AAMT

The Asia-Pacific Association for Machine Translation

Journal

No.9
December 1994

アジア太平洋機械翻訳協会

タイ王国 第２回自然言語処理国際会議

日　　程：１９９５年８月２－４日
開 催 地：タイ国バンコック市　ハイアットセントラルプラザ
主　　催：タイ国科学技術環境省国立電子コンピュータ技術センター

趣　　旨：このシンポジウムは自然言語処理に関連するすべてのサブエリアについて、主にアジア諸言語を中心に、その現状把握を図ることを目的に開催する。将来のＮＬＰに貢献するコンピュータ技術にかかわる大規模な研究開発活動の方向づけを論じる場としたい。研究者と参加者の交流を通じてそれぞれ多くの言語人口を抱える研究課題について意見と情報の交換を促進する。

タイ国では９５年を「全国情報技術年」と定め、政府を代表し、国立電子コンピュータ技術センター（科学技術環境省）および会議運営委員会はＮＬＰに関心を寄せる各方面の参加を広く呼びかけるものである。

日　　程：１９９５年３月１５日　論文提出期限、論者・講演主題の推薦期限
４月３０日　論文採用通知の発送
６月３０日　版下必着

＜パネルディスカッション及びチュートリアルの参加について＞

応募方法：主題、および討議者・講演者を募集する。
1,000語以内に主題もしくは論者推薦を提出されたい。

＜論文の公募＞　ＮＬＰ関係の独自の研究について未発表の論文を、以下の主題に限らず公募する。

主　　題：＊自然言語の理解　＊ＭＴ　＊会話処理
＊それぞれの言語における処理技術とその応用
＊文字の認識、文字フォント開発　＊音声の認識と合成
＊文字表記（ワードプロセッサその他）　＊ＮＬＰ製品開発
＊Ｉ／Ｏ処理（ＣＤ記録を含む）　＊データベースの構築とテキスト検索

＜論文提出の手続き＞

提出先：プログラム委員長
期　限：１９９５年３月１５日必着。
原稿量：ダブルスペースで２０頁以内（Ａ４サイズ）。ただし表紙頁を含まない。
体　裁：本文第１頁に１００－２００語の概論を入れる。
表　紙：必ず本文以外に１頁用いること。
表題、執筆者姓名および肩書き、所属、連絡先住所(含電話・FAX番号)E-Mailアドレスを明記
部　数：各３部
使用言語：英語
審査及び採否：すべての論文を審査し、採用分については予稿集に収録する。

問 合 先　Asanee Kawtrakul, Department of Computer Engineering, Faculty of Engineering, Kasetsart University, Bangkok 10900, Thailand(Tel:662-579-4584 Ext 182, Fax: 662-579-6245)

エッセイ

株式会社 十　印　取締役 渡　辺　　学

今から溯ることおよそ８００年。平安の王朝文化は衰退の一途をたどり、大地に密着した武士と行動する民衆が台頭し始めていた。この時代を生きた民衆の騒乱、驚愕、躍動、すなわち中世のエネルギーを現代に生き生きと伝えてくれる世界がある。絵巻物の広く豊かな世界だ。数ある絵巻物の中の傑作、伴大納言絵巻。

絵巻物は、詞書きに語られたストーリーを次々に絵で展開し、一区切りきたところでまた詞書きが入り、また絵になる。読者は、この進行を絵巻物を両手で広げ、右手で巻ながら左へ広げていく。つまり、両手を広げた幅５０～６０センチが見る者の眼前に現れる。切れ目のない時間の流れが、全長約１０メートルにわたって繰り広げられていく。

では、ここで絵巻物の世界を少しのぞいてみよう。……応天門が炎上している。もうもうたる黒煙は火の粉を降らせながら空を覆っている。現場を目指して朱雀大路を急ぐ群衆。大急ぎで朱雀門を駆け抜けた人々は、いっせいに立ち止まって空に巻き上がる火煙を見上げる。壮麗を駆使した建物をなめつくす猛火。美しく強い色彩が見る者の心を躍らせる。群衆のざわめきが聞こえてくる。……。

この絵のことばを駆使した絵巻物の世界が、今また私たちの眼前に姿を変えて現れている。私たちの行動範囲、経済活動はすでに地球規模になり、それにつれて情報を瞬時に、大量に入手できるようになった。情報は、翻訳というフィルターを通すことでボーダーレスになる。歴史は繰り返す。絵巻物に実現された情報伝達方法が、形を変えて繰り返されている。巻物はコンピューター、キーボード、マウスになり、絵と詞書きはテキスト、グラフィック、動画、音声などの要素に変わった。インタラクティブにユーザーと情報交換できるマルチメディアの世界だ。絵巻物の紙はＣＤ－ＲＯＭに代わり、限られた地域の情報は今やグローバルな情報へと拡大した。手段は変わっても根本に流れている人類の英知は今も昔も健在だ。今から１００年後の私たちの子孫は、どんな思いでこの私たちの活動を見てくれるのだろう。

（アジア太平洋機械翻訳協会　理事）

目　次

MMT'94

1987年以来８年の歳月をかけ日本、中国、タイ、インドネシア、マレーシアの５カ国の機械翻訳共同研究は95年3月をもって研究を終了する。この機械翻訳システムは中間言語方式を採用し多言語同時翻訳が可能国際会議などの議事録作成に大きな威力を発揮する。この研究成果を公表する MMT'94国際シンポジウムが12月5～6日の2日間、東京ホテルニューオータニで開催されたが、その中の「研究成果の技術的成果とその応用」に関し各国研究者から発表された。

中間言語方式のＭＴに求められるものとは

清華大学 教授　黄　昌寧（ファン・チャンニン）

かつて、中間言語方式は多言語機械翻訳（MMT）の伝家の宝刀として多くの研究者に支持されてきた[2]。その利点は、トランスファモジュール方式を完全に上回るものとされ、かいつまんでみると、システム設計において、デザイナーは自国の言語と中間言語さえマスターしていれば、他の言語に習熟する必要がない点が強調されてきた。CICCの研究活動が８年目を迎えた期に発表された研究成果報告でも、やはり中間言語方式MMTシステムでさえも、何種類かのトランスファメカニズムが依然として必要との指摘があり、これは将来の改善点として提案されよう。

1. 中間言語方式の仮説

中間言語方式MMTには、少なくとも下記の3つの仮説がある：

(1) n(n-1)個のトランスファモジュールを書き込むタスクは、MMTシステムがn個の言語間で翻訳するためにデザインされる場合、かなり効率化される。
(2)そのようなMMTシステム開発はトランスファ方式システムより容易である。その理由は、開発に参画するデザイナーが、開発にかかわる言語のうち、自国語及びそのシステムのための中間言語だけを習熟すればそれ以外の他言語をマスターする必要がないからである。
(3) CICC版の中間言語の語彙は、概念見出し、リレーション、および属性から成る。中間言語において、概念見出し間のリレーションには汎用性があっても、概念見出しそのものに汎用性が認められない点は重要である[1]。

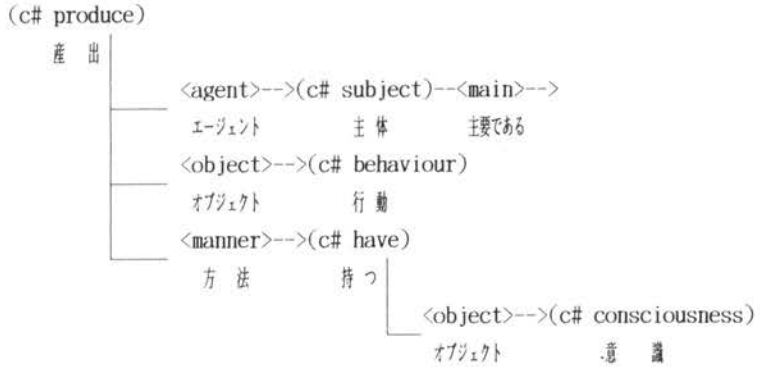
最初の2個の仮説に従えば、中間言語方式によって翻訳する場合、どのような2言語の間でもトランスファー距離は0になるとされる点に注目したい。さらに、最後の仮説により、言語は異なっても同じ意味または命題を表現する等しい文は、固有の中間言語表現を引き起こす可能性があると指摘されている。これはそのような表現で同じリレーションが認める場合、真実とされるわけだが、しかし、これらの仮説の両方について、現在も疑問を感じる次第である。

2. MMTシステムにおける中間言語方式はトランスファメカニズムよりも優れるか？

私はこの件に関して肯定できない。理由を以下に述べる。
(1)　言語AとBが翻訳されると仮定する。言語Aにおける文(a)および言語 Bにおける文(b)は対応する翻訳ペアである。通常、それらの2文の含むワード数あるいは構文構造は違っていることもある。解析フェーズを終えたシステムは、文(a)および文(b)に対して、個々の含むワードが 概念見出し ノードと解釈された場合、異なる構文構造が導かれ、中間言語表現の関係arceへの影響によって、それぞれの文に別個の中間言語表現をもたらす可能性がある。その場合、トランスファメカニズムを用いないシステムでも、なおそれら異なる表現によって文(a)を翻訳した結果が文(b)に、あるいはその逆になるのは正しいとされるだろうか？　以下の単純な例を考察したい。中国語の入力(1a) および英語の入力(1b) は等しい翻訳ペアであるけれども、それらに含まれるワード数は異なる。

(1a)[中国語] 有　意識　地　産生　行為　的　主体
[発音] ユウ　イシ　デ　チャンセン　シンウェイ　デ　ヅゥーティ
[英語] have consciousness <aux> produce behaviour <aux> subject
[日本語] 持つ　意識　<aux>　産出　行動　<aux>　に主体の
(1b)[英語] a subject of consciously producing behaviour
[日本語] （意識的な産出行動の主体）

[図1：中国語の入力(1a)のための中間言語表現]



[図2：英語の入力(1b)のための中間言語表現]

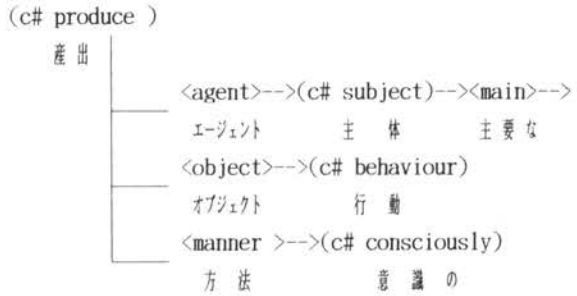


図1と図2にそれぞれ示した中国語の入力(1a)および英語の入力(1b)の中間言語表現にみるように、概念見出しノードは図1に5個、および図2に4個認められ、さらに、上図の'manner'（方法）リレーションの端には、違った表現が表れた。従って、それらの中間言語表現を使い、適当な翻訳を生成するには、システムに何種類かのトランスファメカニズムが必要となる。もしシステムが2個以上の言語を扱い、それぞれに異なる表現を伴う場合、状況はさらに悪くなる。

