人間の言語知識と機械支援を融合させる機械翻訳インタフェース：Source and Target Language Mixed Stage

亀井 真一郎　村木 一至　佐藤 研治　浜田 和彦 (*)
NEC 情報メディア研究所、(*) NEC 情報システムズ

1 はじめに

近年、パーソナルコンピュータの普及に伴って電子テキストが大量に蓄積・流通するようになり、また電子ネットワークの発達によって日本人が外国語と直接接触する機会が急速に増大している。これに伴い、日本人が海外に向けて情報を発信する状況も急増しているが、従来の機械翻訳 (MT) システムはこの需要に充分に応えるに至っていない。

従来のバッチ処理型日英MTシステムでは、結果を機械任せにしてしまい、書き手の持っている言語知識と機械による支援が充分に融合しているとは言い難い。MT単独で充分な翻訳品質の得られない現段階の技術では、結果を相手に読んでもらう必要のある日英翻訳の出力結果は、利用されることなく棄却されてしまう場合が多く、機械による支援技術が充分に生かされていない。

現在のMT技術を有効に使って日本人の外国語での情報発信を促進するため、我々は日英翻訳システムを情報発信支援ツールと捉え直し、インタラクティブ型日英変換インタフェースを提案し [1, 2, 3]、PC上のフロントエンド型ツールとして実現した [4, 5, 6, 7]。

このインタフェースの最大の特徴は、原文 (日本語) と訳文 (英語) の間の段階として、原言語と目的言語が混在した「日英混じりの段階」(Source and Target Language Mixed Stage: STLM Stage) を設けたところにある。この段階においてユーザは、仮名漢字変換で慣れ親しんだ入力・変換・選択の操作を繰り返すことで、部分部分の変換結果を確かめながら、全体の英語を構築してゆくことができる。本稿ではこの STLM Stage の特徴について述べる。

2 Source and Target Language Mixed Stage：語句変換と構文変換の分離

人間のもっている言語知識と機械による支援との融合を図るため、我々はインタラクティブな外国語作成支援インタフェースを提案した [1, 2, 3, 5]。

図1は日英変換の基本的段階を示している。入力された日本語 (a) は自動的にリアルタイムに形態素解析され、自立語部分が抽出されてその部分が英単語に置き換わる (b)。この段階では語順は日本語のままで単語が英語に置き換わった、いわば、日本語英語混じり文の状態である。この段階を Source and Target Language Mixed Stage (STLM Stage) と呼ぶ。この段階で使用者が構文変換を指示すると全体が英語語順に並び換わる (c)。

(a) 私は彼に論文を送った
(b) I は he に paper を send た
(c) I sent him a paper

図 1: 日英変換の基本手順

我々の提案するインタフェースは、上記の日本語英語混じり文の段階 STLM Stage (b) という中間段階を設けたところに特徴がある。日本語と外国語 (英語) の差は、語句の変化と構文の変化とに分離できる。従来の機械翻訳では、二つの変化を一度に処理して結果を出力していたため、その訳出結果が充分でない場合、その原因が語句選択の不備によるものか、構文選択の不具合によるものかが判断しにくかった。特に日本語と英語のように語順の大きく異なる言語対の場合、対応する語句の位置が大きく変化するため、原文と訳文とを対照させて訳文の妥当性を判断する際に、視点の移動が大きく、確認作業を困難にしていた。

新しいインタフェースでは、原文と訳文の中間にこの STLM Stage という原文と訳文の混在した段階を設けることによって、翻訳を訳語変換と構文変換を分離して操作の透明性を増している。この段階では文中の各単語は英語に置き換わっているが、語順および助詞・助動詞など文の骨格は日本語のままである。したがってシステムの使用者はこの段階 (b) において、大きな視点の移動をすることなしに訳語を確認しながら全体の英文を作成することができる。この日英混じりの段階を介してユーザはごく簡単にシステムの処理に介入することができる。

またこのインタフェースは、広く普及している仮名漢字変換インタフェースを日英変換に拡張したものとなっている。したがって、対話制御を行なうにあたり、使用者が全く新たな操作法を憶えなければならないという負担が軽減されている。仮名漢字変換システムにおいては、システムが提示する候補の漢字群の中からユーザが漢字

を選択することが出来る。そのインタフェースと全く同様に、訳語を選択するためのこの中間段階 (b) では、システムが候補となる英単語を表示し、ユーザがそれを選択できる構成とっている。

