Introduction to UTX, a Specification for a Shared User Dictionary

Okura Seiji†, Yamamoto Yuji††, Murata Toshiki‡, Uchimoto Kiyotaka‡‡,
Michael Konin Kato+, Shimazu Miwako*, Suzuki Tsugiyoshi**, Francis Bond‡‡

Sharing/Standardization Working Group, MT Research Committee,
Asia-Pacific Association for Machine Translation

†Fujitsu Laboratories Ltd., ††CosmosHouse, ‡Oki Electric Industry Co., Ltd., +Learning consultant
‡‡National Institute of Information and Communications Technology,
***Toshiba Solutions Corporation, **Cross Language Inc.**

1. Foreword

In order to use a machine translation system for computer aided translation, careful tuning of user-created dictionaries is indispensable. These user dictionaries, however, may not be always compatible between different MTs, often rendering the effort to create such dictionaries futile. To address this issue, AAMT (Asia-Pacific Association for Machine Translation) [1] has undertaken to establish a specification of sharable dictionaries, which can be used across different machine translation systems. AAMT created its first version of specification, UPF, with support from IPA (Information-technology Promotion Agency, an institute in Japan) in 1995. In 2006, AAMT started to create a new specification to reflect and incorporate the subsequent advancement of technology and the changing usage of MT. In 2007, the new format received a new name "UTX," short for universal terminology exchange. As of 2008, AAMT is working to establish UTX-Simple, which is the simple, stripped-down version of UTX before building the full XML version. As for the future plan in 2008, AAMT plans to advance the project by revising and expanding UTX specifications, selecting the actual translation domain, collecting and compiling the terminological data of the domain, and seeking collaboration with translation-related projects led by other organizations or individuals. In addition, AAMT is conceiving of creating a user community for producing, sharing, and accumulating user dictionaries in a sustainable way. This article introduces the specification of UTX-Simple, and describes our process and methods of sharing and reusing dictionaries.

2. Basic of UTX

As I mentioned earlier in the foreword, careful tuning of user-created dictionaries is indispensable in order to use a machine translation system for computer aided translation. When we use commercial high-end translation software for computer-aided translation workflow, specialized terminology, names of persons, and place names in the target document are often not included in basic system dictionaries, and they are not translated as well as one would expect. These terms may be missing from optional specialized dictionaries. It is known that if these terms are registered into a user dictionary, the precision of a machine translation system can be improved. For example, let's assume that the following English sentence is translated into

Japanese with translation software.

XML declaration may contain information about character encoding and external dependencies.

The result would be:

「XML 公表が文字符号化と外部の属国についてのインフォメーションを含んでいるかもしれません。」

In this translation, certain terms are used, such as 公表, 符号化, 属国, and インフォメーション, but these are not suitable for its context and domain. If the suitable terms are properly registered to a user dictionary, however, a more suitable translation is generated as follows: [2]

「XML 宣言が文字エンコーディングについての情報と外部の依存関係を含むことがあります。」

User dictionary compilation is important, because addition of compound words such as "XML declaration" into a user dictionary decreases failures in syntax analysis[3], and it also improves the accuracy of statistics translation.

However, user dictionary compilation is a very time-consuming process for an individual user, and its effect is also not immediately clear for relatively small translation projects. In addition, the formats of individuals' dictionaries are varied if they do not share the same MT, and sharing these dictionaries can be difficult.

TBX [5] of LISA [4] is an existing standardized specification for terminology, but it is intended to be used by professional terminologists only. Its complex nature denies commitments by average MT users, and it is not widely in use. In order to address these problems, AAMT is establishing a new standard called UTX.

The features of UTX are as follows:

1. **"Dictionary for the user" - simple and easy to use:** A complicated specification merely increases users' burdens, and it will be forgotten eventually. A specification must be simple and practical to reflect and include actual MT users' needs, viewpoints, and scenarios.
2. **Entry as a "technical term":** UTX clearly defines the domain of a dictionary, and adheres to the principle of "one word, one meaning." The number of entries in a dictionary should be well-chosen, and differences of word usages must be clearly defined. An entry must be a unique term within an applicable domain.
3. **Improvement in translation accuracy:** When UTX dictionaries become widespread, sharing dictionaries can be drastically simplified. Users can compile user dictionaries more efficiently by exchanging existing data, and UTX will eventually contribute to improvement in translation accuracy.
4. **Multilingual and monolingual dictionary:** UTX is designed for dictionaries not only between two languages but also among multiple languages. In addition, UTX also includes a specification for a monolingual dictionary, which can be used for proofreading tools for terminological standardization, etc.
5. **Inclusion of the information to support managing and sharing dictionaries:** The XML version of UTX includes detailed information, such as the dictionary creator, entry creation timestamp, information intrinsic to discrete

machine translations, etc., and it enables users to manage and share UTX dictionaries effectively.

