4K－7

日英機械翻訳システムのための前編集支援ツールの開発 ～ 開発の目的と概要 ～

亀山 三穂*　伊藤 昭典*　松平 正樹**
*(株)沖テクノシステムズ ラボラトリ　**沖電気工業(株)

1 はじめに

現在の機械翻訳システムは，プリエディット，ポストエディットなど人間の介在が必要不可欠である。

しかし，日英翻訳で訳質を左右する重要な要素であるプリエディットに関しては，どこをどのようにプリエディットすればよいのかが不明確で，作業者の経験に基づいて試行錯誤的に行われているのが現状である。

我々は，実用的な日英翻訳システムの開発を目標として，日英機械翻訳システム PENSÉE のためのプリエディット支援プログラムの開発を試みた。本稿では，まず，現状のプリエディット作業を分析し，プリエディットでどのようなことをすべきかを説明する。次に，我々が開発したプリエディット支援プログラムの構成および機能について述べる。

2 プリエディットの現状

実際にプリエディットしなければならない文はどのくらい発生するのであろうか。

8 種類のマニュアル文 134 文 (116 文 + 36 名詞句だが，名詞句を 0.5 文とした) を専任のオペレータに正しく翻訳できるようにプリエディットしてもらい，どのようにプリエディットしたかを以下の項目に従って分類してカウントした。

1. 文の分割
2. 省略要素の補完
3. 構文的曖昧性の指定
4. 意味的曖昧性の指定 (多義語の言い換え)
5. その他の翻訳困難な表現の修正

結果を表 1 に，プリエディトの例を表 2 に示す。

結果から，プリエディットは平均 1 文につき 2 ヶ所行われ，そのうち「 5. その他の翻訳困難な表現の修正」が全体の 80% 近くを占めていることがわかる。このことから，従来考えられていた構文的，意味的な曖昧性に加え，特定の単語 (列) あるいはパターンに依存する言語現象に応じたプリエディット支援が重要であることがわかる。

表 1: プリエディット数

文書	文数	分野	プリエディット数					
			総数	分割	補完	構文	意味	その他
A	26.5	計算機	48	1	7	2	1	37
B	24.5	計算機	39	1	3	2	0	33
C	15	電子	28	5	1	2	1	19
D	11.5	化学	35	3	2	0	2	28
E	23	機械	55	8	0	3	4	40
F	10	電気	22	2	1	1	0	18
G	10	電子	21	3	1	3	1	13
H	13.5	機械	20	0	0	2	1	17
合計	134	-	268	23	15	15	10	205
(%)				(8.6)	(5.6)	(5.6)	(3.7)	(76.5)

表 2: プリエディット例

分類	原文	プリエディット原文
分割	‥の順に組立て、モータ側面を‥	‥の順に組立てる。モータ側面を‥
補完	‥規定するもので、版下作成時‥	‥規定するものです。これは、版下作成時‥
構文	‥流体の温度、圧力、密度、粘度の‥	‥流体の [温度、圧力、密度、粘度] の‥
意味	‥メッセージを出す‥	‥メッセージを出力する‥
その他	‥取得が可能であること‥	‥取得できること‥

3 プリエディット支援プログラム

正しく機械翻訳するためには，プリエディットすべき部分を検出し，どのように変更すればよいかを指示してくれるようなプリエディット支援システムが必要であると考える。

一方では，正しく翻訳できない語句を自動的に書き換えるような自動プリエディットの方式が提案されている [4]。しかし，この方式では書き換えのルールは開発者が作成するものであり，文法に記述することと本質的にあまり違いがない。すなわち，文法では記述が困難である曖昧性の解消や，開発者がカバーできないさまざまな単語の言語現象を翻訳できるようにするためのプリエディット (支援) とは基本的に異なる。

我々は，上記の問題点を考慮し，以下のような機能をもつプリエディット支援プログラムを開発した。

1. プリエディットすべき箇所の指摘
2. その部分に対するプリエディット候補の表示
3. 使用者が候補を選択することによる原文の自動修正

Development of Computer Aided Pre-Editing Tool for Japanese-English Machine Translation System - Purpose and Overview -

Miho KAMEYAMA*, Akinori ITO*, Masaki MATSUDAIRA**
*Oki Technosystems Laboratory, Inc.
**Oki Electric Industry Co.,Ltd.

