

機械翻訳ユーザ辞書の共通フォーマットの設定 ----- アジア太平洋機械翻訳協会における活動中間報告 ----- Common Formats for User Dictionaries of MT Systems

伊藤 悦雄^{*1} 村木 一至^{*2} 桧山 努^{*3} 赤羽 美樹子^{*4}
斎藤 由香梨^{*5} 平井 徳行^{*6} 亀井 真一郎^{*2}

Etsuo ITOH, Kazunori MURAKI, Tsutomu HIYAMA, Mikiko AKABANE,
Yukari SAITOH, tokuyuki HIRAI, Shin-ichiro KAMEI

*1 東芝, *2 NEC, *3 NEC 情報システムズ, *4 ノヴァ, *5 富士通研, *6 シャープ

etsuo@sp.tokyo-sc.toshiba.co.jp, {hiyama,muraki}@hum.cl.nec.co.jp,
aka@nova.co.jp, yukari@ling.flab.fujitsu.co.jp,
nnd6@isl.nara.sharp.co.jp, kamei@hum.cl.nec.co.jp

1. はじめに

翻訳に使用される辞書データの流通・相互利用を促進するため、アジア太平洋機械翻訳協会(AAMT)に加盟する機械翻訳(MT)メーカーが中心となり、各社のMTシステムのユーザ辞書と相互変換可能なユーザ辞書共通フォーマット(Universal PlatForm; UPF)の設計を開始した。本稿では活動の目的と方針概要を述べる。

商用のMTシステムが製品化されて10年余が経過した。当初、MTシステムは翻訳の専門家向けに設計・開発され、主に技術翻訳に用いられてきたが、近年のインターネットの発達とパソコンの普及により、MTシステムは一般ユーザにも急速に浸透し始めている。MTシステムを有効活用するためには、各ユーザ毎に、頻繁に使用する語彙を「ユーザ辞書」として蓄積し、システムの基本辞書と合わせて使う必要がある。しかし辞書作成は一般に時間と労力がかかる仕事であり、個人ユーザー一人一人が辞書を個々に充実させるのには限界がある。

この問題の具体的な解決方法として、個人が個別に蓄えている辞書データを流通させ、相互利用するための環境の整備が挙げられる。現在は国内の20数社が機械翻訳システムを商品化しているが、それらの機種の違いを越えて、共通にユーザ辞書を交換できる仕組みがあれば、各

人がユーザ辞書を作成するコストが大幅に削減できる。このことによりMTの利用が促進され、ひいては日本人の外国語文書受発信が促進される。

このような環境整備の具体的な活動として、AAMTでは、今年度(平成8年度)から来年度にかけ情報処理振興事業協会(IPA)の創造的ソフトウェア育成事業の予算補助を受け、各社のMTシステムに共通のユーザ辞書記述フォーマットUPFの開発と、ホームページによる仕様公開の活動を開始した。仕様は検討段階においても適宜一般に公開する方針である。

2. UPF設計の基本方針

異種システム間で辞書データを交換できるようにするため、以下のような環境を開発する方針である。

- (1) 共通フォーマットの設計
- (2) 共通フォーマットで記述された辞書を蓄え流通させるための、一般アクセス可能な電子環境の提供

上記(1)のフォーマット開発は、現実に発売・利用されている複数のシステム間でそのユーザ辞書情報を比較することにより行なうこととした。この方針をとることで、共通フォー

マツトが現実のシステムから遊離してしまう危険を回避できると考えている。またこの仕様に基き作成されたデータを参加各社のMTシステムにおいて実証評価を行なう。このために、共通フォーマットと各システムのユーザ辞書の間の双方向コンバータを各MTメーカーがそれぞれ独自に開発することになる。共通フォーマット設計の際には、共通フォーマットから各システムのユーザ辞書への変換（ダウンロード）と、各システムのユーザ辞書から共通フォーマットへの変換（アップロード）との両方が可能となるよう考慮する必要がある。

上記（2）の辞書共有環境としては、AAMTのホームページを想定している。また直接UPF形式で辞書を記述するための辞書エディタも提供する。各ユーザは共有環境に置かれているUPF形式の辞書データを自分の使っているMTシステムのユーザ辞書のフォーマットに変換して使用することができる。また逆に各ユーザが自分の使用しているMTシステムで作成したユーザ辞書は、UPF形式に変換してこの辞書共有環境に置き、他ユーザ（異システムユーザも含む）と共有できる。

