

共有ユーザー辞書仕様 UTX の現状と今後の展開

大倉清司[†]、山本ゆうじ^{††}、村田稔樹[‡]、内元清貴^{‡‡}、
加藤マイケル孝仁[‡]、島津美和子[‡]、鈴木次良^{‡‡}

アジア太平洋機械翻訳協会 機械翻訳課題調査委員会 共有化・標準化ワーキンググループ

[†]富士通研究所 ^{††}秋桜舎 [‡]OKI ^{‡‡}ラーニングコンサルタント

^{‡‡}情報通信研究機構 東芝ソリューション ^{‡‡‡}クロスランゲージ

1. はじめに

機械翻訳システムを実用的に使用するには、ユーザー辞書が必須だが、辞書の仕様が異なると相互利用できない。そのため、AAMT（アジア太平洋機械翻訳協会）[1]はどの機械翻訳システムでも共通に利用できる、共有辞書の仕様を策定している。より具体的には、1995年にIPAの支援を受けて策定されたUPFをベースに、その後の技術や利用方法の変化を反映し、2006年から新しい仕様策定を開始した。2007年には「UTX (Universal Terminology eXchange)」という新名称に改め、現在はそのシンプルな形式であるUTX-Simpleの仕様策定を行っている。2008年度は、実際の翻訳分野を選定し、その分野の辞書データの作成・収集を行い、翻訳プロジェクトと連携しつつ、UTX仕様の修正・確定を行う予定である。また、辞書を継続的に作成・共有・蓄積するためのコミュニティの創設も構想している。本稿では、UTX-Simpleの仕様を紹介するとともに、我々が想定している辞書の共有・再利用の過程と方式について述べる。

2. UTXについて

翻訳ソフトなどの機械翻訳システムを実用的に使用するには、ユーザー辞書の利用が重要である。翻訳対象文書で使われている専門用語や人名・地名などは、基本辞書に登録されていないことが多い。これらは専門用語辞書を使っても適切に翻訳できないことがある。そこでこのような用語をユーザー辞書に登録すれば、機械翻訳システムの精度向上を見込むことができる。例えば、以下の原文をある翻訳ソフトで翻訳したとする。

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

結果は次のようになる。

「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。」

この訳文では、公表、符号化、属国、インフォメーションなど、原文の文脈と分野に合致しない訳語が使われている。しかし、適切な用語を辞書登録すれば、以下のように、より適切な訳文が生成される[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。」

特に「XML declaration」のような複合語をユーザー辞書に登録することにより構文解析の失敗も防げるようになることが知られており[3]、統計翻訳の精度向上にもつながるため、辞書登録は重要である。

しかし、個人が大量の辞書登録を行うのは時間がかかり、辞書登録による効果もわかりにくい。また現状では個人の辞書を集めてもフォーマットが不統一であり、共通化・共有化できない。

翻訳辞書の標準化された仕様としては、LISA[4]のTBX[5]が挙げられるが、仕様が複雑で広く普及していないのが現状である。これらの問題を解決するために、AAMTではUTXを策定中である。

UTXの特徴は以下の通りである。

1. **分かりやすく使いやすい「ユーザーのための」辞書**：複雑な仕様はむだにユーザーの負担を増やし、結局は使われない。ユーザーの立場に立ち、実際に使われるように単純かつ実際の仕様が心がける。
2. **「専門用語」という観点**：辞書の分野を明確化し、「一語一義」とする。無意味に訳語を増やさず、使い分けを厳密に規定する。該当分野で訳語が一義的に定まる語を登録対象とする。
3. **翻訳精度の向上**：UTX辞書が普及すれば共有も簡単にできるため、翻訳精度向上に欠かせない辞書の構築がより効率的にできるようになる。
4. **多言語辞書、単言語辞書としても有効**：二言語間だけでなく、多言語にも対応する。また単言

語辞書として、用語の統一など校正支援ツールなどの入力としても使えるように仕様を策定している。

5. **共有のための情報も保持**：XML 形式においては辞書作成者、エントリー作成日時や個々の機械翻訳固有の情報など、細かい情報も保存できるため、UTX 辞書は効果的に共有できる。
6. **ソフトウェアのローカライゼーションの促進**：特にオープンソース系のローカライゼーションにおいては、翻訳を個人単位で行っているために訳語の統一が難しい。UTX の導入により、共通の辞書が普及して翻訳作業の効率化が期待できる。