例えば上記の例の場合、段階 (b) では入力日本語文中の「論文」という単語は「paper」という英単語に置き換わっている。ここで「paper」にカーソルを合わせると次のような訳語候補が表示される。

論文		
paper	[名詞]	[典型訳語]
dissertation	[名詞]	[学位～]
essay	[名詞]	[一般～]
thesis	[名詞]	[卒業～]
・・・・・・・・・・・・		

図 2: 候補英訳語の表示

つまり日本語から見た英語の類義語が提示されることになる。ユーザはその中から、どの英単語を用いればよいかを考えて適切な訳語を選択できる。ここで品詞（図中の例では［名詞］）の次に表示されているのは、候補の英単語を選別するため補助情報である。英単語を選択する場合にはどの単語を選択すべきかがあらかじめ分からない場合が多いので、このような選択のための補助情報を表示することで使用者の便宜を図った。

3 日英構文変換の実現

3.1 多段階構文変換

提案したインタフェースでは、ユーザは STLM Stage を介して語変換、句変換、文変換という変換の各段階で結果を確認しながら段階的に変換を行う。このことを文レベルにも拡張し、複文・重文の場合にも、原則として用言を一つ含む単文区間の変換を繰り返して文全体を構文変換する多段階構文変換を特徴としている。次図に「私は彼が買った本を読んだ」を「I read a book he bought」に変換する手順を示す。

(a) 私は彼が買った本を読んだ
(b) I は he が buy た book を read だ
(c) I は a book he bought を read だ
(d) I read a book he bought

図 3: 複文の変換手順 (1)

変換の途中で文に部分的に引かれた下線は、次の段階で構文変換を行う範囲を示している。システムが提示した範囲をユーザが確認・修正したのち構文変換を指示するとその範囲が構文変換され英語表現に変換されて元の文に埋め込まれる。図の例では、段階 (b) でまず「彼が買った本」という英語の関係節の相当する単文相当部分が英語に変換される。その後その部分の変換結果を含んだ文全体が英語に構文変換される。

このインタフェースでは、一文は多段に分けて構文変換され、 変換区間は一般には入れ子となる。このため一度変換した結果を外側の 構文制約に従って生成しなおさなければならないことがある。例えば 「私は彼が本を読むのを助ける」という文は、最初に埋め込みの区間 「彼が本を読む」を変換し、次に全体を変換するという手順を踏むが、 最初の変換で得られるのは「he reads a book」という文である。 これを「助ける = help someone to do」という構文と組み合わせて 「I help him to read a book」という結果を得るためには、最初の単文区間の変換結果を不定詞句として生成しなおす必要がある。 このような再生成を可能にするために、システムの構文変換部は 生成結果の英語構文木をすべて保持している。構文木中に保持された生成途上の情報を添えてシステムの英語生成部を起動することで再生成が可能となる。

(a) 私は彼が本を読むのを助ける
(b) I は he が book を read の を help
(c) I は he reads a book を help
(d) I help him to read a book

図 4: 複文の変換手順 (2)

またこのシステムは、多段階変換の各段階からその前段階に戻ってエディットし直すＵＮＤＯ機能を備えている。例えば日本語英語混じり文の段階 (b) では、表示された各英単語を元の日本語に戻すことができる。同様に単位文が英語構文に変換された状態 (図中の (c)) では、全く同じインタフェースによって、単位文を元の日本語英語混じりの状態に戻すことができる。多段階変換の各段階におけるＵＮＤＯを、常にその一段階前の状態に戻るという操作に共通化することによりユーザが新たに憶えるべき手順を最小限にすることができる。

3.2 日英対訳慣用表現

このインタフェースでは語句の訳語はユーザが選択するのを前提としているが、訳語を全て指定するのはユーザにとって負担が大きい。これを軽減するため、このインタフェースは、複数の単語の出現によってそれらの単語の訳語が定まるような慣用表現処理機構を同時に備えている [7, 8]。

日本語では「電話をかける」「コピーをとる」といった多くの慣用的な表現が日常的に使用されている。これらの表現では、句を構成する要素の組み合わせによって

各単語の語義(訳語)が決定される。例えば「電話する」という意味の「電話をかける」という表現は英語の「make a phone call」という表現に対応し、「かける」という動詞は英語の「make」に対応している。しかし動詞「かける」は「hang」「lack」等に対応する非常に多くの語義をもち、通常は「make」に対応することは少ない。「電話をかける」という表現の場合には、「かける」が「make」に対応するという情報を慣用表現用の辞書に蓄積しておき、それを元にシステム側が訳語選択を行なう機構を実装した。さらに「壁に電話をかける」のように「電話をかける」が構成要素の語義通りの意味を持つ場合と、全体で「電話する」という慣用的な意味を持つ場合とを、通常の語句の訳語選択と同一のインタフェースで切り替えられるように操作手順を実現した[6]。

