機械翻訳ユーザ辞書データ流通・相互利用のための共通フォーマット設定活動
－アジア太平洋機械翻訳協会の活動報告－

6Q－1

桧山 努 *1 亀井 真一郎 *2 平井 徳行 *3 斎藤 由香梨 *4
伊藤 悦雄 *5 藤井 美樹子 *6 高橋 雅仁 *7 村木 一至 *2
*1 NEC情報システムズ *2 NEC *3 シャープ *4 富士通研 *5 東芝 *6 ノヴァ *7 九州松下電器

1．はじめに

アジア太平洋機械翻訳協会(AAMT)では、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ひとつとして、機械翻訳システム(MTシステム)の機種の違いを超えて、ユーザ辞書を相互に利用可能とする仕組み(UPF:Universal PlatForm)を構築する活動を平成8、9年度の2年間をかけて行なってきた。本稿ではこれまでの活動について報告する。

2．UPFの全体構想

MTシステムはユーザ辞書を充実することによって、有効に活用することが可能である。しかしながら、ユーザ辞書の作成は一般に時間と労力がかかる仕事であるため、個人個人で充実させるには限界がある。現在は、異種のMTシステム間でユーザ辞書の相互利用ができないため、同じ語彙に対する辞書を別々に作成しなければならない。そこで、図1に示すように、ある人が作成したユーザ辞書を共通のフォーマットに変換し、インターネットを介して、そのデータをやり取りすることができる環境を構築し、個人が使用できるユーザ辞書を飛躍的に増大させるための活動を行なった。このような環境により、ユーザ辞書作成の負担が軽減され、機械翻訳の翻訳品質の向上にもつながる。このような仕組みを構築するために、以下のような活動を行なった。

(1)ユーザ辞書共通フォーマット(UPF言語変換標準)の設計
(2)UPF言語変換標準に準拠した辞書作成を支援するためのソフトウェアの開発
(3)UPF言語変換標準に準拠した対訳辞書の開発
(4)UPF言語変換標準に準拠した辞書を交換できる環境の構築
(5)UPF言語変換標準の有用性の評価実験

UPF活動の成果であるUPF言語変換標準の記述仕様、対訳辞書エディタ等のソフトウェア、2万語の日英・英日対訳辞書、および電子環境は無償で提供する。各社は、これらを利用して各MTシステムとUPF言語変換標準との間のコンバータやUPF言語変換に準拠した対訳辞書を提供することができる。また、MTシステムユーザは、自分の作った辞書の提供や他人が作成した辞書の利用が可能である。

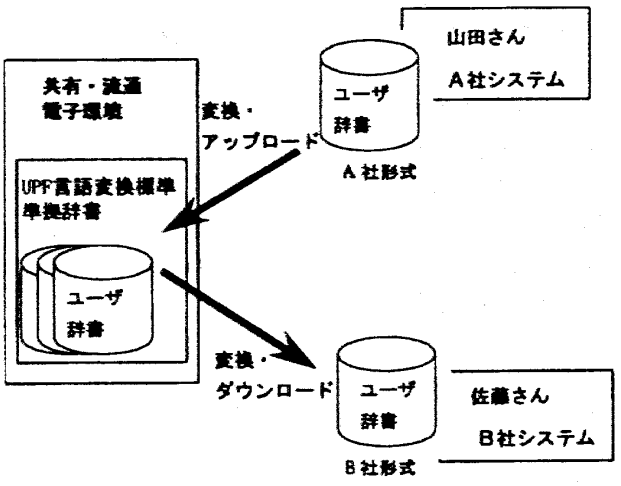

図1 UPFの全体構想

3．拡張言語変換標準の設計

3.1基本言語変換標準と拡張言語変換標準

UPF言語変換標準としては、各MTシステムのユーザ辞書で共通に取り扱うことができ、UPF言語変換標準との間で相互変換可能であることを推奨する辞書情報を記述する形式である「基本言語変換標準」と、より詳細・広範な辞書記述を行なうための記述形式である「拡張言語変換標準」に分けて設計を行なった。これらは、実システムとの乖離を避けるために、各MTシステムのユーザ辞書の辞書情報を比較分析して決定した。基本言語変換標準では、ユーザ辞書に登録される辞書情報の頻度、ユーザにとって辞書記述内容の明確さ等を考慮して制限を設けた。例えば、日本語の品詞は「名詞、動詞、形容詞、形容動詞、副詞」を、英語の品詞では「名詞、動詞、形容詞、副詞」のみを設定した。各MTシステムとの相互変換性は、これらの制限によって阻害される可能性があったが、ユーザ辞書に最低限必要で、相互変換性も損なわれない辞書情報を設定することができた。これに対して、拡張言語変換標準は、

Common Formats for User Dictionaries of MT Systems
Tsutomu HIYAMA*(hiyama@ats.nis.nec.co.jp),Shin-ichiro KAMEI,Tokuyuki HIRAI,Yukari SAITOH,Etsuo ITO,Mikiko FUJII,Masahito TAKAHASHI, Kazunori MURAKI
* NEC Informatec Systems, Ltd.