海外の例では、欧州委員会（EC）が 2008 年 1 月に、約 100 万の文章を 22 言語に翻訳した翻訳データを無料で公開すると発表している。日本では、依然として厚い言語の壁に阻まれて、ソフトウェアやコンテンツの翻訳がなかなか進まず、多くの潜在的機会が埋もれ、日々莫大な損失が発生している。翻訳資産の共有化がすべての業界で推進されることが、今後の産業振興の鍵となる。

UTX のメリットとして、ユーザーにとっては、分かりやすく作りやすいフォーマットであること、個々の特定分野で翻訳精度の向上が期待できること、インターネット、LAN 上のコミュニティを通じた辞書の共有と再利用ができることなどがある。翻訳ソフト開発メーカーにとっても、ユーザー辞書利用の促進により機械翻訳が活性化し新規需要の掘り起こしが期待できること、またメーカー固有の辞書項目も完全に保持することなどが挙げられる。

UTX の仕様策定にあたって、以下の 2 つの作業が重要である。

1. 仕様の策定（UTX-Simple および UTX-XML）

- 1 つの「ヘッダ」と 1 つ以上の「エントリー」からなる
- 文書の最初の行に「ヘッダ」を記述する
- 2 行目以降、1 行に 1 エントリー記述
- # で始まる行はコメント行

【ヘッダ】半角文字でそのファイルに関する記述を入れる。区切り文字は半角スペース。

#UTX-S バージョン番号 原言語/目的言語 最終更新日付時間 4 カラム目以降の情報

例) #UTX-S 0.9 en-US/ja-JP 2007-12-03T14:28:00Z+09:00

UTX-S の S は簡易版の意味。言語名 (ISO 639) と日付時間 (ISO 8601) は ISO 準拠 [2, 3]。

単言語辞書の場合、目的言語を省略する。

4 カラム目のフォーマット例：

source: plural/3sp/past/pastp/presp/comparative/superlative/target: .../optionXXXX/optionYYYY

(見出し語の情報は、source: の後に記述。plural=複数形、3sp=三人称単数、past=過去形、pastp=過去分詞、presp=現在分詞、comparative=比較級、superlative=最上級、訳語の情報は target: の後に記述)

【エントリー】1 行のフォーマットは、タブ区切りでカラムに分ける。各カラムに記述する内容は以下の通り：

1カラム目	2カラム目	3カラム目	4カラム目以降
原言語の単語	目的言語の単語	品詞	その他の属性 (任意)

図 1 UTX-Simple の仕様

2. 実際の辞書データの作成と、辞書を継続的に作成・共有・蓄積するためのコミュニティの創設

本稿では、上に挙げた 2 つの作業内容について説明するとともに、今後の展望・課題についても述べる。

3. UTX-Simple の仕様

UTX では、以下の 2 種の仕様を定める。

- UTX-XML (XML 形式)
- UTX-Simple (タブ区切りテキスト形式)

UTX-XML (XML 形式) は、辞書に必要な情報を全て含む。UTX-Simple (タブ区切りテキスト形式) は、最低限、原語、訳語、品詞の 3 種の情報だけで使用可能な仕様を規定する。仕様策定に関しては、最終的には XML 形式を策定するが、普及という観点から、シンプルな形で共有できる UTX-Simple をまず策定している。

標準化し、さらに普及させるためには仕様をユーザーに分かりやすく使いやすい形式にする必要がある。また、機械翻訳システムでユーザー辞書をフルに活用するには、現状のシステムよりも簡単に登録できなくてはならない。UTX-Simple 形式の辞書は仕様が単純で作成の手間が減り、辞書の共有が促進される。なるべく記述する情報を少なく、しかし実用的になるように仕様を検討した。現在のバージョンは 0.9 である。基本的な仕様は図 1 の通りである。