(2)　周知のとおり、ソース言語に固有の語義に対応する適切な目的言語対を選ぶことは、長年、機械翻訳研究の重要な問題として認められてきた。たとえソース言語において語義の曖昧さ解消に成功したとしても、それら言語対は非常にまれにしか1対の対応を示さないからである。例えば、中国語の辞書には動詞『訂』（ディン）に4個の語義が示されている。この動詞の語義は、ある中国語文では『約 定』（意味=「～と一致」）と取られ、中間言語表現における概念見出しと認められるだろうが、生成段階においてその適切な英語訳を選ぶにはなお、その文脈まで調べる必要がある。つまり、『訂(報紙)』　は「(新聞を)定期購読する」、『訂（ 火車票)』は「(列車の切符を)予約する」とそれぞれ翻訳され、『 訂(貨)』の訳は「(品物を)注文する」である。MMTプロジェク

トでそのようなトランスファ処理をするには、むしろ多言語訳し分け辞書を使うほうが適切であって、一定の概念見出しに基づき１対１の対応を図ろうとするよりも、明らかにより良い。後者（１対の１対応)の手法による処理は、実に困難である。あらかじめ、関与するすべての言語について、すべての概念見出しに伴う適切なコンテキストを定義しなければならないからである。これまで、理想的な中間言語は未だに見いだされていないこと、また、そのようなゴール達成の追求はマンパワーと時間の莫大な消費のため、避けるほうが賢明だと言わざるを得ない。また、MMTプロジェクトに言語の専門家(自国語だけではなく種々の目的言語に習熟した者)の参加を必要とする点で、後者は中間言語方式の2番目の仮説に相反する。

3. CICCプロジェクトのより一層の発展を目指して

CICCのMMTシステムはアジアの5つの言語を伴う世界最大の中間言語方式に基づく MMTの研究開発計画である。同時に、過去8年間にこの分野におけるすばらしい成果を上げ、すでに報告されたように、20対の言語ペアの翻訳の確率は約 50パーセントに達した。他方、中間言語方式MMTリサーチの実施によって、CICCプロジェクトには多くの改善の余地があることが明らかにされた。この計画のより一層の発展を目指し、以下に提案したい。

(1)個々のソース言語の分析フェーズでは、品詞と語義の曖昧さ解消メカニズムが必要である。例えば、例(1a)のほとんどすべての単語に、1個以上の品詞または語義がある。最初の単語『有』(ユウ)は動詞または形容詞として使われる。動詞として使う場合、さらに、最低５個の語義がある。頻繁に使われる単語『地』(デ)でさえ補助動詞と認められたが、しかし、それを『ディ』と発音すると名詞になり、8個の意味を持つ。従って、CICCプロジェクトにおいて、ソース言語とその個々の単語のために正確な概念見出しを得るには、個々の分析モジュールについて、いくつかの曖昧さ解消アルゴリズムを設ける必要がある。

(2) 23個の汎用リレーションが例文と共に定義されたCICCの中間言語であるが、たとえ同じソース言語の単文であっても、２人の人間が処理した場合、同じ中間言語表現を生み出すとは保証できない。また、翻訳タスクがコンピュータにより行われるとき、さらに結果は思わしくない。そこで、特定の中間言語によって規定される汎用リレーションに基づき、関与する言語別にそれぞれの動詞使用辞書をコンパイルすることは、充分熟考に値する。例えば、最近、中国の言語学者 Lin Xingguang教授が、現代の中国語に関する前述のような動詞辞書を出版したばかりである。同辞書にはおよそ3,000の動詞が収録され、個々の見出し語が語義、サブカテゴリー、引数構造および周辺の名詞句との意味関係が付された。このような動詞辞書を用いてコンピュータで解析する場合、与えられた動詞の語義に対して一貫した中間言語表現を決めれば、さらに有効である。

(3) 前項にあるように、究極の概念見出し語い、即ちかかわるすべての言語に止まらず、個々の概念見出しが起こった文脈をも網羅するものは、近い将来に使用可能になると期待することはできない。従って、１言語対ごとに、適切な２言語訳し分け辞書を選ぶ作業に取り組むほうが妥当性が高い。すでに、いくつかの単言語および２言語訳し分け辞書が出版または機械可読バージョンで出回っている。しかしながら、そのような辞書は、取り上げた言語及び基本にした言語もばらばらで、しかも多くの目標言語・訳語の自由な組み合わせを含み[3]、現在の問題を解くのにあまり有効ではないようである。従って、CICCプロジェクトの個々の言語ペアのためにそれぞれ２言語訳し分け辞書をコンパイルすることが重要である。さらに、それらの開発の最も重要な問題は、言語ペアの有益な目標言語・訳語の組み合わせを獲得する、新しい手段を開発することが求められる。

4. 結　論

MMT研究における中間言語方式にはいくつかの利点が認められるが、その根拠となる仮説は誤解を導きかねない。8年にわたる CICCプロジェクトの成果に基づき、適切な翻訳を得るには、たとえ中間言語方式 MMTシステムでさえも、数種類のトランスファメカニズムが不可欠だと強調したい。また、将来CICCプロジェクトを改善する方策を検討した。即ち、

第１に個々の分析モジュールに対する品詞と語義のための曖昧さ解消メカニズムの開発。第２にそれぞれの言語のための動詞使用辞書の編集。第３に、かかわる各言語対ごとに２言語訳し分け辞書の作成。以上３点を提案するものである。

［参考文献］

[1]『Interlingua』Ver.3、1992年3月、CICC発行。
[2]公開討論会「中間言語は多言語機械翻訳の救い主たりえるか」（『MMT'90会議プログラム及び講演原稿』（1990年11月5~6日）、東京、CICC発行。
[3]Benson、M.「訳し分け辞書の構造」、『International Journal of Lexicography』、(国際辞書学会報) Vol.2 No.1、1989年。

インドネシアにおける研究成果

インドネシア国立技術アセスメント適用省（BPP TEKNOLOGI）
人工知能オートメーション技術研究所
副主任　ダルマワン・スクマドジャジャ

日本のCICCとその周辺国の協調プログラムは、8年間にわたり実施された。多言語機械翻訳システムの開発をめざしたこのプログラムも、今年、システムの著しい改良をもって終了する。このリポートはインドネシアにおける中間言語方式による多言語機械翻訳システムの開発を要約する。インドネシア語電子化辞書、翻訳結果、およびサポートシステムについても論じたい。

1 インドネシア語マスタ辞書

機械翻訳システムの開発および他の自然言語処理システムでは辞書作りが必須である。このためインドネシア語マスタ辞書（IMD）を適宜作成した。IMDは一般語の辞書であり、それは、分析と生成処理をサポートするための、他の各種辞書の基本として使用される。IMDにおいて、個々の単語入力は以下の5

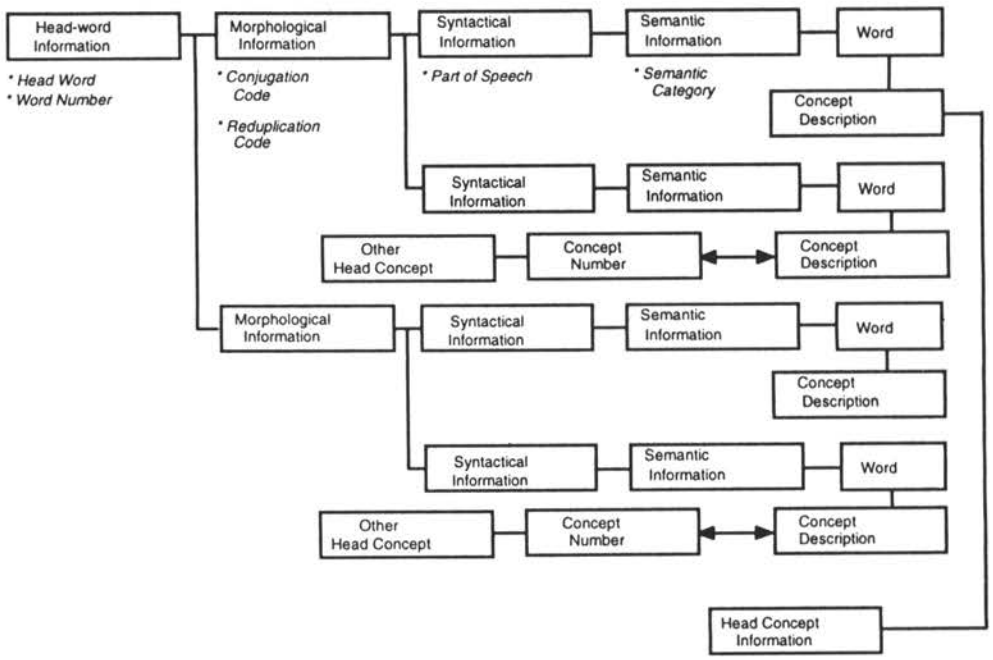


図 1 インドネシア語のマスタ辞書の階層構造

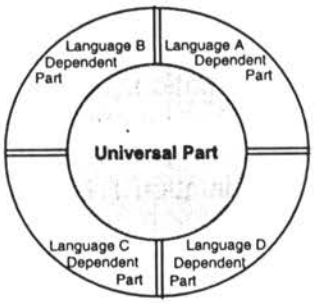


図2 中間言語方式の説明

つの重要な情報からなる。

1）見出し語情報
2）形態素情報
3）構文情報
4）意味情報　および
5)概念および中間言語対応情報。

これらの5種類の情報は、図1に例示するように階層構造を成形するために結合される。

一般に、見出し語は、ルートワードを形成する形態素である。この語は、文のその機能に依存する品詞に分類できる。品詞は様々な意味ももち、このように、それは他の4個の情報の形成リレーションの形成を阻害する。図において描かれるような階層構造が用いられる。その構造から一連の分析を行っている。すなわち、見出し語→形態素→構造→意味と解析を進め、語およびその概念を生成する。

見出し語情報は、ルート語、語番号、およびそのバリエーションからなる。ルート語は、語構成のベースとして利用される形態素である。実用性のために、慣用語もまた、ルート語とみなされる。語番号とは見出し語の指数をさす。バリエーションは、その派生という形であっても、その意味の変化しない語の修正である。

形態素情報は活用コードと再重複コードの項目を用いる。活用コードは次の3種類すなわち、能動態活用コード、受動態活用コード、および他の活用コードに分割される。

能動態および受動態活用コードはともに動詞に適用される一方、他の活用コードは他の品詞に適用される。再重複コードは、たいてい、名詞再重複がインドネシア語の複数フォームを表すように用いられる。

シンタックスは言語の重要な情報の一つである。構文情報は文中の語の詳細な品詞を含む。この細部情報によって自然言語システムは文を精密に示し、分析できる。

意味情報は語の意味的なカテゴリを示す。それは完全な品詞情報をもち、ポリセミ語の正確な意味を得るために有益である。ポリセミ語は、文のコンテキストに依存するいくつかの意味をもつ語である。

