SHARABLE FORMATS AND THEIR SUPPORTING ENVIRONMENTS FOR EXCHANGING USER DICTIONARIES AMONG DIFFERENT MT SYSTEMS AS A PART OF AAMT ACTIVITIES

Shin-ichiro KAMEI(*1), Etsuo ITOH(*2), Mikiko FUJII(*3),
Tokuyuki HIRAI(*4),Yukari SAITOH(*5), Masahito TAKAHASHI (*6),
Tsutomu HIYAMA(*7), and Kazunori MURAKI(*1)

(*1) NEC Corp.
4-1-1, Miyazaki, Miyamae-ku, Kawasaki, 216 Japan,
(*2) TOSHIBA Corp.
3-22, Katamachi, Fuchu, Tokyo 183 Japan,
(*3) NOVA Inc.
1-29-1 Takadanobaba, Shinjuku-ku, Tokyo, 169 Japan,
(*4) SHARP Corp.
492 Minosho-cho, Yamatokoriyama-shi, Nara, 639-11 Japan,
(*5) Fujitsu Laboratories Ltd.
4-1-1, Kamikodanaka, Nakahara-ku, Kawasaki 211 Japan,
(*6) Kyushu Matsushita Electric Co., Ltd.
2-4-16, Momochihama, Sawara-ku, Fukuoka 814 Japan,
(*7) NEC Informatec Systems, Ltd.
KSP R&D 6F, 3-2-1, Sakado, Takatsu-ku, Kawasaki 213 Japan

kamei@ccm.cl.nec.cl.jp, etsuo@sp.tokyo-sc.toshiba.co.jp, aka@nova.co.jp,
nnd6@isl.nara.sharp.co.jp, yukari@ling.flab.fujitsu.co.jp,
takahasi@mmm.kme.mei.co.jp,
hiyama@hum.cl.nec.co.jp, and muraki@ccm.cl.nec.co.jp

Abstract

We, machine translation providers, as members of Asia-Pacific Association for Machine Translation (AAMT), are now establishing environments for sharing and exchanging user dictionaries among different machine translation systems. In order for users to utilize machine translation systems more effectively, we define common formats of user dictionaries, and establish electronic environments available for users to exchange their user dictionaries using these common formats. This task started in 1996, and the formats will be fixed in March of 1998.

1. Introduction

This paper presents an AAMT activity for defining common formats and establishing of electronic environments for sharing and exchanging user dictionaries among different machine translation systems.

It has been more than 10 years since commercial machine translation products appeared in Japan. At first, machine translation systems were designed on main frame computers for professional translators, to be utilized for translating technical documents. However, recently machine translation systems have been rapidly penetrating into the market of non-professional users in Japan, due to the popularization of personal computers and global electronic networks. Today, more than twenty (20) companies in Japan develop their own machine translation products, and about five hundred thousand (500,000) software packages are sold in a year. Most of these systems are priced between about ten thousand yen (= about a hundred US dollars) and a hundred thousand yen (= about a thousand US dollars). In addition, machine translation systems are pre-installed in most popular personal computer sets (more than six million (6,000,000) personal computer sets are sold in Japan per year.). Machine translation systems are one of the essential tools for personal computers along with global networks such as the Internet.

It is necessary for users to build and use their own user dictionaries in order to use machine translation systems effectively. In general, there are three ways

for users to improve the quality of translation results. The first is pre-editing the input sentences. To shorten and simplify input sentences makes translation results better. However, pre-editing needs long experiences to obtain high skills and its methods differ according to each system. Although pre-editing is effective, it is difficult to accumulate knowledge which can be used again by the machine. This means that pre-editing is difficult for users to share their knowledge. The second way is post-editing of translation results. To proofread and rewrite the result sentences is necessary for obtaining high-quality translation. However, users have to read both input sentences and translation results. This means the post-editing is a time-consuming process, which is valid only when the translation quality of systems is high enough.

Contrary to the two ways mentioned above, the third way of users' participation is to enrich entries of user dictionaries. It is a basic and effective way for obtaining high-quality translation results. Quantity and quality of dictionaries are the basis of the quality of translation results. In addition, dictionary entries can be accumulated and used again later once they are created. They also can be shared by different users. However, building dictionaries needs much time and labor. Users have to know equivalents of unregistered words and have to know how to make a user dictionary that differs from an MT system to another, and have to spend much time to make entries of each word. This means it is difficult for an individual user to make his/her dictionary large and good enough.

