1998). 'Computer-based translation tools, terminology and documentation in the organizational workflow: report from recent EAMT workshops'. In *International Conference on Professional Communication and Knowledge Transfer*. 255-268.

Nakagawa, H.; Mori, T. (2003). 'Automatic term recognition based on statistics of compound nouns and their components'. In *Terminology* 9 (2). 201-219.

三省堂グランドコンサイス英和辞典. 東京: 三省堂, 2001.

Sato, S. (2009). 'Crawling English-Japanese person-name transliterations from the Web'. In *WWW* 2009. 1151-1152.

Takeuchi, K.; Kanehila, T.; Hilao, K.; Abekawa, T; Kageura, K. (2007). 'Flexible automatic look-up of English idiom entries in dictionaries'. In *MT Summt* XI. 451-458.

Utiyama, M.; Abekawa, T.; Sumita, E.; Kageura, K. (2009). 'Hosting volunteer translators'. In *MT Summit* XII.

Utsuro, T.; Kida, M.; Tonoike, M.; Sato, S. (2006). 'Collecting novel technical terms from the Web by estimating domain specificity of a term'. In Matsumoto, Y.; Sproat, R.; Wong, K-F.; Zhang, M. (eds.). *Computer Processing of Oriental Languages: Beyond the Orient: The Research Challenges Ahead*. Berlin: Springer. 173-180.