最後に、中間言語対応情報も重要である。この情報は、多言語翻訳システムにおいて種々の言語間の単語をリンクするために用いる概念記述と概念番号からなる。

2 中間言語

中間言語は多言語機械翻訳に理想的と認められたため、このプロジェクトに採用された。それ自体がミディアムもしくは普遍言語として作用するので、

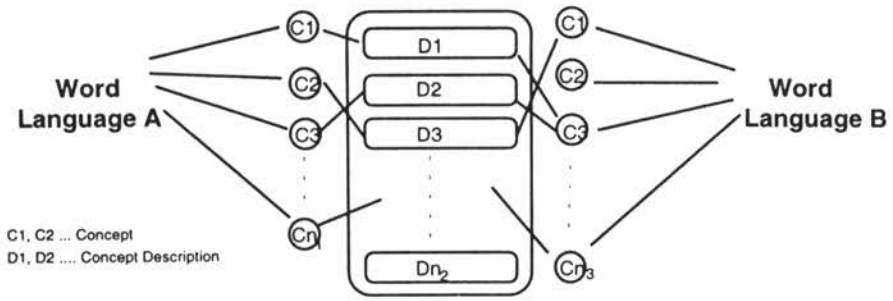
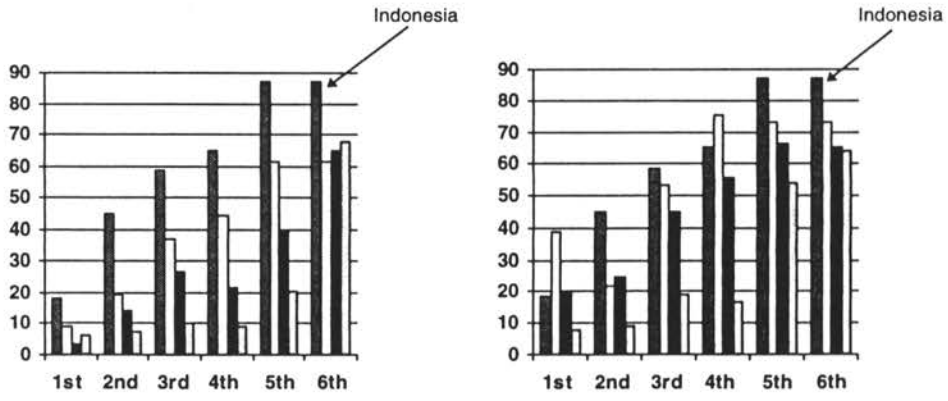


図 3 概念リンク



Indonesia to others Languages
(Analysis point of view)

From other Languages to Indonesian
(Generation point of view)

図 4 翻訳結果(出典:CICC)

中間言語はより望ましい。結果として、1国の1言語のための機械翻訳システムには、その言語の解析システムと生成システムのみを用意すれば実現可能である。

実用段階では、機械翻訳システムの中間言語の実行は簡単ではない。言語依存部分は図2に示されるが中間言語の実行中に起こる障害のうちの一つになる。言語依存部分とは、そのような概念がターゲット言語に存在しないために起こる、1言語から別の言語に翻訳するのが困難なことを説明する言葉である。これは言語がある文化の一部であり、また、まったく同じ文化は二つとないために引き起こされる。

中間言語の限界とは、他の言語の概念と関連させる問題を含む。解決策として、CICCにより提供された英語の概念記述は、MMTプロジェクト参加者の言語のリンクにおいて概念インタフェースとして利用される。概念リンクの複雑性は図3に見るとおりである。現在、リンクは個々の概念毎に手動で行っている。結果として、計画の進行には恐ろしく多くの時間と労力を必要とする。従って、この非能率性によって概念リンクの有効な方法の追求が行われた。

3 翻訳結果

リサーチに、中間言語のある方式が多言語機械翻訳システムの実行に取り入れられた。このため、翻

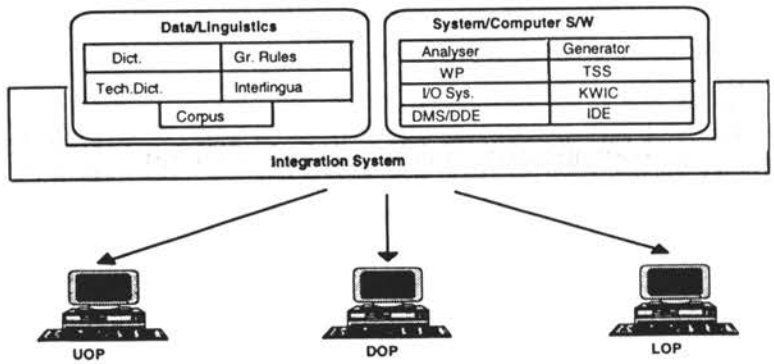


図 5 インテグレーションシステムの概要

訳の品質は中間言語生成の能力に依存する。プロジェクトにおいては各国家の言語の中間言語生成は、各国がコントロールした。例えば、インドネシアのBPP Teknologi はインドネシア語について責任を持った。結果として、インドネシアにおいてインドネシア語辞書、解析システム、および生成システムが開発された。その他の参加国すなわち中国、日本、マレーシア、およびタイも同じ責任をもち、従って翻訳の品質を決定するために、参加者間の協調が必須であった。

システムの性能評価について、3,225対のコーポラ文が各クオータごとに選ばれた。図4は2タイプの評価の翻訳結果を示す。すなわち、1) インドネシア語を他の言語に訳する性能評価、および2)他の言語からインドネシア語に訳する性能評価である。

4 インテグレーションシステム

機械翻訳システムそのものための開発に加え、いくつかのサポートシステムが開発された。

- 辞書保守システム（DMS）
- 辞書不選手エディタ（DDE）
- ワードプロセッサ（WP）
- I／Oシステム（IOS）
- 中間言語ディスプレイとエディタ（IDE）
- インテグレーションシステム

翻訳システムとサポートシステムの間のインテグレーションは 3つのそれぞれのパネルにおいて実行された。これらのパネルには、エンドユーザ、開発者、および研究者の3つのセグメントに対応した。

ユーザオペレーションパネル(UOP) はエンドユーザ専門であり、開発者オペレーションパネル（DOP）はシステムの継続的改善を目的に開発者を、また言語学的研究を目ざし、言語オペレーションパネル（LOP）が設けられた。それらのインテグレーションシステムの説明は 図5のとおりである。

5 結びとして

MMTS開発に関する日本のCICCとの協力プロジェクトがこの度、完成した。本プロジェクト完了によってもたらされた成果は、以下にメッセージの形で示したい。

1. プロジェクトの成果そのものが、より一層の研究開発へ発展させ得る有意義な基本概念を示した。
2. MMTS開発において現在上げられた成果に加えていくつかの要検討問題が指摘された。

――中間言語は中間表現に適する。しかし、言語依存部分にはまだ問題が認められる。中間言語をさらに機械翻訳へ適用させるためにも、言語依存部分の影響を減らすリサーチの実施が望まれる。

――概念リンクは中間言語方式において未解決の課題である。概念リンクの現在の技法に評価と修正が必要である。

――もう一つのR&Dとして、他言語へのよりよい翻訳性能を獲得するために、コオーディネーションまたは標準を修正するべきであろう。

謝 辞

プロジェクト実施に参加されたインドネシア、日本、およびその他の参加国の研究者に深謝したい。研究者間に育まれた協調関係がプロジェクト達成に非常に貢献したと信じるものである。

参考文献

- Interlingua Ver.3, March 1992, CICC.
（Interlingua Ver.3、1992年3月、CICC、日本）
- Indonesian Master Dictionary Technical Report, BPP Teknologi.
（インドネシア語マスタ辞書技術報告書、BPP Teknologi、インドネシア）
- Expression Method in Interlingua for Bahasa Indonesia, BPP Teknologi.
（Bahasaインドネシア語の中間言語方式の表記、BPP Teknologi）
- Research Leader meeting report documents.
（研究リーダー会議の報告書）

タイにおける言語処理研究の次の10年に臨んで

タイ国立電子工学コンピュータ技術センター
言語学知識科学研究所 所長
ウィラット・ソーンラートラムワーニッチ

本稿では、多言語機械翻訳システムの開発をサポートする辞書およびコーパスの構築について記述する。

中間言語方式は、翻訳結果の改善を妨げるいくつかの独特な問題を派生するため、多言語間の単語毎にマッチングをとるための1組の中間表現を定義する必要がある。問題は言語間の分岐と不対応にありここでは、生成プロセスにおける意味単位および表層形態の選択の定義に関する問題に注目する。

まず、最初の問題として、異なる言語に対する意味単位の範囲は、必ずしも同じである必要はない点が指摘される。同一の単語も、ある言語においては他の言語よりも広い語義をもつ場合がある。この場合、概念の合成と分解の処理を実現する、概念に関する一連の操作を提案したい。

第二の問題点として、ある文に対する中間言語は抽出すべき意味を表す点をあげる。表層情報は、概念、関係詞、または構造に翻訳される。従って、表層の適切な単語選択も、文生成における重要なタスクだと言える。文化的に、または日常生活でさまざまの言葉が使用される。タイ語の類別詞の使用の場合には、1個以上の候補語がある時に、ある単語がなぜ他の単語よりもより適格かを説明することは困難である。それゆえ、ここでは類別詞割当てへの統計的手法の適用を提案する。

1 辞　書

辞書は、中間言語的アプローチに基づく多言語機械翻訳システムをサポートするために作成される。単語利用の分野に基づいて、２種類の基本辞書が構築された。すなわち、基本語5万エントリのほかに専門用語（情報処理分野）2万5千個のエントリが行われた。解析と生成モジュールにおいて単語情報のほとんどが共有できるので、辞書は1個以上のインデックスタイプによるアクセスを行えるように設計された。文解析と生成モジュールに用いる基本辞書が単一である利点は、必要メモリが少ないことと、保守管理の容易さである。

生成モジュールは、解析モジュールとは逆で文を生成する。格マッピングにより構文および意味格のマッピングに用いる情報を組織的に提供する[3]。

品詞曖昧さ解消プロセスによって明確化された格を生成モジュールは構文格と一致する単語位置に割り当てる。単語情報を一度アップデートすると、解析と生成の両方に影響する。両側の規則を相互にチェックすることで、辞書の単語情報の整合性がいつも保たれる。

個々のレコードは、図1に例示するように情報を含むように定義される。

辞書は、機械翻訳で使用するためだけに設計されるのではなく、それをアップデートした状態に保ちかつ利用を拡大するよう、保守に関しても配慮されている。従って、単語のレコードは、タイ語の単語形態と概念 IDを提供する形態素情報(MORPH)により構成される。すべての語義はEDRの定義に従い、固有の概念IDと記述を与えられる[6]。概念IDは、各言語の単語形態をリンクする語義の定義を提供する。構

MORPH	::= <thai>,<concept-ID>,<english>
SYNTAX	::= <word-type>,<pos>,<vp>,<classifier>, <case-mapping>
SEMANTIC	::= <ako>,<similarity>
OPERATION	::= <constraintp>,<constraintc>,<prune>, <pruneall>,<replace>

図１単語情報

<constraintp>	: constraint on parent concept
<constraintc>	: constraint on child concept
<prune>	: drop the determined child branch
<pruneall>	: drop all the child branch
<replace>	: compose with the determined child concept

図 2 概念オペレータ

文情報(SYNTAX)は、<pos>(品詞) 、<vp>(動詞の動詞パターン)、<classifier> (名詞の共起類別詞)と、<case-mapping>を含んでいる。単語の意味の意味的関係は概念の階層において定義され、<ako>(a-kind-of)となる[1]。