In order to solve this problem, we have begun establishing environments that enable sharing each user's dictionary by multiple users. If we have environments for exchanging different types of user dictionaries among different machine translation systems, the cost for building dictionaries of each user will be reduced substantially. This kind of environments promote the use of machine translation systems, and encourage users to communicate in foreign languages.

Our task is supported by the Foundation of IPA (Information-technology

Promotion Agency, Japan). We began to define common formats for exchanging user dictionaries among different MT systems last year, in 1996, and we are now establishing electronic environments for users to exchange their user dictionaries using the common formats. At present, we defined the basic parts of the formats, and made the first version of dictionary editors for the formats and converters which from/to dictionaries in the formats to/from user dictionaries in individual machine translation systems. We are now examining specification of extended formats. We are planning to fix the formats and make the total environments available to the public in March of 1998.

2. Basic Policy of Universal Platform (UPF)

In order to exchange user dictionaries among different machine translation systems, we develop the following environments.

(1) common formats of user dictionaries

(2) electronic environments available to the public for accumulating and sharing user dictionaries in the common formats

We call these environments Universal PlatForm (UPF). The conditions we adopted for defining UPF are as follows.

(a) Compatibility with real machine translation products

(b) Bi-directionality of conversion from/to dictionaries in UPF common formats to/from dictionaries in the formats specific to machine translation systems

(c) Human readability of the UPF common formats

We define common formats of user dictionaries by comparing different user dictionaries of several kinds of machine translation systems that are in the market and are actually utilized by users. This makes the formats compatible with real systems so that they can be used actually. Each machine translation system provider offers converters that transform user dictionaries in the common formats into dictionaries in its own format (download) , and transform user dictionaries in its own format into dictionaries in the common formats

(upload). After evaluation using converted dictionaries with real machine translation systems, UPF formats will be fixed and be opened to the public.

We will create a WWW home page as an environment to accumulate and exchange user dictionaries . Users can use dictionaries in the home page with their machine translation system after converting the formats. Also, users can share their own user dictionaries with other users (including users who use different kinds of machine translation systems) after converting the formats and put them in the home page. In addition, we offer an editor for making entries of user dictionary in the common formats. Users can directly build dictionaries in the common formats.

At present, we are working on two languages, Japanese and English, for defining actual formats and features of the dictionaries considering use and needs of machine translation in Japan. However, the formats themselves are designed for multilingual dictionaries.

SGML-like tags are adopted for description of dictionary features from the viewpoint of readability. UPF formats are for user dictionaries for machine translation systems. However, we think dictionaries should be able to read easily without any special tools so that users including usual human translators who are not familiar with machine translation systems can create new entries of user dictionaries. In other words, the user dictionaries in UPF formats can be utilized directly without machine translation systems. We think this is the key to accumulation and utilization of user dictionaries. For this reason, we decided that master dictionaries of UPF is in a plain text format. Users can write and read UPF dictionaries without any special tools, but can also use UPF editors we provide for convenience.

Figure 1 shows the overview of the UPF concept. Electronic environments for sharing user dictionaries are called ‘H2-Environments (Helpful and High-quality Environments)’, which consist of a WWW home page and dictionary editors for UPF formats.

