UTX 1.11, a Simple and Open User Dictionary/Terminology Standard, and its Effectiveness with Multiple MT Systems

Seiji Okura
Fujitsu Laboratories Ltd.

Yuji Yamamoto
CosmosHouse

Hajime Ito
Inter Group Corporation

Michael Kato
Learning Consultant

Miwako Shimazu
Toshiba Solutions Corporation

Francis Bond
Nanyang Technological University

AAMT (Asia-Pacific Association for Machine Translation) Sharing/Standardization Group
http://aamt.info/english/utx/index.htm

Abstract

We have formulated a dictionary/glossary format UTX 1.11 and released it in May 2011. UTX 1.11 is a simple format that is friendly to both computers and humans. UTX dictionaries can be used not only as machine-readable dictionaries for rule-based machine translation (MT) systems, but also for computer-aided translation by human translators. The initial objective of UTX-Simple 1.00, released in 2008, was to improve the accuracy of various MT systems by specifying a common format. A key feature of its latest version UTX 1.11 is a term management mechanism by introducing four term statuses ("provisional," "forbidden," "approved," and "non-standard"). We show that a UTX 1.11-based dictionary originally created as a glossary is highly effective for improving the accuracy of MT. UTX can be widely and successfully applied in various fields with specialized terminology, such as localization, open source, education, administration, medicine, and law.

1. Previous Work

A number of terminological formats have been created in the past, such as TBX (ISO 30042, Term-Base eXchange), OLIF (Open Lexicon Interchange Format, Lieske et al., 2001)[1], and LMF (Lexical Markup Framework, Francopoulo et al., 2006). After the release of TBX-Basic[2], an even simpler format, TBX-Glossary[3], was proposed in 2010. However, these formats are primarily designed for exchanging existing glossaries with additional properties, which are not always used or useful. These are all XML-based formats, and their creators require detailed knowledge of XML, linguistics, and the specifications when they compile a glossary from scratch. These formats are designed for lexicographers to define large-scale terminology in enterprises and organizations. Though they have a wide variety of functions, these XML-based formats are very complex and require huge cost, time, and efforts to compile and manage. They are not designed for individual translators and therefore not suitable for effectively gathering translation knowledge and sharing/reusing them.

In 1995, AAMT (Asia-Pacific Association for Machine Translation) released a common format for machine translation user dictionaries, UPF (Universal PlatForm, Kamei et al., 1997), with the support of IPA (Information Technology Agency, Japan). AAMT is a non-profit organization to promote machine translation technologies. It is one of three regional associations of the IAMT (International Association of Machine Translation) along with AMTA (Association for Machine

[1] http://www.olif.net

[2] https://www.socialtext.net/data/workspaces/terminology-sig/attachments/tbx_basic:20081024215407-0-19440/original/TBX_Basic_datacategoriesV2.pdf

[3] http://www.ttt.org/tbxg

Translation in the Americas) and EAMT (European Association for Machine Translation). Its members include major Japanese manufacturers of packaged translation software for consumers. It should be noted that the standardization effort of AAMT has focused on creating a dictionary format for consumers, not for MT specialists. Several MT tools in Japan supported UPF, but its use was very limited. This is partly because its specification was based on SGML and was too complex for consumers to handle with. Also, MT users didn't realize the benefits of sharing language resources (user dictionary). Moreover, tools to use UPF were not sufficiently developed nor became commonly used. Therefore, it was difficult to share dictionaries by manually converting them for translation. There was a need for a simple, light-weight dictionary format that is easy to create and share.