2 概念のオペレーション

一般に翻訳の品質は、以下の主要な機能を実行することによって一般に高められる可能性がある。[5]
すなわち、

1. 単語の訳語の選択
2. 単語の再配列、および
3. 文スタイルの改善

文法規則の改善のほかに、単語のリンクは、翻訳結果を改善するもう一つの有効なテーマである。中間言語的アプローチにより、1つの文に対して様々な表現が認められる。いつも、ある言語に対する意味表現の単位が必ずしも別のものと同じではないので、これは理論上真実である。定義された概念は基本の概念である必要はない。ある概念は 1個以上の概念的単位から構成される可能性がある。概念のこの特性により、ある言語の特定の単語に対する概念の単位が、他言語の単語より大きい場合がある。

E (英語) : It rains.(雨が降る。)
T (タイ語) : /fon/ /tok/
(rain) (drop)
(雨) (落下)

E : to whiten st.(stを漂白するように)
T : /tham/ /hai/ st. /kwau/
(cause) (white)
(起こす) (白)

いくつかの複合概念が文構造を複雑にした。原文の構造を目的とする言語と一致する概念を含む構造に変えるために、再構造するオペレータが必要である。概念合成と分解のためのオペレータは次の通り準備される。

親と子供の概念が構造を制約する。構造内のある概念はその親概念と、または子供概念によって指定される。英語において、"whiten"とは「より白くなるようにする」を意味するが、タイ語においては、違った概念構造において表現する必要がある。"whiten"(漂白)の概念が分解され、"cause"(起こす)および"white"(白)となる。

E:Toothpaste whitens the teeth.
(c#toothpaste←agt-c#whiten-obj→c#tooth)
(ねり歯みがきは歯を漂白する。)

```
@entry_type{key,
        <Required fields>
                field_name = {field text}, field_name = {field text}, .....
        <Optional fields>
                field_name = {field text}, field_name = {field text}, .....
        <Ignored fields>
                field_name = {field text}, field_name = {field text}, .....
        }
```

図 3 エントリタイプの一般的シンタックス

(คณะกรรมการ_111, คณะ_2, 11)	(ทหาร_111, นาย_1, 9)
(คณะกรรมการ_111, กลุ่ม_2, 5)	(ทหาร_111, ฝ่าย_2, 1)
(คณะกรรมการ_111, คน_1, 6)	(คนงาน_111, คน_1, 6)
(นก_13111, ตัว_1, 9)	(ส้ม_13114, ลูก_1, 12)
(นก_13111, ฝูง_2, 4)	(ส้ม_13114, ผล_1, 3)
(ไก่_13111, ตัว_1, 10)	(แตงโม_13114, ลูก_1, 8)
(ไก่_13111, เล้า_2, 3)	(ทุเรียน_13114, ลูก_1, 9)
(นกกระจอก_13111, ตัว_1, 7)	(โค_13111, ตัว_1, 7)
(คน_111, คน_1, 67)	(หมา_13111, ตัว_1, 13)
(คน_111, กลุ่ม_2, 1)	(หมู_13111, ตัว_1, 5)
(ทหาร_111, คน_1, 17)	(ช้าง_13111, เชือก_1, 3)

図 4 類別詞集合(NCA)

T:Toothpaste causes the teeth to be white.
(c#toothpaste←agt-c#cause-obj→
(c#white-a-obj→c#tooth))
(ねり歯みがきは歯を白くさせる。)

この場合、c#whitenは<replace>(置換)オペレーションによって(c#cause-obj→c#white)と置換される。その時、構造はc#whiteの格必要条件に従って再構築される。

3 タグ付きコーパスベース

タイ語の研究に使用できる電子化テキストデータがほとんどない。多くの研究所は特定の使用のために自らテキストベースを作っている。そのようなテキストのデータベースは非常に小さく、かつ情報不足であって、システム評価および統計的研究に用いる、十分に大きなデータベースを得るためにも、収集したタイ語のテキストデータを加工し、言語研究に役立つ付加的な情報を蓄える構造を与えた。テキストの構造は現在はローカルに定義されるが、将来ある標準にしたがって拡大される見込みだ。

4 テキストベース構造

テキストデータを、記事、本、論文集などの15のカテゴリに分類した。データのフィールド情報を現在のテキストベースに入れる必要がないよう、データの分野は情報科学に制限した。文献データは、図3において示されるシンタックスに従って付与された。

確率品詞タグをテキストデータに与えた。一般のタイ語テキストでは、単語境界だけでなく文の範囲を示す明白なマーカが全然なく、単語は連続して書かれる。単語の区切り、または句の並び、文節または文の境を示すために、時々、スペースが用いられる場合がある。これによって、いくらか読取りが楽になることもある。

従って、単語分割処理によってテキストを前処理する必要がある。単語分割プログラムは、キャラクタ結合規則にも合うような、最も長く極少な単語カウントの単語を発見する規則を使用する[2]。

正解率は、テキストの8,428単語を用いた試験の結果に基づけば、ほぼ95.91%と高い。品詞のタグ付け

Semantic class	Unit classifier	Collective classifier
animal	ตัว_1	ฝูง_2
human	คน_1	คณะ_2
plant	ต้น_1	-
fruit	ลูก_1	-

図 5 代表的な類別詞のためのNCA

は品詞選択プログラムによって半自動的に行われる。品詞は辞書から選ばれる。

5 名詞類別詞の訳し分けの抽出

タイ語では、類別詞にいくつかの利用法がある。類別詞は、例えば序数、代名詞を表現するために、名詞による構造について重要な役割を果たす。特定のパターンに従って類別詞句が参考表現(N−CL-DET)、名詞修正表現 (CL−N)などによって構文的に生成される[4]。

図4に例示したとおり、名詞と類別詞を抽出し、訳し分け類別詞集合 (NCA)と呼ばれるような表にまとめた。

上記の集合は、与えられた名詞の適当な類別詞を決定するのに有益である。コーパスに存在する名詞のための選択肢は、直接的に、関連する代表的な類別詞 (他のどの類別詞よりも頻繁にコーパスに存在する) を使って決定される。また与えられた名詞がコーパスに存在しない場合、選択肢は、概念階層における格の代表的な類別詞を使って決定される。

例(1)および例(3)では、コーパスにおける名詞の格が示される一方 、(2) と(4)では、シナリオが異なる。(2)において/appern/の単位類別詞は格 'fruit' (果実) の代表的な単位類別詞を使って得られるが、図 5に従えば*** /luuk/である。同様に(4)の単語/gangken/ の集団類別詞は、格 'animal' (動物)の代表的な集合的類別詞において、*** /fuung/と決定される。

6 結 論

辞書とテキストデータベースも、重要な資産である。タイ語の多くの言語現象はまだ不明瞭である。多くの問題は答えのないまま放置されている。例えば、何を単語の単位と規定するか、何を文とするか、タイ語文に句読点を使用するべきか、などである。これらの規定がされないまま、タイ語を電気的に処理する場合、あるいは人と人のコミュニケーション、特に技術的な分野または高いエンドリサーチにおいて、非常に問題である。実用的な言語データの研究に起因する、新しい文法規則が必要であろう。大規模データベースに基づく確率論的な方式は、タイ語の言語処理の新しい時代の幕開けを告げる、解決策の一つである。そのためにも、データベース拡充と、各々データを最もアクセス可能な方式に整備することを計画している。

本稿において、適切な類別詞を割り当てるためにNCAを使用する有効性を測った。これは、大規模言語データベースの貴重な運用法のほんの一例である。

謝 辞

計画のすべての段階でリサーチを円滑に進められたことは、ひとえにタイ国立電子・コンピュータ技術センター(NECTEC)、および日本の国際情報化協力センター(CICC)のお蔭と感謝申し上げたい。

本プロジェクトに参加された多くの研究者および諸氏がご所属のチュラロンコーン大学、カセサート大学、キングムングット研究所にはたいへん寄与していただいた。もしLINKS研究所の同僚から、最大限および無限のサポートが受けられなければ、本プロジェクトの成功は危ぶまれたことから、諸氏に感謝したい。

参考文献

[1]:村木一至、Sornlertlamvanich, V. 他著(1989年)、「多言語機械翻訳システムのためのタイ語辞書」、Computer Processing of Asian Languages(CPAL):211-220頁。

[2]:Sornlertlamvanich、V(1993)、「機械翻訳システムのためのタイ語の単語分割」、Machine Translation国立電子コンピュータ技術センター(タイ語)

[3]:Sornlertlamvanich、V.、Phantachat、W(1993)、「タイ語の情報ベース言語解析」、Asean Journal on Science & Technology for Development:181-196頁。

[4]:Sornlertlamvanich. V. Phantachat. W. Weknavin S(1994)、「コーパスベースアプローチによる類別詞割当て」、COLING 94論文集:556-561頁。

[5]:堤泰治郎、(1990)、「機械翻訳における中間表現の広範囲再構造」、Computational Linguistics Vol. 16、No. 2:79-78頁。

[6]:安原均、(1993)、「テキストコンパイラおよび概念タグ付きコーパス」、KB&KS93論文集:251-256頁。

遠隔勤務環境におけるMMT

マレーシア国立翻訳協会
主任技師　モハマド・ザイニ・ユヌス

マレーシア国立翻訳協会（ITNM）の技術開発部門は、研究開発、技術移転および情報検索のための機関であり、情報技術を用いて、翻訳の一括処理機能をもつ機関として遠隔勤務の環境を整備してきた。この種の就労形態の概念とは、世界中に散らばった内部および外部の翻訳スタッフも顧客も、ファックスとコンピュータネットワークを経てITNMと関係を保つというものである。また、ITNMの翻訳者のほうも、多言語機械翻訳システム（MMT)や翻訳支援ツールの利用を激励される。これらのデバイスによって、翻訳プロセスが、より効率的にずっと短い時間で実行するができるからである。MMTは、ITNMで開発されているツールのうちの1つであり、また将来、遠隔勤務環境において使用されるであろう。本稿では、ITNMにより提供された遠隔勤務環境において開発されるMMTの使用を詳述したい。

1　技術開発部門

この技術開発部門として、以下にあげるようなITNMの設立趣意に沿った活動方針を掲げる。即ち、
(i) 多言語を用いた翻訳、解釈、および情報検索に付随する国立の研究開発センターとして役立つ。
(ii)国立の技術移転センターとして、芸術、科学および技術の分野における技術移転計画に対し、専門的な支援と相談を提供する。これらのプログラムは、多言語を用いた翻訳、通訳および情報検索と関係があり、また、マレーシア政府または他の適切な組織により、地元及び国外機関へ技術移転を行う。
(iii) この分野の国際的、多言語的情報検索センターの国立の受け皿として役立つ。

これらの活動方針は、質の高いインフラストラクチャー、人的資源、技術、およびサービスの創出および開発、実行、管理、および提供において、マレーシアの発展を妨げる言葉の壁を取り除くために寄与している。これらはいずれも、国内および国際的なレベルにおいて多言語を用いた翻訳、通訳、および情報検索に付随する。