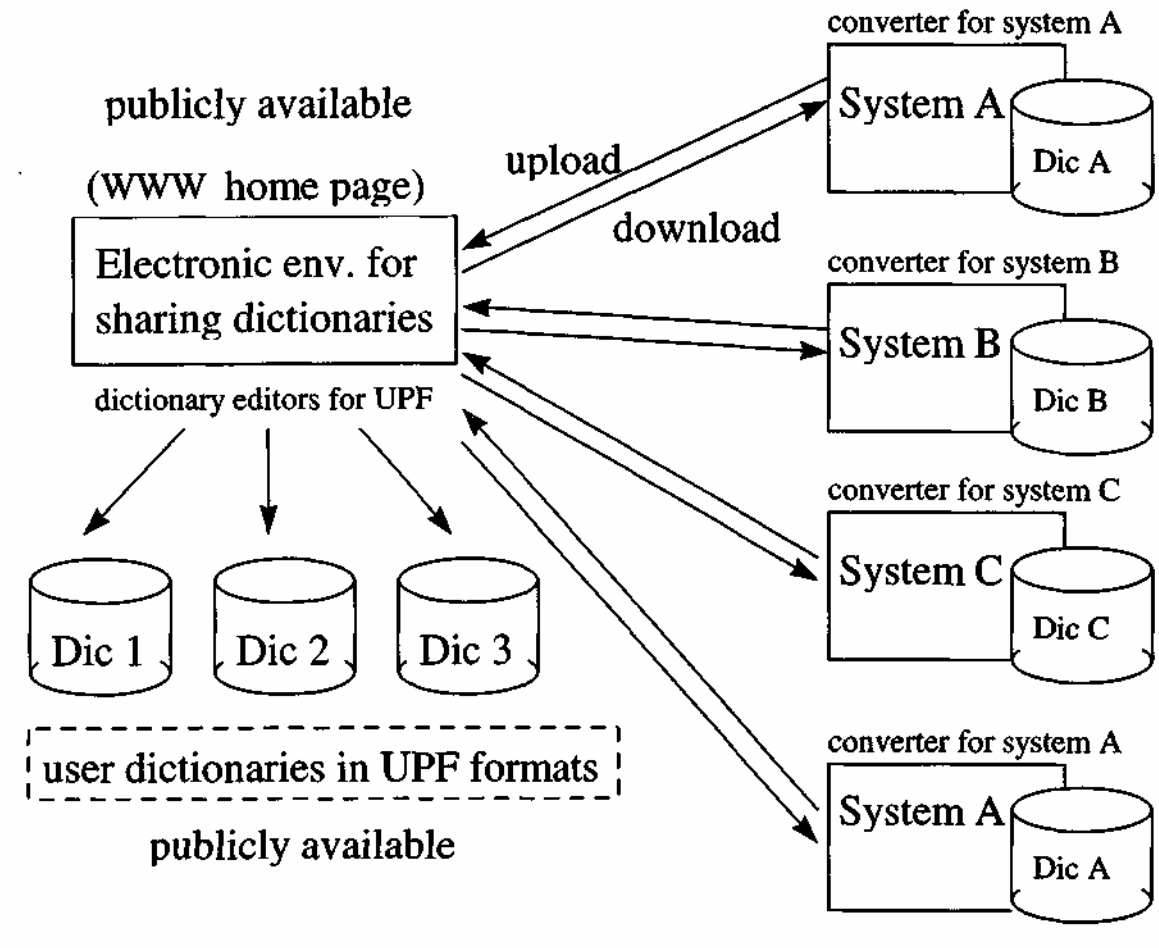

Figure 1: Overviews of UPF Concept

3. UPF Common Formats of MT User Dictionaries

3.1 Basic format (μ-sheet) and extended format (κ-sheet)

This section describes how we actually define UPF common formats of user dictionaries. The condition we have to consider is bi-directionality of conversion, that is, uploading and downloading. However, we easily found that parts of speech and dictionary features of each user dictionary of each machine translation system differ from each other. If we take the bi-directionality of conversion in a narrow sense, parts of speech and dictionary features of UPF formats become intersection of all machine translation systems. However, this is too narrow to handle various linguistic information in the dictionaries. In order to solve this problem, we define two formats:

(1) A basic conversion format (Minimum sheet; μ-sheet)

(2) An extended conversion format (Keystone sheet; κ-sheet) .

μ-sheet (a basic conversion format) is a format that all machine translation systems can handle. Bi-directionality is strongly recommended for μ-sheet conversion. κ-sheet (an extended conversion format) is a format for describing all information in user dictionaries of machine translation systems.

3.2 Steps of fixing UPF basic formats (μ-sheet)

For defining a basic format (μ-sheet), we compared user dictionaries of five (5) machine translation products which were actually sold and used. First thing we had to do was to fix common technical terms for describing linguistic information, such as parts of speech and other features. What is important here is that even definitions of parts of speech vary from an machine translation system to another. For example, so-called 'na'-adjective or 'keiyo-doshi' (ex. 'kirei-na hana' = beautiful flowers) in Japanese can be treated linguistically as a kind of adjectives, as a kind of nouns, or as an independent part of speech. In fact, some systems treat this as a kind of adjectives and other systems treat this as an independent part of speech. We took into consideration standard grammars that are taught at school, and modified these from the viewpoints of preciseness and easiness of word registration.

The second step is selection of parts of speech for μ-sheet (UPF basic format). There are many parts of speech in general, but we selected some parts of speech for μ-sheet after considering the following conditions of μ-sheet.