6. **Promotion of localization of software:**
   Especially in open source localization projects, translation is carried out individually, and terminological standardization can be difficult. By exchanging and sharing translation resource systematically through the use of UTX, more and more common dictionaries become available, and they increase the efficiency of translation exponentially.

In January, 2008, the European Commission announced that it will release the translation of about 1 million sentences into 22 languages for free. In Japan, the thick language barrier hinders the translation of software and other contents, resulting in immense economic educational, and cultural losses every day. Sharing of translation assets in all industries will be the key to future industrial development.

The major merits of UTX for MT users are: it is a collection of plain formats that are easy to create; it improves the translation accuracy in various domains; sharing and reuse of dictionaries is possible through user communities on LAN or the Internet. The major merits of UTX for translation software manufacturers are: the entire market of machine translation is enlivened by the promotion of user dictionaries; thus new demands and applications of MT can be explored. UTX-XML format retains entry properties that are proprietary to manufacturers, and no data is lost during the conversion.

The following two tasks are important for the specification establishment of UTX.

- Establishment of specifications (UTX-Simple

- A dictionary file consists of one header and one or more entries.
- The header is located in the first line of the file.
- One entry is stored in each subsequent line.
- The line starting with a sharp "#" is a comment line (which only occurs at the beginning of the file).

[Header] The description of the dictionary file. The delimiter is an "en" space.
#UTX-S <version number> <source language>/<target language > <last updated> <other properties>
Example: #UTX-S 0.9 en-US/ja-JP 2007-12-03T14:28:00Z+09:00
* UTX-S means "UTX-Simple". The language name (ISO 639) and the date/time (ISO 8601) formats conform to ISO[2, 3].
In case of a monolingual dictionary, the target language is omitted.
Example of the fourth column:
source:plural/3sp/past/pastp/presp/comparative/superlative/target:...../optionXXXX/optionYYYY
(The information on a source word is described after "source:".) Abbreviation denote as follows: the plural = plural form, 3sp = third person singular, past = past form, pastp = past participle, presp= present participle, comparative= comparative degree, superlative= superlative degree, the translation is described after target: .

[Entry]One line is delimited to columns by tabs. The contents described in each column are as follows:

First column	Second column	Third column	Fourth column and more
Word of source language (Headword)	Word of target language (Translation)	Part of speech	Other properties (optional)

Fig. 1 Specification of UTX-Simple 0.9

and UTX-XML)

- Creation of actual dictionary data and the community for sustainable creation, sharing, and accumulating of dictionaries

In this article, we illustrate these two tasks, as well as the future plan of UTX and its issues.

3. Specification of UTX-Simple

The following two formats of specifications are defined in UTX.

- UTX-XML (XML format)
- UTX-Simple (tab-delimited text format)

UTX-XML (XML format) includes comprehensive information required for a user dictionary. UTX-Simple (tab-delimited text format) is a specification that requires only three types of information: a source word, its translation, and the part of speech of the source word. All other properties are optional. AAMT will eventually establish XML format specification, however, it has started by creating UTX-Simple, which is highly practical and can be easily shared to expand and sustain the community of UTX.

In order to make a specification usable, it is necessary to make it plain and easy to use. In order to encourage the use of user dictionaries with a machine translation system, entry addition must be simpler and more straightforward than the present system. As the UTX-Simple dictionary is simple, it can significantly reduce the time and effort necessary for compilation, and the sharing of dictionaries is accelerated. The specification is designed so that it carries only essential information while keeping its usability. The current version is 0.91. The basic specification is shown in Fig.1.

The following issues were discussed in the process of establishing UTX-Simple:

(1) Should priority be given to readability for human, or to ease-of-processing for machines?

We also discussed about what kind of information to include, and reached to the conclusion that specification requires only three essential types of information: a source word, its translation, and the part of speech of the source word. For example, the information on the plural form of a noun may be required in the translation from/to English. However, the plural form is not an essential piece of information for all the parts of speech of all languages. In some situations, information of the plural forms could improve the translation result, but it also significantly increases the burden of the creator of the dictionary if it were obligatory. In order to ensure smooth creation, we concluded that the information of the plural forms as optional, but not essential.