提供されたサービスとは、具体的には：
(i)研究および技術にかかわる施設を提供し、高い品質の製品や研究成果およびシステムの発展を創成するために、継続的な研究開発活動と共同研究を行わせる。
(ii) 情報と研究成果を発信する。
(iii) 専門的相談、訓練、職業的および技術的なサービス、および技術移転を提供する。

2　多言語機械翻訳システム(MMTS)

MMTシステムにおいて、コンピュータは翻訳者の役割を果たす。しかし、人は、コンピュータがよりよく翻訳することを補助する重要な役割をまだ果たしている。さらに、人間の翻訳者はコンピュータに手渡される前に、テキストを前編集する。また、人は翻訳プロセスの終了後、テキストを後編集する。MMTシステムはまた、完全な自動翻訳システムとして使用され、コンピュータは、人的な介入を全く必要とせずにテキスト全体を翻訳する。

前述の2つのシステムはいずれも、性質上ルーチンである情報や、さらに、ニュース速報や形式的な書簡、証券取引および外国為替または商品取引などの特定され限定された分野の情報に関して、翻訳者に特に非常に有益である。

MMTSは実用のためのいくつかのオプションモデルをもっており、大きく3つののカテゴリに分割できる。独立操作可能なモデル、分配されたモデル、およびネットワークモデルである。

MMTプロジェクトは未だ研究開発段階であり、プラットホームとして本計画のコーディネータである日本とともに、参加国間の協力で進行している。ITNMは、MMT研究開発を促進するために、主要な役割を果たす一方、独自に研究開発を続けており、また国内および国際的なレベルにおいてシステムを十分に利

用している。

3 遠隔勤務の環境

ITNMの翻訳者とは、常勤の内部スタッフだけではなく、全国各地にいる契約あるいはパートタイムのスタッフから校正される。地理的な距離に打ち勝つために、ITNMは業務に遠隔勤務の概念を利用し、情報技術を大幅に取り入れている。遠隔勤務とは、ITNMの登録翻訳者が全国至る所または世界の他の地域に散在おり、ファックスとコンピュータネットワークを通じてITNMと結ばれている。つまり個々の登録翻訳者は、コンピュータシステム、電話回線、モデム、必要なネットワーキング用ソフトウェア、電話、およびファックスを装備しなければならない。

ITNMは、宅配便やファックス、またはコンピュータネットワークによって、顧客から翻訳依頼原稿または文書を受け取ると、まずその翻訳必要条件を決定するために迅速な判断を行う。いったん適当な翻訳者または翻訳者グループが翻訳者データベースから選ばれたら、ITNMは、翻訳させる依頼原稿を選ばれた翻訳者または翻訳者グループに送る。またより短いテキストの場合、ソーステキスト（原文）も翻訳者に直接ファックスまたは電子メイルによって送信する。

翻訳が完成すると、翻訳者は訳文をコンピュータネットワークを経たITNMに送信し後編集を受ける。その後、訳文はDTPシステム上でレイアウト作業に回される。ITNMは最終の植字あるいは印刷工程を行わずに、しない。品質の信頼できる外部の印刷会社へ業務委託する。印刷の終わった文書は、ITNMの一環作業の概念に則り、顧客に納入する。

ビジネスセクタの顧客から、時間が勝負の翻訳を特急で求められた場合、ITNMはエラー削減および経費節減だけでなく時間短縮についてもあらゆる手段を追求するが、ここでもITNMの遠隔勤務の利用は、正規の郵便システムまたは宅配便サービスを利用する従来の方法と比べても、翻訳にかかわる輸送費軽減と納期短縮を実現する。また、キーボード入力の間違いや校正にかかる時間をかなり減らすことができた。従来は、訳文を翻訳者から受け取った後に、わざわざ再入力していたのである。遠隔勤務によって、翻訳者およびエディタが、相互に電子メイルサービスを通じて効果的に通信可能になった。

4 インプリメンテーション

遠隔勤務の概念は、すでに、ITNMがその顧客への効率的なサービスを提供するために実用されてきた。コンピュータネットワークを通じて接続できる多言語を用いた翻訳者の集団を登録し、クライアントサーバ環境のバックグラウンドにおいてヒューレット・パッカード社HP9000G50サーバを用いて、ネットワークサーバ、通信サーバ、およびデータベースサーバとして作動させている。さらにまた、ITNMのためのゲートウェイとして動作している。

フォアグラウンドでは、以下のエキュゼクティブ情報システムが部分的に運用されている。すなわち、人的資源マネジメントシステム、金融情報システム、市場情報システム、翻訳情報システム、ライブラリ情報、宣伝活動情報システム、および運用監視システムである。X-D EISソフトウェア、Paradoxおよびオラクル7リレーショナルデータベースと四次元データベースシステムなどの様々なソフトウェアを利用している。　これらの情報システムは、ITNM職員に重要な情報を提供するため、意志決定のため、あるいは距離を問わずそれらの仕事を効率的にこなすために、利用される。例えば、翻訳情報システムはITNMについて登録トランスレータの情報や、多国語辞書データバンクを求める機能、用語データバンク、スペルチェッカ、翻訳作業のための文法チェッカおよび常用句データバンクシステムを提供する。これらを使用して、翻訳作業のコーディネータは、潜在的な翻訳者を迅速に見つけることができる。翻訳者にとってのシステムは、翻訳タスクを補助するために有益な情報が見つかるツールである。ITNMは、統合された翻訳環境の翻訳システムだけでなく人的な翻訳者、およびツールを利用するニーズがある。

対内的、対外的に様々なコンピュータプラットホームを結ぶ複合電子メールを維持管理するために、メールハブがインストールされる。メールハブは、統合電子メッセージシステムの役割を果たす場合、LANに基づく電子メールシステムの不整合を解く。

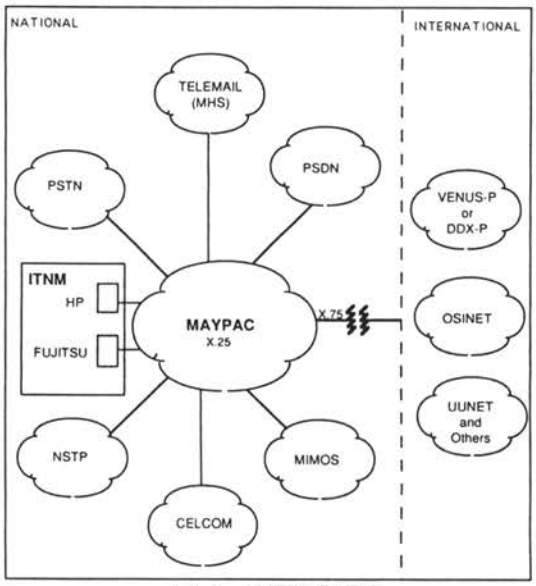


図 1　遠隔操作環境

電子メールおよびメール利用の可能なアプリケーションをサポートするインフラストラクチャを整えるものだ。

さらに、それはファックス能力も統合できる。大衆あるいはどの研究開発機関のメンバーも、“特定のユーザ”と呼ばれるが、彼らもMMTS計画の中で、彼らはシステムへのアクセスができる。さらに内部使用以外にも、システムを改善するために作動させ、十分に利用することができる。この場合、ITNMはMMTSを遠隔勤務環境に適合させる必要がある。MMTSはローカルエリアネットワークまたは広域ネットワークからホストまたはネットワーク環境を通じてアクセスできる。内部では、アップルマッキントッシュとIBM PCとともに互換のパーソナルコンピュータとSUNワークステーションを接続するLANが、MMTシステムのための富士通ワークステーションと東芝ワークステーションも接続する。知的なハブは、MMTシステムのイーサーネットバックボーンをコンピュータの残りと接続する。TCP/IPプロトコルを利用することによって、コンピュータのうちのすべてはUNIXとMMTシステムへのアクセスを特に得ることができる。内部のユーザとはコンピュータプラットホーム上のすべての翻訳者または研究者を含んでおり、彼らはMMTシステムを最も最適の方法で使用している。

ローカルエリアネットワーク以外に広域ネットワークも利用されている。電子メイルによって外部からITNMにアクセスするには、ダイアル呼び出しシステムが現在、JARINGを経たマレーシア・マイクロエレクトロニクスシステム協会（MIMOS)から提供される。もうすぐMIMOSが専用回線を利用している。JALINGとは、UUNET(USA)と MGVAX（オランダ）とMUNARI（オーストラリア）を接続するグローバルなインターネットの一部である。開発中のMMTシステムは、この国立のデータ流通インフラストラクチャの部分に当たる。さらに、ネットワークは、教育的および研究活動を主にサポートし、促進すると見込まれる。

情報提供業と情報レシーバの一員であるITNMは、地元の新聞社である New Straits Times Press社からオンラインのライブラリ情報システム（LOL)に参加している。現在、NSTPは技術を含むほとんどすべて分野の情報を提供し、リサーチ会社および私的な教育的な研究機関を含む幅の広い層から加入者がある。MMTシステムは、部門特定のユーザに利用可能な技術分野のシステムのうちの1つである。

また、ゲートウェイとしてITNMを東京の富士通ワークステーションと直接に接続するために、専用回

線がインストールされた。東京の国際情報化協力センター（CICC）の機械翻訳研究所、タイのバンコク、インドネシアのジャカルタおよび中国の北京へ通じる他のコンピュータリンク、また計画されている。MMTSモデルに限らず、この遠隔勤務環境は、参加諸国のすべてで計画されて、達成される。さらに、この環境は、MMTシステムの改良と未来を保証するために、メンバーの間で支えられるべきである。

別の専用回線はITNMとマレーシアの一番上のセルラー通信会社のうちの 1つであるCELCOMを連結している。その会社はセルラー電話機能だけでなく、データ通信も提供し、セルラーモデムを使用する。このモデムは無線機能をもち、それが遠隔勤務環境について非常に役立っている。この技術の最も極端な使用法のうちの 1つは、クライアントがITNMに送信した至急翻訳の業務である。MMTSのユーザは、通信インフラストラクチャの整備が遅れた遠隔地からこの機能を利用できる。

図 1 はITNMの遠隔勤務環境を説明する。

5 結 論

ITNMは、マレーシアの翻訳産業を形成するために、訳者の地位向上を目指す使命をもっている。翻訳作業を補助する生産システムとツールは、確かに、ITNMの初期の目標達成を助ける。それは、情報のマレーシアとのスムーズな受発信を妨げる言葉の壁を取り除くことができるだろう。

MMTSは、翻訳プロセスを高速化し翻訳者を助けるツールのうちの1つであり、かつ他のITNM目的を達成することを助ける。そのため、促進および共同のR&DによるMMTSの普及以外に、他のプロジェクト部門、特に、通信関連技術はITNMの遠隔勤務環境におけるMMTS関連の部分の閑静に応用されるだろう。

参 考 文 献

Ahmad Zaki Abu Bakar博士"翻訳専門家および非常勤翻訳家のための翻訳ツール"言語学の応用に関する国際会議の予稿集、USM、マレーシアペナン島、1994年7月. 206-210頁。