4 Front-End Processorとしての実現

システムに機能を盛り込みすぎると、使用者が憶えなければならない操作が増え、ユーザに負担を強いることになる。使用者が使い慣れた文書作成環境(ワープロやエディタ)と融合できる構成にすれば、ワープロやエディタの備えている種々の機能、たとえば文書フォーマット整形機能などを利用することができ、利用者は同じ機能に対して似て非なる操作を憶える必要がなくなり、作業に専念できるようになる。またMT機能と組み合わせたいソフトウエアは通常のワープロやエディタだけに限らない。例えば、自分の使っている電子メールソフトに外国語作成支援ツールが add-on できれば非常に効果的である。

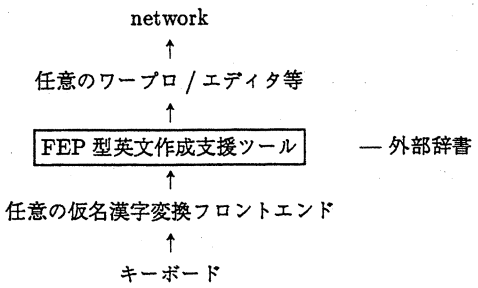

図 5: FEP 型英文作成支援ツールの構成

以上の点を考慮して、我々は日英変換機能をフロントエンドプロセッサ型のツールとして試作した。すなわちこのツールは他の任意のアプリケーションプログラムから起動をかけることができ、変換結果の文字列を元のアプリケーションプログラムへ送信することができる。さらに任意の仮名漢字変換ソフトとも組み合わせ可能とした。すなわち使用者は自分の慣れ親しんだ文書作成環境の上に、英文作成支援環境を容易に構築することができる。また、外国語文書を書く際には、外部の資料、例えば英和・和英辞書や文法書、例文集などを頻繁に参照するのが普通である。そこでこのＦＥＰ型ツールからはワンタッチで外部辞書を検索できる機能も装備した。図はこのＦＥＰ型英文作成ツールを使用する際の他のソフトウエアとの関係を示したものである。

5 ディスカッション

5.1 日英システムと英日システムの非対称性

まず、MTシステムを情報の受発信の観点から分析してみる。単にモジュール構成の視点からは、日英システムと英日システムとは日本語と英語とが入れ変わっただけの鏡対称のように見える。実際、現在多数製品化されている従来型のMTシステムのほとんどは日英も英日も同様のインタフェースを持っている。

しかしながら、MTシステムとそれを使用するユーザとを含めた全体を大きな一つのシステムで考えると、日英と英日とは全く異なるシステムであることがわかる(図6)。現在、非常に多くのMTシステムが製品化されているが、現実に利用されているのは英日システムである。英日システムの場合、

- とにかく訳してみる
- でてきた日本語で概略をつかむ
- 詳しく読みたければ原文に戻る

という使い方が可能だからである。出力結果が母国語であるため、多少結果が不適切でも、人間の母国語把握能力の高さと原言語類推能力との組み合わせによって英日MTシステムは情報収集ツール足り得ている。

図 6: 日英翻訳と英日翻訳

日本人にとって	日本語 → 英語	英語 → 日本語
情報の	発信	受信
原文は	母国語 エディット可能	外国語 エディット不可能 (時間かければ読める)
結果文は	外国語 良否判定困難	母国語 悪文でも解釈可能
処理は	インタラクティブ 支援処理向き	一括バッチ お任せ処理向き

一方、日本人が英語を書く場合には事情が全く異なる。出力結果の英語は、日本人ユーザ以外が読むためのものであるから、内容を誤解なく伝えるだけの品質が求められるが、MT技術は残念ながらまだその訳質を得るに至っていない。英語は日本人の母国語でないから、システムのMTシステムの出力結果の英語が適切か否かを即時に判断するのは困難である。日本語と英語は表現も語順も大きく異なるから、一旦英語に翻訳された結果を