(2) Comparison of the XML version (UTX-XML) and the simple version (UTX-Simple)

As UTX-Simple does not retain detailed information of its creator, the timestamp for each entry, etc., it is not suitable for a compilation of a permanent, versatile dictionary. However, it is practical and easy to create, use, and share. It may be used by a variety of users regardless of their background knowledge in advanced linguistics. The establishment of the XML version may require some more time.

(3) Should we fix the order of columns (Column X always corresponds to Property Y etc.)?

A column accommodates both information common to all languages, as well as information unique to a specific language. The first to third columns are assigned to three essential properties, while language-specific properties are added after the forth column as they are necessary.

(4) Should we define language-specific columns for every language?

The main languages of focus are currently on Japanese, English, and Chinese. We plan to include all existing languages defined in ISO in the future. It is not realistic or practical to define all columns required for all languages altogether at this time. Therefore, we left the liberty of defining necessary optional columns to users. Feedbacks from actual usages will be incorporated to decide the optimal definitions of properties in the future. We would recommend using certain name for certain property for particular language pairs.

(5) About creation of guidelines

It is necessary to establish guidelines for notation, i.e. how to describe each entry, for each target language. The tools which read UTX-Simple would issue warnings following the guideline, if they find problems in the dictionaries. For example, if one or more mandatory properties are missing, a warning is issued according to the corresponding option of the tool. To cope with human errors, the tool could be set to skip an entire line, if the format has problems.

The following is the basic guideline:

- Difference between a missing value (default) and "not applicable"

  Other missing values (defaults) are subject to handling by MT systems.

  To what extent a tool uses the information contained in a dictionary depends on the specification of the tool itself.
- The user-defined columns which are not identified by the tool will be kept intact.
- An English noun phrase starts with a lower-case letter (except for proper nouns etc.).

An example of UTX-Simple is shown in Fig.2.

4. Creation of a user community

As the basic policy of creating and collecting dictionary data, the dictionary must belong to a specific domain, such as sport, IT, medicine etc.No "generic" dictionary should be permitted.

In open source localization projects, translation is carried out individually, and dictionaries are not shared as they should be. Dictionaries are scattered across various providers, and their licenses and formats are also varied. If these scattered language resources are centralized, the localization between different languageses is significantly accelerated.

In order to spread UTX dictionaries in which anyone can participate in creation, and in order to realize "open dictionaries for everyone," a shared dictionary community should be established. For that purpose, the following efforts are required:

- AAMT will establish two types of dictionary communities for producing, sharing, and accumulating dictionaries. AAMT will also establish a framework for distributing UTX dictionaries.
- The official dictionary community (managed by AAMT or its delegate) offers supervised dictionaries with guaranteed quality for a fee.
- The open dictionary community offers free dictionaries with open source license and promotes mutual exchanges. AAMT or its delegate provides hosting service only, but no management or guarantee.

#UTX-S 0.9 en-US/ja-JP 2007-12-03T14:28:00Z+09:00 source: /target:plural/3sp/past/pastp/presp/comparative/superlative									
syllable	音節	noun	syllables						
new	新規の	adj						newer	newest
go	行く	verb		goes	went	gone	going		
prosody	韻律	noun							

Fig.2 Example of the basic notation of UTX-Simple

In order to meet various needs, official dictionaries and free dictionaries should be distinguished. The corresponding community for managing and sharing each type of dictionary must also be created. The official dictionary community (managed by AAMT or its delegate) offers supervised dictionaries with a guaranteed quality for a fee. Free dictionaries can be used for no fee, although the correctness of the contents is not guaranteed.

Sharing/Standardization Working Group of AAMT has already introducing UTX to open source community by participating in events such as Mozilla 24 (hosted by Mozilla). We exchanged opinions on various topics concering UTX with open source contributers.

Although the format of UTX is simple, it effectively centralizes dispersed language resources through a common format, to be used as a collective intelligence. Therefore, we must maintain and stimulate the motivation for the dictionary creation among the community participant, and create a framework to ensure fairness for the participants with different degrees of contribution.

5. The prospect and issues

We are currently committed to the following tasks: the establishment of specifications and guidelines for creating dictionaries, production, and collection of dictionary data, development of various tools and how to use them, evaluation of the effectiveness of UTX, and collaboration and cooperation with other parties. Eventually, our goal is to make UTX into one of ISO standardizations.