ザイニ・B,ユニス"マレーシアのMMT/OSIシステム" OSI報告書、ＩＮＴＡＰ,日本,1994年3月、102-106頁。

「TMI'95」

（第５回機械翻訳の理論的、実践的諸問題に関する国際会議）

開催月日　平成７年７月５日～７日
開催場所　University of Leuven, Belgium

論文募集　テーマ　－ＭＴとコンピュータ意味論　　－ＭＴと音声言語
　　－サブランゲージ　　－制限言語とＭＴ
　　期　　限　－1995年１月１５日（コピー６部提出）
　　審査完了　－1995年４月１５日
　　使用言語　－英　語

参 加 料　1995年5月15日まで申し込み--
　　6,500 BF(ベルギーフラン）バンケット込　8,000 BF
　　5月16日以降申し込み--
　　10,000 BF(ベルギーフラン）バンケット込 11,500 BF

問　合　先 University of Leuven, Center for Computational Linguistic
　　Maria-Theresiastraat 21, B-3000, Leuben, BELGIUM
　　(Tel.+32-16285088　E-Mail tmi95@et.kuleuven.ac.be)

MTって，知らなかった
使ってみて欲しくなった

商店街ではちらほらクリスマスのイルミネーションが灯り出した11月25日、名古屋商工会議所と共催で機械翻訳実務講習会を開催した。会議所国際部の管轄で、会員の中から国際的業務を担当している企業1,500社に対し実務講習会の開催案内を送付した。

会議所としても経済セミナー等は常時開催してはいるものゝ実務講習会は初めてとのこと、まして当日は給料日、しかも花の金曜日とあってどの程度の参加が得られるか心配されたが、それでも110人の人々が会場につめかけた。

講習第１部は機械翻訳の種類、方式、構造、機能などの“概要説明”から始じめた。今夏以来機械翻訳はハード、ソフトとも大きな変革をみせている。ＰＣ翻訳ソフトは従来品の半額以下へと低価格化が進み、中には1万円を切ったものまで発売されているし、ハードもWindowsの採用によって、機器の操作性は抜群に向上し、ゲーム感覚で使用できるようになった。またユーザ利用情報のメーカへの積極的なフィードバックによって、年に2~3回もバージョンアップが繰り返され、翻訳品質も著しく改善、“機械翻訳は使えない”という言葉は最早神話になりつゝある。

講習第２部はW/S、PC各2社による翻訳実務デモンストレーション、高速大量翻訳、豊富な翻訳支援機能など各社商品の最も得意とする機能を中心にインストラクターが模範作業を実施した。

講習第３部はユーザ参加による実習、会場の広さ、電源容量の関係から20台の機材しか準備出来なかったが、参加者は交替で翻訳実務に取り組んだ。時間的制約もあり、初期設定の仕方や、長文分割、未知語登録、用語選択、品詞指定など前編集項目を盛り込んだ実習例文をFDに各社毎に準備し、これをベースにして親切な指導がおこなわれた。パソコンなんか今まで触ったこともないという50代の男性も、インストラクターの指導で翻訳できて、思わずにっこりという光景もみられた。

総時間数は僅か３時間という強行スケジュールで心行くまで機械に慣れ親しんでもらうまでにはいかなかったが、途中退席者は殆どなく、その熱心さに参画メーカ側ももっと早い時期に実施すればよかったとの声が出された。

講習参加者の声をアンケート(回収35枚)から拾ってみよう。

「講習案内を貰って初めて機械翻訳というものを知った」人が10人（28.6%）「商品名だけは知っていたが、内容までは知らなかった」が7人(20%)であり、まだまだ知名度が低いのが実態である。

業務上での翻訳の必要性については「大量にある」8人、「少量多種ある」が6人、「海外情報を収集したいものが多量にある」が7人、「今後翻訳需要が増える」が9人となっており、あらゆる分野において翻訳需要の拡大がうかがえる。

現在翻訳業務は「翻訳業者に依存」が7人、「社内専門部門で処理」が13人、専門外社員が自分で翻訳処理が15人となっている。

実習後の感想については講習時間が短すぎるという声はあったものの、機械翻訳は「使える、慣れれば使いこなせる」との声ばかりで、「難しい、使えこなせない」という人はいなかった。また名古屋地区での講習会や機械翻訳ショールームの所在を尋ねる人も少なくなかった。

個人ユースとしてはまだまだ高価ですネという声はあるにせよ「早速購入使用したい」が6人、「早急に検討したい」が6人、「将来使いたい」が18人であり、「必要性を感じない」という人は皆無であった。

参加者は製造業(9社)貿易業(5)商社(2)情報処理業(5)公務員(2)翻訳業(2)など多彩。職階は経営職(2人)管理職(4)専門職(11)業務職(10)その他である。

職種は海外管理(5)特許(1)ソフト開発(5)秘書(1)事務(3)翻訳(2)教官(1)営業(4)その他となっている。

年代別には60代(1)50代(4)40代(3)30代(13)20代が(14)である。(無回答は除外)

この講習会開催に際しては講習カリキュラム検討など参画メーカ4社で3回にわたり、綿密な事前調整会議を実施した事が成功に繋がったと確信している。

これを機会に各方面にわたって普及啓蒙活動を展開して行きたいと念願している。

技術早分かり

自動通訳システムと音声認識技術

日本電気(株) 情報メディア研究所　坂井　信輔

1. はじめに

今回は、自動通訳システムと、その中で用いられる音声認識技術の話をいたします。みなさんにおなじみの機械翻訳システムは、普通、ある言語のテキストを入力したら、それを翻訳し、別の言語のテキストを出力するというものですが、自動通訳システムは、ある言語の音声を入力したら、それを翻訳(通訳)し、別の言語の音声で出力するシステムで、音声翻訳システムとも呼ばれています。

交通機関の発達により、世界はますます狭くなり、他の言語を話す人々とのコミュニケーションを必要とする機会はますます増えています。マニュアル、ニュース、技術文献などのテキスト情報の大量の翻訳のニーズを解決するためには、機械翻訳技術が用いられていくことでしょう。一方、異なる言葉を話す人と人とのコミュニケーションを可能にするには通訳が必要となりますが、通訳サービスを利用できる機会は、コスト的にも大変限られています。私たちが仕事や余暇で外国に出かける時は、言葉が通じないために、自分の意志を伝えられずに困ることはよくありますね。いちかばちかでメニューを選び、変なものを注文してしまったというような失敗談も良く聞きます。通訳の機能を自動的に行なうシステムが容易に利用できるようになれば、異なる言語を話す人々同士のコミュニケーションは飛躍的に容易になり、また、コミュニケートする機会も増えていくことでしょう。

自動通訳のサービス形態は、大きく分けて、

(1) 遠隔地間の対話の通訳
(2) その場に居合わせている人と人の間の通訳システム

に分かれます。(1)は、たとえば自動翻訳(通訳)電話サービスとなります。これが実現すると、たとえば、ニューヨークのコンサート情報を東京から電話で確認して、チケットの予約を入れるといったことが可能になるでしょう。また(2)は、レストラン、ホテル、旅行案内所などいろいろな場所における通訳サービスに用いられるでしょう。これらについてはこれまでに研究システムが試作されています[1][2]。

さらに、ハードウェアの技術の進歩にともなって、携帯用の自動通訳機も登場するかもしれません。さて、この自動通訳システムは、一体どういう技術から構成されるのでしょうか?

2. 自動通訳システムの構成

人間の話す音声を聞きとってテキストや意味表現に変換する技術は音声認識と呼ばれています。またテキストや発音記号などから人間の声を生成する技術は音声合成と呼ばれています。機械による音声から音声への翻訳である自動通訳を行なうためには、機械翻訳技術に加えて、この音声認識・合成の技術が必要です。

図1は、日本語と英語の間の通訳を例にとって、自動通訳システムの構成を表したものです。図の左半分が日本語から英語への通訳を受け持ちます。図1(a)は、日本語の音声による発話文の入力を、文字テキストなどに変換する日本語音声認識部です。(b)の日本語意味理解部は、音声認識部の出力を解析して発話の意味を抽出します。(c) の英語文生成部は、この抽出された意味表現から英語テキストを生成し、英語音声合成部(d)は、これを人の声に変換して、英語話者に対して出力します。(e)～(h)のモジュールは、同様の処理を、英語から日本語に向かって行ないます。図1の真中の4つのモジュール(b), (c), (f), (g)において、機械翻訳の技術が用いられています。

自動通訳システムの構成を眺めてみると、人間の通訳が行なっている仕事が、どれだけたくさんの複

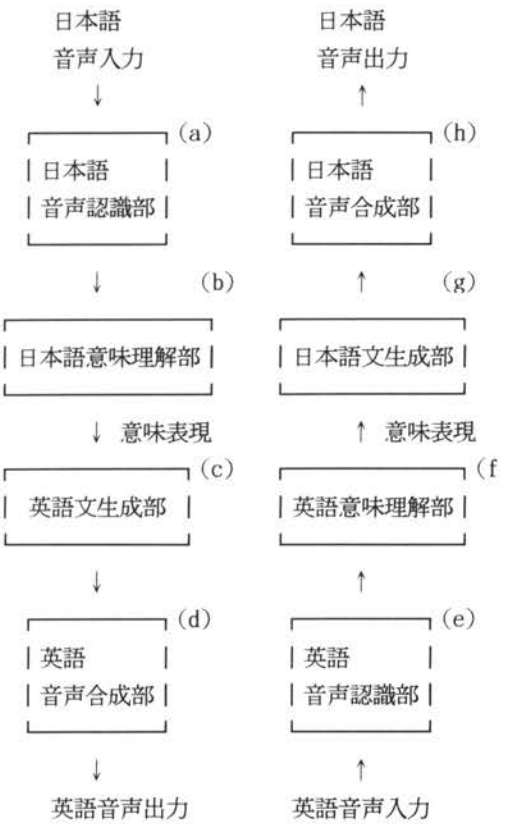


図1 日英自動通訳システムの構成

雑な処理からなるのかということに改めて気がつきますね。さて、今回はこれからこの自動通訳システムの要素技術となる、音声認識について見ていくことにします。

3 音声認識

自動通訳を構成する要素技術は、音声認識・機械翻訳・音声合成の3つに大きく分けることができますが、音声認識は、この中でも高性能化、大規模化が著しく難しく、また計算量が大きいという特徴があります。音声認識技術を開発あるいは利用する場合以下に述べる観点から、どういう仕様にするか選択する必要があります。

(1) 話者の限定 --- 特定話者 / 不特定話者

一般に、ある特定の話者の音声認識にくらべて、不特定の話者の音声の認識は、男女差、年齢、声道の形や喋り方の個人差などのバリエーションがあるために、より難しくなります。不特定の話者の認識を可能にするためには、何百人もの話者の音声データから、認識単位の音響モデルを学習するということがしばしば行なわれています。また、不特定話者のために学習されたシステムを、特定の話者の音声を精度良く認識するようにチューニングする、話者学習(図2(c))も、大変重要な技術です。

自動通訳システムでの利用形態を考えると、公共のシステムでも、またパーソナルなシステムの場合も、少なくとも片方の話者はあらかじめ特定できないことが多いので、不特定話者の認識技術や、短時間で話者学習する技術は大変重要なものとなります。

