# Sharing Dictionaries among MT Users By Common Formats and Social Filtering Framework

**Shin-ichiro Kamei**

Human Language Technology Group
C&C Media Research Laboratories
NEC Corporation, Japan
kamei@ccm.cl.nec.co.jp

## Abstract

MT users have to build "user dictionaries" in order to obtain high-quality translation results. However, building dictionaries needs time and labor. In order to meet the speed of the information flow in the global network society, we need to have common formats for sharing dictionaries among different MT systems, and a new way of dictionary authorization, that is "social filtering".

## 1 User dictionaries for MT systems

It is essential for MT users to build and use their own "user dictionaries" in order to use MT systems effectively. MT systems have "system dictionaries" which cover basic words and phrases, but these entries are not sufficient for translation in actual situations. Real documents contain topical and technical terms in special fields, and translation quality depends strongly on these terms. However, most of them are not registered in system dictionaries in advance, since they are produced daily in various fields.

Hence users need to make new entries for these terms in their own dictionaries and build their own user dictionaries. This is a basic and effective way for obtaining high-quality translation results. Quantity and quality of dictionaries controls the quality of translation results.

However, building dictionaries needs much time and labor. Users have to know equivalents of unregistered words and process of making a user dictionary that differs from an MT system to another, and have to spend much time to make new entries of each word and phrase. This means it is difficult for an individual user to make his/her dictionary large and good enough to meet the needs.

In order to solve this problem, we need a new framework for building and using MT dictionaries. 1n particular we have to change our way of thinking for authorization of dictionaries. In addition, environments for sharing user dictionaries among users who utilize different MT systems play an important role.

## 2 Authorizing dictionaries by Social Filtering

In general, dictionaries have been approved by authorities of linguistics and published by famous publishers. Users have relied on these authorities when using dictionaries. However, these authorized dictionaries are often useless for MT users in the present day and in future. New words and phrases in documents which users want to translate are not registered in these dictionaries, since speed of generating new terms is getting faster and faster these days, and this way of making dictionaries takes too much time for MT systems.

In order to catch up with the speed of the multi-lingual information flow in the global network society, we have to find another way of authorizing dictionaries. One of the choices is voluntary building and social filtering of dictionaries.

A trial of exchanging dictionaries was executed by one of MT vendors in Japan, NEC. They have opened a web site [1] for users of their MT products, and users

have been exchanging their own user dictionaries through **it.** Users can use dictionaries at the web site with their MT systems. Also, users can put their own user dictionaries on the web site and share them with other users.

Within a year since the web site opened, more than ten enthusiastic users have uploaded the dictionaries they have made. The dictionaries cover various fields, such as cuisine, SF (Science Fiction), sports, computer technology, Fl (car race), proper nouns of persons, places, companies, etc. Words in these fields are difficult to obtain from usual dictionaries. Many users of the products have downloaded the user dictionaries and used them on their personal MT systems.

What we want to emphasize here is that this environment for accumulating and exchanging dictionaries depends not on traditional authorization but on "social filtering". There was an incident where a user uploaded some inappropriate entries on the web site, but these entries were soon omitted within a week, since other users have pointed out the inadequacy.

This means that this voluntary way of building dictionaries is effective in using MT systems. A large volume of dictionaries in various fields are accumulated on the web site. Individual users only have to build dictionaries in the fields they are familiar with, and are able to use dictionaries in other fields which is created by other users. The social filtering framework makes the qualities of these dictionaries reliable.

## 3 Common formats for sharing dictionaries

As reported before [2], we, MT providers in Japan, as members of the Asia-Pacific Association for Machine Translation (AAMT), have established environment for sharing and exchanging user dictionaries among different MT systems. We have defined common formats of user dictionaries "UPF (Universal PlatForm)," and established electronic facilitators available for users to exchange their user dictionaries [3]. This task was supported by the Foundation of IPA (Information-technology Promotion Agency, Japan).

The trial of sharing dictionaries described above is for the same MT system of a single company. MT users are able to exchange and share user dictionaries more widely when they use the UPF format and environments, since more than five MT vendors in Japan including MediaVision, Sharp, Nova, Toshiba, and NEC at present have embarked their dictionary converters which convert dictionaries in the UPF formats from/to dictionaries in their own formats.

Common formats such as UPF have advantages for users, venders, and dictionary vendors of MT systems. Users can obtain high-quality translation results without much effort. MT vendors can reduce costs for building their own dictionaries in various fields. Dictionary vendors can obtain a wider market, since they only have to make dictionaries in a single format which are used by many people.

Dictionary entries can be accumulated and used again later once they are created. Users are able to use dictionaries described in common formats for comparing qualities of MT systems, and reuse the entries when they replace their old MT systems with new one. This means the dictionaries will promote sound competitions and growth of MT systems.

Also, the dictionaries will influence a wide range of fields related to languages. The accumulated dictionary entries for MT systems, that is the accumulated multilingual knowledge, can be used in other fields of natural language processing, language education, and so on.

## 4 Conclusion

User dictionaries are essential for using MT systems effectively. In order to reduce costs for users to build their own user dictionaries, we need common formats for sharing dictionaries among different MT systems, such as UPF which is defined by AAMT, and social filtering framework in place of traditional authorization of dictionaries. Accumulated dictionaries will promote growth of MT systems and will be a basis of a wide range of language-related fields.

## References

[ 1 ] http://meshplus.mesh.ne.jp/CRV2/dic/club/down.html (in Japanese)

[2] Shin-ichiro Kamei, Etsuo Itoh, Mikiko Fujii, Tokuyuki Hirai, Yukari Saitoh, Masahito Takahashi, Tsutomu Hiyama, and Kazunori Muraki. (1997). "Sharable Formats and their Supporting Environments for Exchanging User Dictionaries among Different MT Systems As a part of AAMT Activities". In proceedings of MT Summit VI.

[3] http://www.jeida.or.jp/aamt/upf/Upfindex.html (in Japanese)