5.1 License system

Make the distribution system of dictionaries simple and clear, both for paid and free dictionaries. Especially, a free/open dictionary must allow mutual use and changes, and its commercial uses need to be permitted. We will start with available royalty-free dictionaries.

5.2 Establishment of UTX-XML

Based on UTX-Simple, the details of the specification of more sofisticated UTX-XML will be determined.

5.3 Establishment of a notation guideline

Notation of dictionaries must be standardized through guidelines for each language and its dialects, if applicable. If notation is not standardized among dictionaries, we must unify them each time we merge one dictionary to another, and this process would be time-consuming. For instance, in case of Japanese, the notation of prolonged sound marks, "middle dots," etc. must be standardized. Equivalent in English would be variations of British and American spellings, punctuation rules, hyphenation etc. In the actual projects of translation, significant time is wasted only to absorb the differences of the notations difined by different parties. Therefore, it is necessary to define and observe the standardization guideline of dictionary notation, if necessary, for different specialized areas.

5.4 Production and collection of dictionary data

We will select some domains for which translation is needed, and build and collect actual dictionary data in accordance with the specification of UTX-Simple and UTX-XML. By carrying out translation with the dictionary and collecting feedbacks, the UTX specification will be further improved.

5.5 Development and use of various tools

We will need to develop following tools:

UTX converters (including parsers)

We will need tools which convert from a format unique to a translation application or a translation site to UTX format, and vice versa. A parser which verifies the conformity to the UTX specification must also be included. In addition, a converter between UTX-XML and UTX-Simple is also required.

Term extraction and dictionary building tools

Tools to analyze multiple documents, extract terms, and add them to a user dictionary [6, 7, and 8] or make a new one instantly, not by building a dictionary one-word-at-a-time.

Dictionary search tool (glossary search tool)

We will need tools to perform a direct search of a dictionary and glossary to see the translation of a word.

5.6 Evaluation of the effectiveness of UTX

In order to promote UTX, it is necessary to prove that the performance and accuracy of translation are improved by the use of UTX. As for evaluation, the effectiveness of a UTX dictionary is due to be checked with test sets created by AAMT.

5.7 Collaboration and cooperation with other parties

Besides the co-authors (which include most of MT developers in Japan) of this article, collaborations with other parties has already been began for UTX, including a cooperation with Oki's community-oriented machine translation site "Yakushite-net" [9] in which a user can add his/her own technical terms. In the future, collaboration and cooperation with Edict, the other projects which connect many dictionaries and systems (such as Language Grid), and organizations (GSK) which collect, manage, and distribute language resources etc. are also envisioned. While providing UTX to other parties and verifying its effectiveness, we are also interested in collaborating in terms of tool developments.

The organizations and individuals who support the UTX project

(In no particular order, titles omitted)

Fujitsu Laboratories, CosmosHouse, OKI, National Institute of Information and Communications Technology, Toshiba Solutions, Cross Language, NHK, Sharp Corporation, NEC, Sakamoto Yoshiyuki

6. Bibliography

[1] http://aamt.info/

[2] http://tran.blog.shinobi.jp/Entry/316/

[3] Fuji, "Evaluation experiment to reading comprehension of an English-Japanese machine translation sentences," the collected papers of the 2nd annual meeting of Association for Natural Language Processing, 1996.

[4] http://www.lisa.org/

[5] http://www.lisa.org/standards/tbx/

[6]http://www.r.dl.itc.u-tokyo.ac.jp/~nakagawa/resource/termext/atr.html

[7] Koyama, Kageura, and Takeuchi, "The compound term extraction from a Japanese specialized-area text corpus," Information Processing Society of Japan, Natural Language Processing Study Group, 176-NL-2006, pp.55-50, 2006.

[8] Hino, Sasaki, Utsuro, Tsuchiya, Nakagawa, and Sato, "Translation inference of technical terms using the related term collection technique from the Web," The collected papers of 11th annual meeting of Association for Natural Language Processing, pp.21-24, 2005.

[9] http://yakushite.net/

AAMT (Asia-Pacific Association for Machine Translation) invites organizations and individuals who could collaborate with us to establish the specification of UTX to provide and/or build dictionaries, and to perform evaluation. At present, priority is given to the Japanese, English, and Chinese languages. If you are interested, please do not hesitate to contact us from the following page:

About UTX:
http://www.aamt.info/utx/english/

UTX mailing list:
http://groups.yahoo.co.jp/group/UTX
(Anyone can participate in this mailing list, but the correspondence is mostly in Japanese. We are planning to start another mailing list in English.)