2. A New Format, UTX, and its Features

User dictionaries are indispensable for effective use of MT systems. Although most of statistical MT systems cannot reflect their users' needs through user dictionaries, they ultimately need to rely on external terminological management systems to produce commercially usable translations. In contrast, rule-based MT systems integrate terminological needs in their translation process through user dictionaries. If the specification of user dictionaries for each MT system is different, it is not possible to share dictionaries across systems. Therefore, there have been efforts among AAMT members to establish a common format specification for dictionaries that can be incorporated by various MT systems. Following the changes of technologies and dictionary use after the release of UPF, AAMT began to explore a new dictionary format as a successor of UPF in 2006. The driving force was AAMT Working Group 3 (Sharing/ Standardization), also known as the UTX team. In 2007, the format was formally named as "UTX (Universal Terminology eXchange)." In 2008, the specification of UTX-Simple 1.00, a simpler version of UTX, was released (Bond et al., 2009). Subsequently, we successfully converted three dictionaries - in medicine, law, and computational linguistics - into the format of UTX-Simple 1.00 and made these available to the public.[4]

UTX-Simple was renamed to "UTX" in 2011, since UTX in its simple tabular format is the best suited for our purpose and the current situation. We realized that the initially planned "UTX in XML format" would not be very useful. In this paper, we use the formal name of each version, but UTX-Simple is equivalent to UTX unless otherwise stated.

There are four main characteristics of UTX.

1. **Simple**: UTX is easy to use and read. A complex specification needlessly increases the user's burden. UTX is a practical and understandable format from the user's standpoint. UTX is designed to be the greatest common divisor of various tools, not the least common multiple. In other words, UTX share only the essential information that is useful and usable to many tools.
2. **Open**: The UTX specification is an open standard. Everyone can freely compile a dictionary based on the specification and can freely use it. A UTX dictionary would include a clear license in its header to encourage sharing.
3. **Univocal in a specific domain**: The domain of a dictionary is decided from the viewpoint of "technical terms" as opposed to common, non-technical terms. The UTX format is designed to be a highly enriched technical dictionary for a specific domain. Its principle is "one term, one meaning." A term in a dictionary should be univocal.
4. **Versatile**: It is possible to promptly share and recycle UTX dictionaries with existing text editors and spreadsheet applications. The format enables us to compile dictionaries quickly and easily, which is indispensable for improving translation accuracy.

UTX places emphasis on easy editing in spreadsheet format, at the expense of covering the full range of possible translations. The goal is to avoid complexity and a wide choice of infrequently used functionalities. If long-term terminological management is required, it would be better to use XML-based formats such as TBX, but not UTX. UTX is useful when used for the preparation of such formats.

[4] http://aamt.info/english/utx/index.htm#Download

It is also noted that bidirectionality is supported in UTX 1.11. Moreover, the specification can be also used for a monolingual dictionary for applications such as authoring tools. By sharing and recycling common dictionaries by introducing UTX, efficiency of translation works can improve further.

3. Problems of UTX-Simple 1.00

In principle, UTX-Simple 1.00 was designed for unidirectional translation use. For example, an English-Japanese dictionary was not readily usable for Japanese-English translation. Even if the dictionary contains useful entries, there was no guarantee that they would work in the reversed translation direction.

Moreover, there was a problem from the viewpoint of term management when applying it to computer-aided translation. For example, there was no means to specify that two or more terms (when only one of them is a formal term) apply to the same concept. Also, when several terms have been presented by two or more translators, there was no means to specify which were provisional and which have been approved. In commercial translation, there are forbidden terms for many reasons - a political factor, a social factor, an image of the brand of the enterprise, etc. If such terms cannot be appropriately managed, a glossary's value may be substantially reduced.

The UTX specification has been enhanced in order to address and overcome these problems of term management.

4. Specification of UTX 1.11

The most important improvement from UTX-Simple 1.00 was to introduce "term statuses", enhancing the practicality in translation aid through term management.

Four term statuses, "provisional," "forbidden," "approved," and "non-standard" were introduced in UTX 1.11. Moreover, a notion of a dictionary administrator and a dictionary contributor were introduced from the viewpoint of dictionary management. A dictionary administrator is ultimately in charge of a dictionary, and defines the framework of the dictionary. A dictionary contributor adds new terms to the dictionary. After one or more dictionary contributors add terms, the dictionary administrator judges whether they are appropriate, and decides their term statuses. If a dictionary is created by a single individual, they are both dictionary contributor and dictionary administrator.