(2) 入力の単位 --- 離散 / 連続

また、単語単位の音声を認識対象とする離散単語認識と、複数の単語からなる音声を対象とする連続音声認識では、後者の連続認識の方が、認識性能の面でも、また計算量の面でも難しい問題となります。連続音声認識を行なう場合は、通常、入力音声の中で複数の単語がどのようにつながり得るのかということを制約として示す「文法」が用いられます。

ここでまた自動通訳システムでの利用を考えますと、話し言葉を一つ一つの単語に区切って話すのは、人間にとってかなり困難なので、連続認識の技術が利用できることが大変望ましいといえます。

(3) 認識語彙 --- 小語彙 ～ 大語彙

認識語彙、つまり単語辞書の大きさは、音声認識システムの精度と処理速度に大きな影響を与えます。一般に大語彙になればなるほど、発音の良く似た単語が増えるので、認識性能は低下し、また認識処理量が増えるので、システムの応答は遅くなります。現在、製品化されている連続音声認識システムには、動作するハードウェアに応じて異なりますが、数百～千単語程度の語彙をリアルタイムで認識するものがあります。

これまでに開発されている自動通訳システムのプロトタイプの一例においては、チケット予約をタスクとして、500語程度の語彙でリアルタイム認識を行ないます[1]。

認識に必要とされる語彙の大きさは、タスクにおいて用いられる固有名や、専門語がどのくらいたく

さんあるかによって異なってくると思われます。

(4) タスクと文法的制約

連続音声認識の入力は音声波形を分析して得られる特徴ベクトル（パターン）の時系列であり、あらかじめ /n/, /i/ などの子音や母音の境目やあるいは単語と単語の間ではっきりした切れ目が入っているわけではなく、なだらかに変化しています。このような入力を認識するために、音声認識システムでは、ふつう、テキスト入力の機械翻訳システムや、自然言語インタフェースなどの場合よりも、ずっと制約の強い文法が用いられます。音声認識においては、このような、認識性能を強力にするための文法的制約のことをしばしば「言語モデル」と呼びます。これらには、たとえば有限状態ネットワーク文法、子音や母音などの音素を最小単位とする文脈自由書き換え規則による文法などがあります。また、最近は連続音声認識の言語モデルとして、もっと統計的なアプローチである、N グラム文法がしばしば用いられています。この場合、書き換え規則や単語をアークとするネットワークで文法をあらわすのではなく、ある単語 W(i) が、入力文中で、その前のN - 1個の単語 W(i-N+1) … W(i-1) の後に出現する確率をすべての単語の組合せについて集めたものが文法知識となります。

(5) 音声認識システムの構成

図２に、音声認識システムの典型的な構成を示します。入力された音声は、まずA/D変換されて、音声分析部(図２(a))により、特徴パラメータ（パターン）の時系列に変換されます。フーリエ変換や、線形予測分析に基づく周波数スペクトルの概形をあらわす特徴量と音声の強さ、およびこれらの時間変化量が特徴パラメータとしてしばしば用いられます。

この特徴パラメータの時系列を入力として、認識単位の音響モデルおよび単語辞書、文法を用いて認識処理を行ないます(図２(b))。音響モデルを構成する音声認識の最小単位（認識単位）としては、子音や母音などの「音素」や、ほぼひらがな１文字に相当する「音節」、音節を２つに分けた「半音節」[1]などがあります。モデル化の方法としては、HMM(隠れマルコフモデル）と呼ばれる方法が広く用いられています。

単語辞書には、上で述べた認識単位のならびとして各単語の発音を格納してあります。認識処理部は入力音声に対し、文法（言語モデル）と音響モデルに基づいて、最も発声された可能性の高い単語列を求め、それを認識結果として出力します。

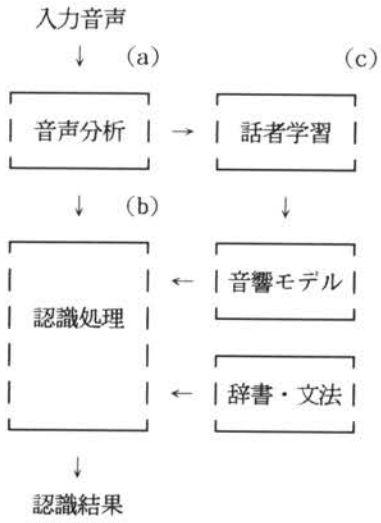


図２音声認識システムの構成

4. 自動通訳システムにおける音声認識の統合

ここで、また図１にもどって見ましょう。音声認識部(a)から、意味理解部(b)に渡される情報は何でしょうか？あるいは何が望ましいでしょうか？音声認識システムは、通常、認識結果として認識した単語列を、たとえば、

マドンナ を 二 枚 下さい
どの 席 が 一番 高い です か

のように、その表記の列として出力します。意味理解部(b)は、これを入力として形態素解析、構文意味解析を行なうことになりますが、音声認識の出力は単語の列なので、形態素分割処理は行なわずに済むことになります。

さらに、音声認識に用いられている言語的制約は、特定の応用ドメインに限定されているので意味理解部が用いる言語的知識に比べて強い制約を利用していることがしばしばあります。たとえば、上の例を

見ますと、「どの席が一番高いですか。」という発話文の中の「高い」という表現には、「位置として高い」という解釈と、「値段が高い」という２つの解釈の曖昧性があり、適切な翻訳を行なうためには、このうち「値段が高い」という意味を正しく選択できなければなりません。もし、音声認識部がより強い応用ドメインの制約をもち、意味の限定された情報を出力するならば、日本語意味理解部での曖昧性（多義）解消の負担は軽くなり、通訳システム全体としての性能は高くなるとという効果があります。また、音声認識部が構文・意味・タスクの制約を総合的にもっとも満足する文の仮説を認識結果として出力する場合、その仮説に対する構文・意味の内部表現がすでに生成されている場合もあり得ますのでそのような内部表現を意味理解部以降への出力とすることにより、重複した処理を避け、システムの効率性があがることも見込まれます。

また、現在、実用の音声認識システムにおいては、イントネーションの情報はほとんど用いられていませんが、

三 枚 あります

のような発話が平叙文であるか、疑問文であるかの曖昧性は、イントネーション情報を用いることにより区別することができるようになる可能性があります。

5. その他の技術課題

自動通訳システム技術は、現在、まだ十分実用化可能な段階には至っていません。実用化のための大きな課題の一つは、実時間の不特定話者大語彙連続音声認識技術の確立であると言えましょう。この他にも、システムを実際に使用するには、以下のような問題にも対処する必要があります。

(1) 耐雑音性

自動通訳システムは、ホテルのフロント、レストラン、旅行案内所など、大勢の人の話し声等による雑音環境の中で使用されることが予想されますので音声用と雑音用の二本のマイクを用いた雑音除去などの雑音対策が必要となります。

(2) 通信回線による歪み

自動通訳システムが、電話回線を通して使用される場合は、電話回線の伝送路特性による歪みが音声認識部の性能を劣化させる可能性があります。認識性能を維持するためには、入力音声からこの歪みを除去する機能が必要となります。

(3) 話者の割り込み

われわれ人間は、相手がしゃべり終る前に、それをさえぎって話し始めることがしばしばあり、またこの割り込みができることが、効率的なコミュニケーションを可能にする場合があります。自動通訳システムでこれを行なうには、通訳結果の合成音による出力が行なわれている際に、それを聞いている側からの音声入力が始まったら、音声出力を最後まで行なわずに途中で停止するという機能が必要となります。

6. おわりに

以上、自動通訳システムの概観と、その中で用いられる音声認識技術について簡単に紹介させていただきました。皆さんはどんな用途に自動通訳を使いたいと思いますか。

参考文献

[1] 畑崎他: 日英双方向自動通訳システムＩＮＴＥＲＴＡＬＫＥＲ, 情報処理学会第44回全国大会, 6P-5 (1992).

[2] 嵯峨山他: 自動通訳電話実験システムＡＳＵＲＡの概要, 日本音響学会講演論文集, pp. 83-84, (1993-05)

ヒアリング

「産業翻訳」革命の主力たる機械翻訳システム

スバルインターナショナル㈱　ＣＡＴＳ事業部長　鳴海　武史

プロフィール

スバルインターナショナル（以下スバルと略称）は1977年に柏原翻訳事務所として設立、85年に株式会社組織となる。資本金は現在3億8千万、社員数は35名である。産業翻訳会社としての新しい方向を実践するため、人手翻訳から撤退、機械翻訳専業に転換（英日翻訳のみ）、翻訳の生産性を高め、1年以内に納期、コスト半減を目指す。言語処理部、機械翻訳部を擁するCATS事業部が、社運をかけて取り組む企業戦略と生産技術はマスコミにも取り上げられ、センセーションを巻き起こしている。

1　産業翻訳業界の現状

●　バブル後の市場低迷

先ず私個人の業界の現状認識を述べたい。昨年の事だが、(社)日本翻訳協会の原田副会長の述べられたことの中に、「業界の市場価格は英文和訳でバブル期の30～40%も低下した水準で推移している」とあった。

一般的には400字詰原稿用紙で3,500～4,000円/枚の翻訳料金が、原田副会長の言によれば1,000円以上も低下したことになる。従来、翻訳会社が翻訳者に仕事を頼む時には恐らく1,500～2,000円前後だったのではないかと思う。市場価格が下がっているため、翻訳会社も翻訳者に値下げのお願いをしなければならないが、やはり相手は個人であるため、30～40%の話はもって行けない。精々10% ダウン程度の協力で終わっているのであろう。話を単純化するが、翻訳会社に残るお金は、例えば2,400円で受注したものを1,500円で翻訳者にお願いするならば900円しか残らない。これで翻訳会社は社員の人件費や家賃などの固定費を賄って行かねばならない。売上利益ベースでなんとかトントンでも、営業利益では全くの赤字であろう。この原因は受注量と市場価格がかつてなく減少したことによる。

産業翻訳業界での最大のお客様はコンピュータ関連業界である。新製品が出ればマニュアルが必須であるが、日本経済の失速で、国内需要は冷え込んだままである。

10年位前までは和文英訳が60%位を占めていたが、現在は80% 近くが英文和訳という話である。英文和訳は発注元が外資系となる場合が多い。外資系もかつて円高によって好況を呈したこともあったが、国内市場の冷え込みによって現在は苦戦している。したがって売れる商品を絞り込むなど、全体的に翻訳需要は減少している。

●　買い手市場の定着

産業翻訳業界は定常的かつロットのまとまった需要を歓迎するが、需要が減退してくると当然、従来の顧客以外に販路を開拓しようとするため、価格競争が起こって来る。翻訳会社にも種々あって、社長自ら翻訳にたずさわっているところから、ある程度の規模で、事務所を構えてやっているところなど様々である。組織だってやっているところほど苦戦しているのではないか。1枚あたり900円程度の粗利では到底やって行けず、「食べていくのがやっと」という現状であろう。しかし安ければ受注できるかというと必ずしもそうではなく、日本の商習慣からすれば、やはり馴染みとか古くからの取引実績がないと難しい。