UTX-Simple の仕様策定にあたり、以下のことが論点として上った。

- (1) 人間にとっての読みやすさを優先するか、機械にとっての処理のしやすさを優先するか。

何を必須の項目にするかについて議論した。最小限原語、訳語、品詞情報があればよいことにした。

例えば英語では名詞の複数形の情報が必要な場合がある。しかし複数形は全ての言語の全ての品詞で必要なわけではない。確かに複数形の情報がないと機械翻訳がうまく訳せない状況も少なくないが、人間が記述すること、また普及の観点から、オプションとして記述することにした。

(2) XML版(UTX-XML)と簡易版(UTX-Simple)の位置づけをどうするか。

UTX-Simpleは日付、作成者などの情報がなく、恒久的・継続的な辞書の構築には適さない。しかし、一方で使いやすさから、容易に作成、共有することができ、普及が期待できる。XML版の策定にはまだ時間を要する。

(3) カラムの内容を固定にするか(Xカラム目=Y、と決めてしまうか)。

カラムは、言語共通の情報と、言語固有の情報を入れる枠組みとした。共通の情報は1~3カラム目に、実際翻訳において必要とされている言語依存の固有情報は4カラム目以降に記述する。

(4) 言語固有のカラムを言語対ごとにあらかじめ決めておくか否か。

当面、UTXでは日本語、英語、中国語を中心に扱うが、本来はISOに従い存在する言語全てを扱えるようにしたい。多数の言語について必要なカラムを現時点ですべて策定するのは現実的ではない。そこで、ユーザーにカラム名を指定できる幅を持たせる。実際の用法で最低限必要な情報を見極めていく。言語対によっては、何カラム目は決まった情報を入れることを推奨する。

(5) ガイドラインの作成について

個々のエントリーについてどう記述すればよいか、表記などのガイドラインを決める必要がある。UTX-Simpleを読み込むツールはそのガイドラインに従い、警告を出すようにする。例えばツールの警告出力オプションを指定しておけば、必要な項目がなければ警告を出す。基本的にフォーマットエラーはなしで、フォーマットが違っていた場合、その行は無視することにする。

基本的なガイドラインとしては以下を取り決めた。

- デフォルト値と「該当なし」との区別
「該当なし」を示す値：-(ハイフン)
それ以外は、自動処理(ツールの判断)
- 辞書に保持されている情報をツールがどこまで使うかはツール自体の仕様に依存する。
- ユーザーが定義したカラムは保持するのみ。

- 英語の名詞句は語頭を小文字で書く。複数形のみの場合、見出し語を-(ハイフン)にする。

UTX-Simpleの具体例を図2に示す。

4. 辞書を継続的に作成・共有・蓄積するためのコミュニティの創設

辞書データの作成と収集の基本方針として、漫然とした分野の辞書は作らず、スポーツ分野、IT分野など、辞書ごとに特定の使用分野を限定する。

現状では個人単位の翻訳において、特にオープンソースの翻訳に関して辞書が共有できないことが問題となっている。辞書はさまざまな提供元に拡散しており、それぞれにフォーマットや使用ライセンスが異なる。言語資源を集約できれば、各国語間のローカライゼーションを一挙に加速できる。

UTXは仕様がオープンであり、UTXの辞書を真に普及させるために誰でも参加できる「みんなで作る辞書」共有辞書コミュニティを構築する。そのためには以下の取り組みが必要である。

- 辞書を作成・共有・蓄積するための辞書コミュニティと流通インフラの確立
- 公式辞書コミュニティでは、品質を保証した辞書を有償提供(AAMTまたは関連組織が管理する)
- オープン辞書コミュニティでは、オープンソース的な許諾による自由かつ無償で相互利用(AAMTまたは関連組織はホスティングのみを行う)

多様な需要に対応するため、公式辞書と無料辞書を作り、それぞれを管理・共有するためのコミュニティを作る。公式辞書は有償だが、専門家が監修して信頼できる内容を保証する。無料辞書は、内容の正確さが保証されないが無料で利用できる。

すでに共有化・標準化ワーキンググループでは、Mozilla 24などのイベントに参加してUTXについて紹介するなど、オープンソースのコミュニティとも意見交換を行っている。

UTXのフォーマット自体は単純だが、分散した言語資源を集約し、「集合知」として活用できる。そのために参加者にとって辞書作成の継続的な動機付けを維持し、参加者が不公平にならない枠組みの作成が必要になる。