UPFは、さしあたり日本語と英語の2カ国語を分析対象として具体的設計をすすめるが、その形式は多言語に対応できるものを目指している。また一つの言語の生成と解析の辞書はできる限り統一した形式で記述できるように仕様設計することを目標としている。具体的記述形式はSGMLのようなマークアップ言語で記述する。

3. UPF設計の具体的活動

3. 1. 基本変換標準と拡張変換標準

上述したようにUPFは、UPFから各システムのユーザ辞書への変換（ダウンロード）と各システムのユーザ辞書からUPFへの変換（アップロード）との双方向が可能となるように設計する必要がある。現実利用されている複数のシステムのユーザ辞書で扱える語彙（品詞）には相違があるから、狭い意味で上記の双

方向条件を満たすためには、各システムに共通して記述できる語彙（記述可能な語彙の「AND」）だけを対象範囲とする必要がある。一方、そのような「AND」仕様だけでは、詳細・広範な語彙情報の記述を許すシステムが有効活用されないという問題が生じる。そこで上記の両方の要求を満たすため、UPFでは以下の2種の対象範囲を設定することとした。

（a）基本言語変換標準

全MTシステムのユーザ辞書で取り扱うことができ、UPFとの間で相互変換（アップロード・ダウンロード）可能であることを推奨する語彙の情報を記述する形式

（b）拡張言語変換標準

各MTシステムで記述する可能性のあるすべての語彙の情報を記述する形式

すでに商品化されユーザに使用されている5つの異なるシステムのユーザ辞書の比較検討を元にして、現在上記2種の仕様設計作業を進めている。基本的には第1年度に基本言語標準を、第2年度に拡張言語変換標準を設計するが、基本言語標準の設計に際しても拡張言語標準を考慮する必要があるのは言うまでもない。また拡張言語標準の設計段階で再度基本言語標準に戻って改良する必要が生じるかも知れない。

3. 2. 基本言語変換標準の概要

基本言語標準の設計に際しては、まず各システム共通の「記述用語」の設定作業を行なう必要があった。つまり、品詞のセット、品詞の呼称など用語と定義の統一から作業を開始した。特に日本語の場合、基本となる品詞設定についても学校文法では機械翻訳にとって不十分であり、準拠すべき標準が存在しない。具体例としては「形容動詞」という品詞を独立の品詞として立てているシステムと「形容詞」の下位として扱っているシステムが存在した。またその登録単位も語幹登録、終止形登録の二通りがあっ

た。このような用語・形式の統一を行ない、原案作りをすすめている。なお、基本言語標準の設計に際しては、言語学的に厳密な現象記述よりもデータの流通性に重点を置いている。

4. 基本言語変換標準

4. 1. 方針

上記で述べたとおり、基本言語変換標準においては、言語的な厳密さより機械翻訳における情報流通性を重視している。また、機械翻訳用のユーザ辞書としての有効性を得るために以下の方針としている。

1. 従来の機械翻訳ユーザの辞書登録は9割が名詞・固有名詞であるため、この2品詞の登録を容易にする。多くのユーザは補助情報を設定せずデフォルト値で登録しているため、必要最低限の情報のみの登録とする。
2. 日英・英日の登録を1回で済ませるために、英日・日英は共通形式とする。