4. プリエディットすべきパターンとその変換候補の記述ルール提供
5. 使用者も記述可能な変換ルール

3.1 本プログラムの構成

本プログラムは，入出力部，プリエディット部およびプリエディットルール群から構成される。

入出力部は，使用者への表示や入力などを処理する部分である。

プリエディット部は，日本語形態素解析の結果とプリエディットルール群を利用して，プリエディトすべき部分を指摘し，プリエディット候補を出力する。

プリエディットルール群は，プリエディットの対象となるパターンとその変換パターンを集めたもので，システムがあらかじめ持っているものと使用者が記述するものとの 2 種類がある。

日本語の形態素解析は，日英 PENSÉE の日本語形態素解析部を利用し，プリエディット候補の日本語生成は，英日 PENSÉE の日本語形態素生成部を利用している。

3.2 プリエディットルール群

プリエディットルール群は，単語変換ルールと構文変換ルールから構成される。これらは，使用者が自由に登録できる。

単語変換ルールは，プリエディット項目の「4. 意味的曖昧性の指定」に対応するもので，意味的に曖昧性のある単語について，その意味を一意に決定するような言い換え候補を変換パターンとして持っている。例えば，動詞「かける」に対しては,「注ぐ」,「建造する」,「乗ずる」などの言い換え候補を持っている。

構文変換ルールは,「5. その他の翻訳困難な表現の修正」を含むすべてのプリエディット項目を処理するもので，単語 ID，見出し，品詞，変化形，連続数からなるマッチングパターンと変換パターンおよび変換のガイダンスから構成される。それぞれのパターンは複数個記述できるものとする。例えば,「A の B する」を「A が B する」に変換するパターンは以下のようになる。

- マッチングパターン：
  [1] [] [名詞] [] [], [2] [の] [格助詞] [] [],
  [3] [] [動詞] [] []
- 変換パターン：
  s1, a[が] [格助詞] [] [], s3
- ガイダンス：
  主格を示す「の」を「が」に変更してください。
  「A の B する」→「A が B する」

なお，マッチングパターン，変換パターンで未記入の項目は，指定なしあるいは 1 を表し，s1,s3 は，それぞれ単語 ID が 1,3 の単語をそのまま出力することを表す。

現在，システム内部の変換ルールとして約 300 ルールを作成した。

3.3 ユーザインタフェース

本プログラムは Windows 上で設計した。

原文を読み込み，本プログラムの検索機能を起動すると，原文表示画面と編集画面が表示される。使用者が，ガイダンスに従って，プリエディット候補の中から該当するものを選択することにより，自動的に原文ファイルが修正される。この時，プリエディット候補の中に該当するものがない場合は直接入力することもできる。

4 まとめ

日英機械翻訳におけるプリエディットの現状を調査し，従来考えられていた構文的，意味的な曖昧性に加え，特定の単語(列)あるいはパターンに依存する言語現象に応じたプリエディット支援が重要であることがわかった。

このような現状に対し，我々は，プリエディットすべき箇所およびそれに対するプリエディット候補を表示することにより，原文を容易に修正できるようなプリエディット支援プログラムを開発した。本システムでは，ユーザ記述も可能なプリエディットパターン変換ルールを利用可能とした。

今後は、ルールをさらに拡張してゆく計画である。また，ユーザが変換ルールを容易に記述できるようにするためのツールを提供する計画である。

参考文献

[1] アジア太平洋機械翻訳協会：制限日本語 (第 1 版), 研究成果報告書 (93 年度版), アジア太平洋機械翻訳協会 (1993)

[2] 平井 章博, 高岡 紀子, 梶 博行：日英機械翻訳用前編集支援システム (1) 構文的曖昧性の検出方式, 情報処理学会第 36 回全国大会, 2U-2 (1988)

[3] 林 良彦, 菊井 玄一郎：日本文推敲支援システムにおけるの書換え支援機能の実現方式, 情報処理学会論文誌, Vol.32 No.8 (1991)

[4] 白井 諭, 池原 悟, 河岡 司：日英機械翻訳における原文自動書き替え型翻訳方式とその効果, 情報処理学会研究会, NL-95-12 (1993)