商用機械翻訳ユーザ辞書の共通フォーマット設定に向けて

4B－4

亀井 真一郎 *1　平井 徳行 *2　斎藤 由香梨 *3
伊藤 悦雄 *4　赤羽 美樹子 *5　檜山 努 *6　村木 一至 *1
*1 NEC　*2 シャープ　*3 富士通研　*4 東芝　*5 ノヴァ　*6 NEC情報システムズ
kamei@hum.cl.nec.co.jp, nnd6@isl.nara.sharp.co.jp, yukari@ling.flab.fujitsu.co.jp,
etsuo@sp.tokyo-sc.toshiba.co.jp, aka@nova.co.jp, {hiyama, muraki}@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）の創造的ソフトウエ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）　共通フォーマットと各システムのユーザ辞書の間の双方向コンバータの提供
（3）　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、共通フォーマッ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。上記（2）の双方向コンバータは、各MTメーカがそれぞれ独自に開発するものである。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（3）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

Common Formats for User Dictionaries of MT Systems
Shin-ichiro KAMEI, tokuyuki HIRAI, Yukari SAITOH, Etsuo ITOH, Mikiko AKABANE, Tsutomu HIYAMA, Kazunori MURAKI

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に準拠したタグを用いる。

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。現実に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、基本言語変換標準としては、現在のところ以下を対象として原案作りをすすめている。

日本語：　名詞、固有名詞、動詞、サ変動詞、
　　形容詞、形容動詞、副詞

英語：　名詞、固有名詞、動詞、形容詞、副詞

上記の品詞設定は、実際に現在までの１０年間にＭＴユーザが登録した辞書の約９割が名詞、固有名詞であるという各社に共通した経験的データに基づいている。名詞の下位分類とすることも可能な固有名詞を独立の品詞として立てたのは、ユーザが高頻度で登録する可能性のある語群であることがその理由である。固有名詞と同様に、サ変動詞も登録率を考慮して動詞の下位でなく独立品詞とした。このように、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た。

4　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。