逐一原文と目視照合して不適切部分を判断し訳文の切替を指示するのは負担が大きい。

MTシステムの能力を引き出すためには、プリエディットの段階で入力文をシステムが処理できるように書き変える(主として文を短く区切る)のが有効であるが、システムが処理できる文のタイプは一般ユーザにとって自明のことではなく、またMTシステムによっても異なるので、効果的にプリエディットできるのは、システムの操作に慣れた専門家に限られていた。文の構造をシステムに指示して翻訳品質を向上させるためのインタフェースとして、システムの解析結果を木構造として提示しユーザがそれを操作する方法も開発されたが、文の構造を木構造として表示するのは一般ユーザになじみがなく、有効なインタフェースとは言えない。このような事情で、現実問題としてユーザが介在できるのはユーザ辞書の語彙を強化する点に限られていたと言える。

一般のユーザが外国語作成にMT技術を手軽に利用できるようにするために、ユーザがその外国語について持っている言語知識と機械による支援とを融合させるインタフェースが求められていた。本稿でのべた STLM Stage はこの要求に応えるものである。

5.2 対話処理と一括処理

多段階変換インタフェースを単純に実現すると最終的な英文を得るまでの変換操作の数が多くなる。しかし一旦選択した訳語や構文が次回の入力の際に最優先される学習機構を備えているので、使う回数が増すごとに対話インタフェースの繁雑さが減少してゆく。

また、その文を最初に変換する際に途中の段階の変換指示を使用者が明示的に行わずに自動的に処理が進むモードを設定することも可能である。そうすればそれは従来のバッチ型翻訳と変わらない。一旦翻訳結果を得た上で、結果の不十分な文に対して段階を踏んで英文を作成するという手順が考られる。対話(インタラクティブ)処理と一括(バッチ)処理は一見、対立する概念として考えられるが、そうではない。対話処理の途中の段階を自動化することで一括処理を実現できるのであるから、その意味で対話処理は一括処理を含むインタフェースであると言える。

6 おわりに

我々は日英機械翻訳システムを外国語情報発信ツールとしてとらえ直して現状のMTの問題点を考察し、ユーザがシステムと対話しながらユーザ自身がもっている言語知識と機械による支援とを融合するインタフェースを提案し、それを備えた英文作成支援ツールを作成した。

原文から訳文へ至る変換の途中段階に、自立語だけが英単語に変換され語順および助詞・助動詞のような骨格は日本語のままである日本語英語混じり文の段階 (STLM Stage) を設けたのが最大の特徴である。構文変換に関しては単文範囲の変換を繰り返す多段階変換方式をとった。このようにすることで、翻訳における訳語選択と構文変換とが分離でき、変換の透明性が増すので、ユーザは変換の各段階で変換結果を確認しながら最終的な英文を作成してゆくことが可能となる。

また、この多段階変換インタフェースを採用することで、変換の各段階で必要な情報を必要なタイミングで提示することが容易になった。さらに、選択のためのインタフェースを変換の段階や言語現象によらず統一することで使用者が憶えなければならない操作を最小限にとどめ、使用者の負荷を軽減した。このシステムはフロントエンドプロセッサ型ツールとして実現したので、使用者は自分の慣れ親しんだ文書作成環境の上に容易に外国語作成環境を構築できるという利点をもつ。将来は、外国語作成に必要となる様々な種類の情報のインテグレーション・プラットフォームとしてこのツールを拡張してゆく予定である。

参考文献

[1] 赤峯 他「日本語入力による英文作成支援システム ー仮名漢字変換から仮名英語変換へー」第 46 回情処全国大会 4B-1. 1993.

[2] 赤峯 他「翻訳機能付きワープロ ー不安と疲れを感じさせないインタフェースー」第 7 回人工知能学会全国大会 1993.

[3] K. Muraki et al. "TWP:How to Assist English Production on Japanese Word Processor" COLING-94. 1994.

[4] 亀井 他 「FEP 型英文作成支援ツール -外国語情報発信の効果的インタフェース -」第 51 回情処全国大会 5H-3 1995.

[5] 山端 他「FEP 型英文作成支援ツール ー日英構文変換部ー」 第 51 回情処全国大会 5H-2. 1995.

[6] 土井 他「FEP 型英文作成支援ツール ー訳語選択のユーザインタフェースと辞書記述ー」第 51 回情処全国大会 5H-1. 1995.

[7] Yamabana, et al. "Interactive Machine-Aided Translation Reconsidered – Interactive Disambiguation in TWP –", Proc. of NLPRS'95, pp.368-376. 1995.

[8] 田村 他「日英機械翻訳のための大規模慣用表現辞書の構築」言語処理学会第 2 回年次大会. 1996.