Provisional: The term status "provisional" means that an entry is not yet authorized by the dictionary administrator. It is preferable that the dictionary administrator changes the status to one of "forbidden," "approved," or "non-standard," or deletes the term.

Forbidden: The term status "forbidden" means that an entry includes a target term which should not be used from the viewpoint of term management. A target term may also need to be suppressed to avoid conflict with different domain-specific dictionaries, when a translation tool does not properly honor the priorities among multiple dictionaries. Forbidden terms can be extracted to be used for terminological check outside of a translation tool.

Approved: The term status "approved" means that an entry has been approved by the dictionary administrator and a translator must use the term. There is only one entry where the term status is "approved" for the term in the source language. An approved term is always bidirectional, that is, usable for translation from Language A to Language B and vice versa. It is the only effective entry when there are two or more entries corresponding to the same concept ID and the direction of translation is reversed.

Non-standard: The term status "non-standard" indicates one or more non-standard source terms. Non-standard terms are only permitted to accommodate variations of source terms. Non-standard terms should not be used as target terms. The only reason to register a term as non-standard is to enable automatic translation to translate improper terms in the source language.

Concept ID and dictionary ID are defined to manage two or more terms in the same concept. Concept ID is an optional numerical value up to ten digits to specify the same concept to two or more entries. Dictionary ID is optional four alphanumeric characters (case insensitive) to distinguish entries with the same concept ID when multiple dictionaries are merged. The dictionary

#UTX 1.11; en-US/ja-JP; 2011-04-19T19:00:00Z+09:00; copyright: AAMT (2011); license: CC-BY 3.0 #description: This is a sample dictionary for AAMT-related terminology. It is not an official dictionary.				
#src	tgt	src:pos	term status	concept ID
dictionary administrator	辞書管理者	noun	approved	
provisional word	暫定語	noun	non-standard	1
provisional term	暫定語	noun	approved	1

Table 1. Example of UTX dictionary

administrator defines the dictionary ID. Table 1 is an example of the UTX1.11 format including term statuses and concept IDs.

5. Conversion among User Dictionaries, Glossaries, and UTX

In this section, we explain our converter and conversion examples.

5.1 UTX Converter

To use a UTX dictionary for specific translation tool, UTX may need to be converted to a dictionary format of such tool. A converter is a tool that converts the UTX format to/from other text-based formats for various applications. With a converter, a single UTX dictionary can be used for various applications. The UTX dictionary will be distributed widely if there is an online community where the dictionary can be freely uploaded and downloaded. In 2009, we developed and evaluated a converter which converts the UTX format into the user dictionary formats for five Japanese machine translation systems. The following year, Alan Melby developed a converter for glossary formats, including UTX-Simple and TBX-Glossary.[5]

With these converters, UTX can be currently converted into the following formats: ATLAS, The HON-YAKU, Yakushite-Net, LogoVista, SYSTRAN, and TBX-Glossary.

5.2 Conversion Examples

Examples of conversion from UTX into other formats are shown in Table 2 and Table 3.

[5] http://www.ttt.org/tbxg/

名詞－名詞	辞書管理者	dictionary administrator	0	3
名詞－名詞	ユーザー辞書共通フォーマット	common format for user dictionary	0	3
名詞－名詞	暫定語	provisional word	0	3
名詞－名詞	暫定語	provisional term	0	3

Table 2. Example of converting UTX dictionary into user's dictionary of a translation system

dictionary administrator;n;(種類 n);	
	辞書管理者();
common format for user dictionary;n;(種類 n);	
	ユーザー辞書共通フォーマット();
provisional word;n;(種類 n);	
	暫定語();
provisional term;n;(種類 n);	
	暫定語();

Table 3. Example of converting UTX dictionary into user's dictionary of a translation system

5.3 Compatibility of "term status" for Various Applications

Many translation tools have equivalents of the term status in UTX 1.11, but the implementations differ from tool to tool.