#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						

図2 UTX-Simpleの基本記述例

5. 今後の展望と課題

仕様や指針の策定と同時に、辞書データの作成および収集、各種ツールの開発と活用、UTX の実効性の評価と確認、そして各方面との協力・連携を進めていく。最終的には UTX の ISO 標準化を目指す。

5.1 ライセンス体系の策定

有償・無償のそれぞれの辞書形態について、辞書の配布を単純かつ明快で単一のものとする。特に無料・オープン辞書では相互利用可能とし、改変と商業利用が許諾される必要がある。著作権の問題が解決できることから作成していく。

5.2 UTX-XML の仕様策定

UTX-Simple に基づき、より高機能な UTX-XML の仕様の詳細を決定していく。

5.3 表記指針の策定

辞書で音引き、中黒などの表記が統一されていないと、異なる辞書を統合するたびに修正が必要になる。翻訳の現場では、各社の表記の違いを吸収するだけで多大な労力が日々浪費されている。そのため、辞書表記の統一指針を定めて遵守する必要がある。

5.4 辞書データの作成・収集

実際に翻訳が必要とされている分野をいくつか選定し、UTX-Simple、UTX-XML の仕様にそって具体的な辞書データ作成・収集を行う。その辞書を使った翻訳を行うことにより、UTX の仕様の修正・確定を行っていく。

5.5 各種ツールの開発と活用

以下のツール開発が必要になる。

UTX 変換ツール（正規化ツールを含む）

翻訳ソフトや翻訳サイト独自の形式と、UTX 形式を相互に変換するツール。UTX 形式の仕様が正しく実装されているか検証する正規化ツールを含む。また、UTX-XML と UTX-Simple のコンバーターも必要である。

用語抽出・辞書作成ツール

一語一語登録するのではなく、原文を解析して一括して必要な辞書登録を行うためのツール[6,7,8]。

辞書検索ツール（用語集検索ツール）

辞書や用語集を直接検索して参照できるツール。

5.6 UTX の実効性の評価と確認

UTX の普及のためには、UTX を使って翻訳精度が向上し、翻訳効率が改善することを実証する必要がある。評価に関しては、AAMT でもテストセットを作成し、UTX 辞書の効果を確認する予定である。

5.7 各方面との協力・連携

共同執筆者が運営する、ユーザが専門用語を登録できるコミュニティ型機械翻訳サイト「訳してねっと」[9]との連携をはじめ、共同執筆者以外にも、UTX にはすでに各方面からの協力が得られている。今後は、Edict、多数の辞書やシステムをつなぐ他のプロジェクト（言語グリッド）、言語資源を収集・管理・配布する組織（GSK）などとの協力・連携も考えている。実際に UTX を使ってもらい検証していくとともに、ツール開発面でも協力・連携していきたい。

UTX の趣旨に賛同している組織および個人（順不同、敬称略）

富士通研究所、秋桜舎、OKI、情報通信研究機構、東芝ソリューション、クロスランゲージ、NHK、シャープ株式会社、NEC、坂本義行

参考文献

- [1] <http://aamt.info/>
- [2] <http://tran.blog.shinobi.jp/Entry/316/>
- [3] 富士秀. 英日機械翻訳文の読解に関する評価実験. 言語処理学会第 2 回年次大会論文集, 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] 小山, 影浦, 竹内. 日本語専門分野テキストコーパスからの複合語用語抽出, 情報処理学会, 自然言語処理研究会, 176-NL-2006, pp.55-50,2006.
- [8] 日野, 佐々木, 宇津呂, 土屋, 中川, 佐藤. ウェブからの関連語収集手法を用いた専門用語の訳語推定. 言語処理学会第 11 回年次大会論文集, 言語処理学会, pp.21-24, 2005.
- [9] <http://yakushite.net/>

問い合わせ先

AAMT（アジア太平洋機械翻訳協会）では、UTX の仕様策定や辞書作成、評価にご協力いただける方を募集しております。現時点では、日本語、英語、中国語を優先しています。興味のある方は下記ページからお気軽にご連絡ください。

UTX についての説明 URL:

<http://www.aamt.info/utx/japanese/>

UTX メーリングリストについて:

<http://groups.yahoo.co.jp/group/UTX/>

（誰でも参加できます）