4. 2. 記述情報

この方針に基づき、基本言語変換標準では以下の品詞を扱い、これ以外の品詞は拡張言語標準で扱うこととする。

日本語：名詞、動詞、形容詞、形容動詞、副詞

英語：名詞、動詞、形容詞、副詞

また英語－日本語間の品詞組み合わせは表1に示す内容とする。

以下、各品詞において扱う情報を記述する。

日本語にかかわる情報

名詞 種類：普通名詞、固有名詞

表1：品詞組み合わせ

英語	名詞	動詞	形容詞		副詞
日本語	名詞	動詞	形容詞	形容動詞	副詞

意味分類：人、組織、その他具体物、場所、時間、その他抽象物

動詞 活用型：一段、五段、カ変、サ変

格要素：が、を、に

格要素の意味制限

英語の格パターンへのマッピング情報

形容詞

格要素：が

格要素の意味制限

英語の格パターンへのマッピング情報

形容動詞

活用型：なだ型、のだ型

格要素：が

格要素の意味制限

英語の格パターンへのマッピング情報

副詞 記述情報なし

英語にかかわる情報

名詞 冠詞指定：the/none

可算性：可算 / 不可算

数限定：常に複数 / 常に単数 / その他

活用形：複数形

先頭音：母音 / 子音

意味分類：人、組織、その他具体物、場所、時間、その他抽象物

動詞 活用形：三単現、過去、過去分詞、現在分詞

先頭音：母音 / 子音

格パターン：SV、SVC、SVO

格要素の意味制限

日本語の格パターンへのマッピング情報

形容詞

活用型：er-est 型 / more-most 型 / 無変化

先頭音：母音 / 子音

副詞 活用型：er-est 型 / more-most 型 / 無変化

先頭音：母音 / 子音

注意：

- ・日本語の活用語は終止形で、英語の動詞は原形、形容詞・副詞は原級で登録する。
- ・サ変動詞の語幹のみが名詞になる場合は、語

幹を名詞として別途登録する。

- ・ 格要素の意味制限は、要素ごとにその格要素に許諾される意味分類を記述する。

格パターンのマッピング記述方法

英語の語順に従い、各要素が日本語のどの格要素に対応するかを記述する（ただし V はのぞく）

例：SV文型の場合：英語の S に対応する日本語の格を記述する

SVO文型の場合：英語の S, O に対応する格をこの順に記述する。例えば eat の場合、He eats an apple. のように用いられるため、「が格, を格」に関する情報をこの順に情報を記述する。

4. 3. 記述例

本節では以上の情報の記述の例を示す。以下に示す例は、記述例を概念的に示したものであり、マークアップ形式、項目名、項目値など今後変更の可能性がある。

名詞の例

<エントリー>
<日見出し>本</日見出し>
<日品詞>名詞</日品詞>
<意味>その他具体物</意味>
<英見出し>book</英見出し>
<英品詞>名詞</英品詞>
<英可算性>可算</英可算性>
<英複数形>books</英複数形>
<英先頭音>子音</英先頭音>
</エントリー>

動詞の例

<エントリー>
<日見出し>食べる</日見出し>
<日品詞>動詞</日品詞>
<日活用型>一段</日活用型>
<英見出し>eat</英見出し>
<英品詞>動詞</英品詞>
<英先頭音>母音</英先頭音>
<英活用形>eats, ate, eaten,

eating</英活用形>

<格関係>[が]=人

[を]=その他具体物</格関係>

</エントリー>

形容詞の例

<エントリー>

<日見出し>速い</日見出し>

<日品詞>形容詞</日品詞>

<英見出し>fast</英見出し>

<英品詞>形容詞</英品詞>

<英先頭音>子音</英先頭音>

<英活用型>er-est 型</英活用型>

</エントリー>

4. 4. エディタ

以上の情報を簡単に登録できるようにエディタを作成している。このエディタは情報の登録、削除の他、参照ができる。このためMTを使用せずに翻訳を行っている場合でも、他のユーザーが登録した情報を参照することができ、最新語句情報が得られ翻訳生産性を向上できる。

5. おわりに

本稿では、異システム間でユーザ辞書データを交換・流通させるための共通フォーマット（UPF）開発活動の概要を述べた。UPFとしては、基本標準、拡張標準の2つを開発する方針である。ワーキンググループで原案を作成し、AAMT加盟メンバーの承認を経て、MT業界の標準とし、電子ネットワークによってユーザ辞書データを流通させるのが目標である。この活動の第1年度末である現時点では、基本標準の仕様がほぼ固まっている。今後は各システムのユーザ辞書との間の双方向変換の確認を行なうのと並行して、拡張標準の設計を行ない、来年度末に基本・拡張の両仕様をFIXして一般公開する予定である。この活動が、個人のもつノウハウの交換、流通を活性化し、MT技術の普及に貢献し、日本人の外国語情報受発信を促進することを望んでいる。