Term status "approved": Refers to a term that has been approved by the dictionary administrator and which should be used by an MT system whenever possible. While the equivalent implementation of approved terms is different between applications, the common framework is "a term used for translation with top priority." Its priority is higher than terms in the system dictionary.

Term status "forbidden": Only one translation system currently has a function that corresponds to the term status "forbidden." Another system has the notion "DNT" (Do Not Translate). It differs from the concept of "forbidden" in UTX. DNT is for terms that should not be translated at all, such as proper nouns.

Term status "provisional": There is no corresponding notion for any translation systems. However, it can be said that it is supported by almost all systems because creating a user dictionary that contains only provisional terms and registering terms to it enables the system to translate those terms properly.

Term status "non-standard": There is no corresponding notion for any translation system. However, it can be said that it is supported by almost all systems because registering several terms in the target language corresponding to a term in the source language enables the system to translate those terms properly.

Conversion from each user dictionary to UTX differs substantially among systems.

Any system can assign "approval" and "provisional" status to the terms, as long as these terms are grouped into separate dictionaries.

For the systems that do not distinguish the term statuses "approved," "provisional," and "non-standard," we can reproduce support through the following process: A) Create distinct user dictionaries for "approved terms," "provisional terms," and "non-standard terms"; B) Convert these into UTX, place the corresponding term status on the terms in each dictionary; C) Merge these dictionaries in UTX by manually setting concept IDs.

To better address circumstances such as noted above, we intend to provide enhancements to the converter and publish a user's guide for the converter tool.

6. Evaluation

In 2009, we developed a converter from UTX to the user dictionary formats of five Japanese MT systems, and conducted an evaluation. We converted a dictionary made for one MT system by means of UTX-Simple 0.9, and reconverted it for other MT systems. We found that the accuracy of translation improved for about 37% of sentences (Bond et al., 2009).

Our previous evaluation focused on the effectiveness of an existing glossary when fine-tuned as a user dictionary for MT systems. For this evaluation, the base glossary is more closely related to the target document.

In our test, two new evaluations were conducted. First, we converted a UTX-format dictionary into dictionary formats supported by MT systems and checked if each system can properly incorporate term status properties. Next, we measured the improvement of translation accuracy by translating a document with the converted dictionary for each system. For the evaluation of translation accuracy, we used the "UTX 1.11 Specification"[6] (3961 words, 314 sentences). It was originally written in English and has been translated into Japanese by a human translator.

6.1 Systems Used in the Evaluation

The systems we used in our evaluation are as follows.

- **LogoVista PRO 2008 Super Pack** (www.logovista.co.jp/product /honyaku_pro2008/pro2008_st.html)
- **Translation Software ATLAS** (www.fujitsu.com/global/services/software/ translation/atlas/lineup/)
- **The HON-YAKU 2009 Premium** (pf.toshiba-sol.co.jp/prod/ hon_yaku/premium/index_j.htm)

[6] http://aamt.info/english/utx/utx1.11-specification-e.pdf

- **SYSTRAN**
  (www.systransoft.com/translation-products/desktop/systran-7-premium-translator)

In the evaluation results, the systems are anonymized as A, B, C, and D.

6.2 Dictionary Used in the Evaluation

The UTX dictionary used in the evaluation is the "(unofficial) AAMT glossary".[7] It was created when a human translator translated the UTX 1.11 specifications from English to Japanese. This glossary was not specifically designed for MT use. Therefore, terms that improve the accuracy of MT were not specifically registered to the UTX dictionary. A total of 51 entries are found in the UTX dictionary (including entries whose term status is provisional, non-standard, and forbidden).

7. Result: Verification of Effectiveness of UTX

The specification of UTX 1.11 was translated from English to Japanese using each MT system, with and without the converted UTX dictionary.

7.1 Verification of Conversion into the Format of User Dictionaries for Different MT

Here we will mention the term status "forbidden" and noun/verb inflections.