広範、詳細な辞書情報を記述できるよう次のような方針で設計を行なった。

(1)各社MTシステムのユーザ辞書情報のうち、基本言語変換標準から除いた辞書情報を記述できるようにする。
(2)あらかじめ設定した辞書情報以外の情報を、辞書記述者が新規に定義できる枠組みを備える。

3.2 拡張言語変換標準

基本言語変換標準については、すでに報告されているので[1][2][3][4]、ここでは、拡張言語変換標準の概要について述べる。まず、品詞であるが、基本言語変換標準設定したもの以外に、上記(1)の方針に基づき、以下のものを設定した。

[日本語の品詞]
連体詞、接続詞、単位、感動詞、助詞、助動詞、接頭語、文
[英語の品詞]
限定詞、接続詞、単位、間投詞、前置詞、助動詞、代名詞、疑問詞、文

また、基本言語変換標準では、日本語の名詞に対して英語の名詞のみを対応させるなど日英の品詞の対応関係に制限を設けた。これに対して、拡張言語変換標準では、日本語の「原子」に対して「atomic」という英語の形容詞が対応できるように、日英の品詞の対応関係に制限を加えないこととした。

拡張言語変換標準で記述された辞書情報に関しては、各社システムで利用できる情報のみコンバータで変換して使用する。

3.3 新規辞書情報の定義方法

拡張言語変換標準では、各MTシステムのユーザ辞書を比較分析して、具体的な辞書情報を設定した。しかしながら、辞書情報は今後も新たに設定されうる。そこで、辞書記述者が辞書情報を新規に定義すできるようにするために、以下のような記述形式を設定した。

```
<tagdefine>
 <tag_name>新規タグ名</tag_name>
 <tag_descript>タグの説明</tag_descript>
 <parent_tag>親のタグ名</parent_tag>
 <value_sets>値のリスト</value_sets>
 <value_sets_descript>値の説明
 </value_sets_descript>
 <tagdefine_comment>コメント
 <tagdefine_comment>
</tagdefine>
```

例えば、新規に連体詞の型（rentaitype）を辞書中に記述しようとする場合、次のように定義することができる。

```
<tagdefine>
 <tag_name>rentaitype</tag_name>
 <tag_descript>日本語連体詞の型を記述
 </tag_descript>
 <parent_tag>japanese</parent_tag>
 <value_sets>あらゆる型,指示型,の型,たる型
 </value_sets>
 <value_sets_descript>いわゆる型「いわゆる、あらゆる等」,の型「木製の、金属の等」,たる型「堂々たる、絢爛たる等」,指示型「この、その等」
 </value_sets_descript>
 <tagdefine_comment>-
 <tagdefine_comment>
</tagdefine>
```

4．おわりに

本稿では、MTシステムの機種の違いを超えて、ユーザ辞書を相互に利用可能とする仕組み構築する活動の成果について、その概要を報告した。現在、数社が、UPF言語変換標準に準拠した辞書と自社MTシステムとのコンバータの作成を開始している。活動の詳細は、AAMTのホームページ[5]で随時公開しているが、最終的な活動成果は、平成9年度末に公開予定である。本活動の成果が、対訳ノウハウの流通促進、MT技術の普及に貢献し、日本人の外国語情報受発信の促進につながることを期待している。

参考文献：

[1]亀井 他: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 情報処理学会 第54回全国大会
[2]伊藤 他：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-言語処理学会 第3回年次大会
[3]赤羽 他:機械翻訳ユーザ辞書データ流通のための共通フォーマット －アジア太平洋機械翻訳協会の活動報告－人工知能学会第11回全国大会
[4]Kamei et al:"SHARABLE FORMATS AND THEIR SUPPORTING ENVIRONMENTS FOR EXCHANGING USER DICTIONARIES AMONG DIFFERENT MT SYSTEMS AS A PART OF AAMT ACTIVITIES" MT Summit VI 1997
[5]http://www.jeida.or.jp/aamt/index.html