(1) Frequency of entries in user dictionaries
(2) Bi-directionality of conversion
(3) Easiness of users' understanding

The fact shows that about ninety percent (90%) of entries of user dictionaries that are actually built in the past ten years are nouns and proper nouns. In addition, almost all the systems should handle parts of speech in the μ-sheet. Systems are requested to convert their own user dictionaries into user dictionaries described by the μ-sheet, and convert user dictionaries by the μ-sheet into user dictionaries in their own formats. Thus, when selecting parts of

speech and their features to be described in the μ-sheet, we emphasize easiness and effectiveness for actual use rather than linguistic strictness. The following is a set of parts of speech in the case of Japanese and English.

Japanese: meishi(noun), doshi(verb), keiyoshi(adjective),
keiyo-doshi('na'-adjective) , fukushi(adverb)

English: noun, verb, adjective, adverb
(Note: Noun includes proper noun)

In addition, we restrict the pairs of Japanese and English parts of speech. It is true that Japanese parts of speech corresponds to varieties of English parts of speech in general, but in most of cases each of parts of speech has its major counterpart in the other language. This is the reason we restrict the pairs of Japanese and English parts of speech as follows:

Japanese		English
meishi	$\longleftrightarrow$	noun
doushi	$\longleftrightarrow$	verb
keiyoshi	$\longleftrightarrow$	adjective
keiyo-doshi	$\longleftrightarrow$	adjective
fukushi	$\longleftrightarrow$	adverb

μ -sheet requires minimum linguistic features for user dictionaries for users' convenience. The following is features each of Japanese parts of speech has to have in the μ -sheet dictionary.

[Meishi (noun)]
type: common noun, proper noun
semantic category: human, organization, other concrete objects,
place, time, other abstract objects

[Doshi (verb)]
inflection type: ichi-dan, go-dan, ka-hen, sa-hen (including 'sa-hen doshi')
case frame:
semantic restriction of case frame:

case mapping to English case frame

[Keiyoshi (adjective)]

case frame: (restrict to only 'ga')

semantic restriction of case frame:

case mapping to English case frame

[Keiyo-doshi (na-adjective)]

inflection type: na-da, no-da

case frame: (restrict to only 'ga')

semantic restriction of case frame:

case mapping to English case frame

[Fukushi (adverb)]

(no extra information to describe in the case of μ-sheet)

An example of dictionary description using μ-sheet of UPF formats is as follows:

```
<entry>
<japanese>
<jentry>    食べる (taberu)    </jentry>
<jpos>      動詞 (verb)        </jpos>
<jinfl>     一段 (ichidan)     </jinfl>
<jcase>     が(ga), を(wo)     </jcase>                    </japanese>
<trans> (が(ga) = subject; noun phrase; human,)
        (を(wo) = object; noun phrase; other concrete objects;)  </trans>
<english>
<eentry>    eat     </eentry>
<evpresent> eats    </evpresent>  <evpast>  ate   </evpast>
<evpp>      eaten   </evpp>       <eving>   eating  </eving>
<ecase>     SVO     </ecase>
<jheadpron> vowel   </eheadpron>                        </english>
</entry>
```

In μ -sheet, translation directions are not restricted. Dictionary entries in this format can be used in both Japanese-to-English and English-to-Japanese translations. Other parts of speech, other combinations of English and Japanese parts of speech, and more detailed linguistic features described above are treated in the extended format, κ-sheet. κ-sheet is in a sense, the union of the information in various dictionaries. Thus when downloading, converters select features which are usable to particular machine translation systems.

4. Conclusion

We, UPF executive committee, are now defining common formats for exchanging user dictionaries among different machine translation systems. Our purpose is to encourage machine translation users to build their own user dictionaries and to share them in order to obtain high quality translation result with less efforts. As electronic environments for building and exchanging user dictionaries, we create a WWW home page, **http://www.meshnet.or.jp/aamt-upf**, and offer editors of user dictionaries in UPF in the home page. After confirming the effectiveness of common formats by examining translation results of actual systems using user dictionaries which are converted from common formats, we will fix UPF and make it available to the public in March of 1998. At the same time we offer English-to-Japanese and Japanese-to-English dictionaries which have twenty thousand (20,000) entries each. We, as AAMT members, strongly recommend for machine translation providers to develop their converters and integrate them in their machine translation systems in the near future. We really hope our task will contribute to the popularization of machine translation systems, and will help users to exchange their individual know-how, as well as promote information circulation in foreign languages.