For the term status "forbidden," System A can specify those headwords that should not be used in user dictionaries, but not their translations which UTX can. However, other systems do not have such a notion. Even if forbidden terms are included in the dictionary (and indicated as such), they may not be excluded from the top priority of the translation process. Seven terms of the UTX dictionary - whose term status were "forbidden" - were removed for Systems B, C, and D. As a result, the number of entries for those systems totals 44.

Using System B, unless its native user dictionary tool is used, the plural forms of nouns and inflection of verbs cannot automatically be set. It is preferable that the UTX converter has a function to automatically compute inflection forms of terms.

[7] http://goo.gl/DjLy5

Although there are some properties to be set manually, it was generally possible to automatically convert the UTX dictionary with our converter into the formats of user dictionaries for each system. In general, the term statuses introduced from UTX1.1 can be reflected properly in user dictionary of MT.

7.2 Verification of Effectiveness of the UTX Dictionary in English-Japanese Translation

Table 4 is a result of evaluating the translation result of four systems (A, B, C, and D) using BLEU (Papineni et al., 2002). Figure 1 is the graphical version of Table 4. BLEU scores have improved in all systems by using the converted UTX dictionary. The range of improvement is 1.61-2.67, and the average of improvement is +2.14. Despite the limitations that the UTX dictionary used in this evaluation was created as a human translation glossary and there were only 51 terms in it, it demonstrated more than 2 point improvement by BLEU. Further, the effectiveness of the dictionary was verified across all four MT systems.

System	A	B	C	D	Ave.
(A) default	18.31	16.57	16.00	13.06	14.53
(B)=(A)+UTX	20.63	18.40	18.67	14.67	16.67
Diff.=(B)-(A)	+2.32	+1.83	+2.67	+1.61	+2.14

Table 4. Evaluation result (BLEU score)

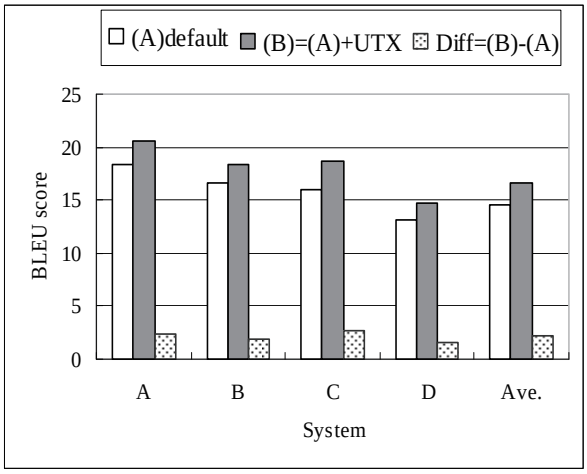

Figure 1. Evaluation result (BLEU score)

This improvement of BLEU scores is mainly due to accuracy of translation of compound words. It is known that registration of compound nouns to MT dictionaries improves MT accuracy (Fuji, 1996).

We should note that not all terms for improving MT accuracy have been registered in MT dictionaries, such as "English-Japanese dictionary," "domain," "computer," and "form" (verb), which have a couple of possible Japanese translations. Accuracy can be further improved by registering these terms. Using some functions such as machine learning or adding more entries to the user dictionary may be also effective.

However, in one system, there was a side effect of an approved entry. An approved entry "term" caused the idiomatic expression "in terms of UTX" to be translated incorrectly. This is because the specialist translation of "terms," 用語 *yougo* was used instead of the idiomatic translation. One could justify this behavior, because in principle, an approved term should always be used. In most of the systems, the multiword expression *in terms of* took precedent over the single word user dictionary entry *term*, which is the desired behavior. Further discussion will be necessary if this principle should always override idiomatic expressions.

8. Conclusion

In conclusion, we have proved that a UTX with only 51 entries significantly improved accuracy of four different rule-based MT systems. This means that a simple, non-XML-based dictionary with a proper domain and appropriate, hand-picked entries is sufficient to improve the accuracy of MT, without huge corpus and massive computing resources. Also, the results show that UTX 1.11 (with term statuses) can be successfully converted into other formats to be used for translation between Japanese and English.

In the previous versions, UTX didn't have an effective way to distinguish which entry is appropriate. Many entries did not match to the content of the translated document at all, and there was no way to improve the appropriateness of entries. In UTX 1.11, the introduction of term status contributes to an increase of accuracy, by clearly marking useful and less useful entries.

9. Future Work

Although we have developed a converter that enables the use of UTX-formatted glossaries as user dictionaries in various MT systems, we have not yet developed a converter for the opposite direction. We plan to address this need in the future, so that user dictionaries can be more easily shared across many translation applications.

In addition to UTX dictionaries for medicine, law, and computational linguistics that we have created and made available to the public, we plan to release dictionaries for other domains. There are problems of motivation or incentives when dictionary entries are added in online communities. An effective motivation or framework of incentives is an indispensable point of research to enhance the contents of dictionaries in the future.

Currently, we are working on the next version of UTX. A main enhancement will be multilingual support. Initially, UTX supported only two languages and one translation direction (from the source language to the target language) in a dictionary. The next version will support three or more languages in a dictionary, and multiple translation directions. While UTX can be unlimitedly extended by adding additional fields, keeping the format simple is the key for its usability and versatility.

It is also necessary to verify whether term status functionality is appropriate and effective, and to improve the method of managing glossaries. Through applying UTX to actual translation projects and receiving feedbacks, we intend to improve UTX further for the benefits of real-world translators.

UTX is still a young standard. Its openness and carefully designed simplicity successfully eliminate the problem of useful data locked in a proprietary format. It is already practical and useful at this stage; however, it will need more real-world feedback from a wider range of users.

References

Seiji Okura, Yuji Yamamoto, Toshiki Murata, Kiyotaka Uchimoto, Michael Kato, Miwako Shimazu and Tsugiyoshi Suzuki. 2009. Sharing User Dictionaries across Multiple Systems with UTX-S. *Proceeding of the Second International Workshop on Intercultural Collaboration (IWIC2009)*, Stanford. 147-154.

Francopoulo, Gil, George M., Calzolari N., Monachini M., Bel N., Pet M., Soria C. 2006. Lexical Markup Framework (LMF). *Proceedings of the 5th International Conference on Language Resources and Evaluation* (LREC 2006), Genoa. 1-8.

Fuji, Masaru. 1996. Experiments to evaluate the reading comprehension of English-to-Japanese machine-translated texts. *2nd Annual Meetings of the Association for Natural Language Processing, Proceedings*, Tokyo, 21-24.

ISO 30042:2008. *Term-Base eXchange (TBX) format specification.*

Kamei, S, S., Itoh, E., Fuji, M., Hirai, T., Ssaitoh,
Y., Takahashi, M., Hiyama, T., AND Muraki, K.. 1997. Sharable formats and their supporting environments for exchanging user dictionaries among different MT systems as a part of AAMT activities. *MT Summit VI. Machine Translation: Past, Present, Future. Proceedings*, San Diego, 132-141.

Lieske, Christian, McCormick, S., Thurmair, G. 2001. The open lexicon interchange format (OLIF) comes of age. *MT Summit VIII, Machine Translation in the Information Age, Proceedings*, Santiago de Compostela, 211-216.

Papineni, Kishore, Rouko, S., Ward, T., and Zhu, W. J. .2002. A method for automatic evaluation of machine translation. *40th Annual Meeting of the Association for Computational Linguistics, Proceedings*, Philadelphia, 146-156.

Wright, Sue Ellen, Nathan Rasmussen, Alan K. Melby and Kara Warburton. 2010. TBX Glossary: A Crosswalk between Termbase and Lexbase Formats, presented at the *Workshop of the Ninth Conference of the Association for Machine Translation in the Americas (AMTA2010)*, Denver.