外資系は随意契約が少なく、相見積りが普通で、価格的に弾力性をもった業者でないと受注できない。つまり完全に買い手市場になってしまった。この傾向はもはや戻ることはないと思った方がいい。その理由は、一旦安くなった価格が理由なくして上がることはあり得ないからである。現在コンピュータ業界の価格競争も熾烈であり、これに連動して翻訳料金の価格競争も続いていくものと考えられる。

●　コスト／スピード／品質の３要素

現在顧客が一番要求しているのはコスト、次いでスピードである。コスト、スピードの要求が満たされるならば品質はある程度目をつぶるというお客様もいる。ある外国の自動車メーカの話であるが、英文で4,000頁の SGML文書を印刷も含めて40日でやって欲しいとの要請があった。 残念ながら、これは産業翻訳会社以外の企業が落札し、実際に完成させた。これは従来ならば、半年はかかる代物である。この受注業者は機械翻訳でこなしたのではなく、手翻訳でやったが、過去の翻訳データを持っていたのである。つまりテキスト処理の技術を使って、恐らく、前にあった対訳データなどを利用して翻訳を短期間で仕上げたのではないかと思う。翻訳会社が応札できなかったのは現実にSGMLに対応できなかったことにもよる。その自動車会社が要求したものはまさにスピードである。品質はある程度我慢するにしても納期は絶対的な要求であった。したがって原文データも米国本社から通信衛星で日本支社に送信されし、即座に業者に手渡されたそうである。

● 大量翻訳に向かない人手翻訳

産業翻訳業界はこのようなお客様の要求するスピードについて行けなくなっているのである。お客様の要求はコスト、スピード、それにある程度の品質である。これに人手翻訳はもうついて行けないのである。 人手翻訳は人間が行うので、安くするにも限度がある。

翻訳者が生活していけないような価格では翻訳を依頼できない。受注する翻訳会社が自分達の利益を削っていくしか方法がない。

翻訳スピードも実績を積み、発注企業の翻訳文化に慣れていないと速く翻訳できない。翻訳者が変わると品質も変わる。翻訳会社が品質を保証できるのは１線級の翻訳者をどれ位抱えているかということになるが、品質管理コストを含めるとやはり限度がある。２線級、３線級の翻訳者を使うと、お客様にすぐ解ってしまう。発注側も「この翻訳は前にやった人以外にはやって貰いたくない」との逆指名さえある。工業製品ならいざ知らず、人が変ると品質も変わってしまうのが翻訳業界である。

新規のお客様は開拓しにくいという点もある。今まで持っていた翻訳文化が新しい業者の出現によって変わってしまう。また人手翻訳は決して大量処理ができない。

● 産業翻訳業界の収入

一般的には1日８時間でのアウトプットは400字詰原稿用紙で15～16枚が限度と思う。新人では5～6枚ではないかと思う。これらから計算しても翻訳者の年収が計算できると思うが…。以下も原田さんの言だが、平均的な翻訳者の年収は400万程度と言われている。これは20歳代だったらいいかもしれないが、40～50歳台では世間相場からして安すぎはしないだろうか。プロとしてならば相当きつい額と思うが、フリーランスであっても将来的な社会保障など考えれば厳しい環境下にあると言える。

産業翻訳会社の1人当たりの売上は年間1,000万位と言われ、30人いれば3億になるが、その内40%を翻訳者に支払えば会社には600万しか残らず、その中から社員の給料や経費を払わなければならない。翻訳会社の社員は当然のことながら低賃金に甘んじているのが現状である。また少ない人員で固定費アップを避けていることから労働過剰にならざるを得ない。まとまったロットには非常な競争があるため、競争に敗れれば利益がでないと知りつつ 5枚～10枚の小ロットのものも受注しなければならないのも事実である。この場合1人のコーディネーターが何十本も注文を抱えねばならず、しかも人が減らされる傾向にあるから労働荷重になる。したがって翻訳業はその業態を含めて歴史的な岐路に差しかかって来ていると言える。

では産業翻訳業界には未来がないのかということになるが、産業界は翻訳者がいなくなると困るし、翻訳者もまた翻訳が好きで続けている人が殆どだから仕事そのものがなくなるということはない。しかしビジネスとして実際にやっていることと、それに見合う収入が有るのかというと私には言葉がない。

2．機械翻訳に専業

● テキスト処理技術／自然言語処理技術

スバルは 6年前から早晩機械翻訳の時代が来るだろうという考えの下に、機械翻訳に取り組んだが、はっきり言ってこれまでは失敗だった。というのは機械翻訳を利用する技術が確立されていなかったのである。利用技術がないまま機械を導入したのである。

ただ業務用というか、プロとして使えるシステムと個人用のシステムは自ずから違って良いと思うが、業務用の機械翻訳システムは現行のものしか選択肢はない。機械翻訳の魅力は高速、24時間稼動できる、記憶能力が抜群で一度教えたことは忘れないという点である。そこでスバルでは機械翻訳システムの周辺に、どれだけの技術を構築できるかということに注力した。

機械翻訳システムの周辺技術とは 1つにはテキスト処理技術であり、1つには言語処理技術である。業務用として売りたいならば開発メーカもそこまで面倒を見ないと、翻訳業者はシステムを使いこなせないのである。例えばUNIXの環境を扱える翻訳会社はまだまだ少ないのではないかと思う。UNIXマシンすらない会社も多い。

各社のカタログを見ると高速翻訳を謳っているがユーザが習熟しない限り期待する成果は出ない。これはメーカの人は解っていると思う。また高品質かと言うとこれもまた違った話であり、本当に高品質ならばすでに翻訳業者は機械を入れて成功しているはずである。やはり「電源さえ入れれば使えるといったものではない」ということをきちんと説明しなければならない。

かつてお客様の求めに応じて翻訳会社は「機械翻訳でやります」と言っても、実際は手翻訳でしかやってこなかったのではないか。それはユーザ持ちの仕事が多すぎるからである。機械翻訳システムの能力を最大限に引き出すための仕事が多すぎるのである。ユーザの仕事を軽減するオプションツールを拡充していけば、もっと普及するのではないか。現状のままだとユーザの期待を満足させることはできない。

● 人手翻訳からの撤退

スバルではこの10月から人手翻訳より撤退し、人手翻訳の受注は一切行っていない。逆にいえば機械翻訳で受注できなければ会社は潰れることになる。それぐらいの決意がないと機械翻訳の専業化はできない。経済的に自分達を追い込むことが最善手であろう。人手翻訳への逃げが打てないのである。普通の営業マンならば、もし両方のカードを持っていたら、数字があがらないと人手翻訳で受注してしまうだろう。ただ機械翻訳でやると言うだけではお客様には納得して貰えないので、「翻訳の生産方式や生産技術について」は十分な説明を行っている。

「生産技術」とは翻訳システム周辺にあるテキスト処理技術と自然言語処理技術である。テキスト処理技術とは英語のドキュメントが作られているDTP、例えばインターリーフとかフレームメーカ、SGML、MSワードなど、スバルは基本的に磁気データしか扱わないが、テキストは勿論、英語DTPにあるレイアウトやフォント（これらを総称して仮にタグという）などの情報もすべて活用して行くという考え方に立っている。したがって最初にやる仕事はプログラム処理である。テキストとそうでない部分を分離する技術である。その後、日本語側で翻訳文と合成する技術もある。

● 辞書構築

ユーザ辞書構築をどうやるかも問題だ。通常ではお客様から用語集を貰うとかインデックスに出てくる用語をベースに構築する方法をとるが、スバルは独自に、テキスト中の単語列を頻度情報とともに抽出するプログラムを開発し、品詞情報なども参照しながら、ユーザ辞書登録候補語を高速に高精度に構築することができる。その候補語の訳語をお客様に確認して貰い、確認後、ユーザ辞書に登録、その辞書を使って翻訳している。これもテキスト処理技術と言語処理技術をミックスしたものである。

またバージョンアップ対応プログラムも開発した。既にシステムそのものに取り込まれているかも知れないが、同一英文の抽出ツールである。あるマニュアルでは 10%位全く同じ英文が出現することがあった。前バージョンの対訳をもっているなら、同一文は翻訳する必要はないので、翻訳費用は頂戴することはない。これもスバルの持っている技術が翻訳のコストダウンに繋がるような形でお客様に提供している。これもテキスト処理技術の1つである。

このようなプロセスを経て機械翻訳し、外部のリライタに日本語の校閲を依頼する。リライト結果をまたスバルでチェックするが、ここでも言語処理技術を使っている。現在はプリミティブなことしかできないが、有り得ない文字列はエラーとして検出できる。漢字変換のミスを検出するのは技術的には難しいが、平がなレベルはチェックできる。例えば「下さい」を「下ささい」と打ち間違えることがある。つまらないミスでもクレームはクレームなので、こ

ういうミスをなくすには従来は人間が時間をかけて校正しなければならなかったが、スバルはすべて機械処理をしている。このような技術がなければ、品質の安定化を図り、納期短縮やコスト低減を目指すのは難しい。スバルの技術をセールス時にプレゼンテーションするが、改めて機械翻訳について関心を示し注文してくれるお客様もいる。スバルの技術は機械翻訳の周辺の技術とも言えるが、このような技術をどれくらい多く作れるかが、これから鍵となる。機械翻訳システムメーカもスバルが開発したようなツールを製品にオプションツールとして付加しないと、ユーザはこのオプションを自ら開発しなければならないことになる。また言語処理技術については現場の人間が工程毎にいろいろなアイデアを出して生産性をどう上げるかといった検討を続け、常に新しい技術を開発している。一般の翻訳会社では技術者もおらず、こういう発想も出てこないと思うが、このような技術がないと機械翻訳は恐らく成功しないだろう。

ユーザ辞書構築は顧客との共同作業になるが、顧客はこれを嫌がる傾向にある。しかし将来的な果実を顧客と翻訳会社が共有するのであるから、顧客を説得して仕事を進めて行かねばならない。スバルはこの夏に英語で 3,000頁のマニュアルを６週間で版下まで完成させた経験がある。見積り段階から辞書を作成していたこともあるが、顧客も辞書構築に協力してくれたからできた仕事である。今後バージョンアップの注文が予定されているため、前に作った辞書が活かせる。当然２回目は辞書構築が楽になる分、コストが削減されるので翻訳費用も安くできるし、顧客も当然それを期待する。納期も前より短縮できる。これらはスバルのやり方に顧客が理解を示すことによって初めて可能になるもので、顧客の辞書資産が蓄積され、コスト、納期が軽減されることにも繋がる。

● 品質の管理

スバルの技術はテキスト処理技術と言語処理技術を組合わせたものである。人間がやらなくていいものは機械にやらせることが基本である。人間がやらねばならないことは翻訳の中身のところだけである。スバルの場合で言えば、翻訳工程中リライトチェックという所のみである。その外はすべてプログラムにやらせてしまう。 組版の大量編集については、テキスト以外の情報もすべて活かすことに重点をおいている。例えばインタリーフのデータで言うと、初稿段階では、日本語側に再現するだけであらかた編集ができている。英語を日本語に置き換えると文字が膨らんだりするので、イメージデータに文字が被さることがあるが、そこは改行をいれる程度で、元のレイアウトを再現できる。図や表の多いデータはオリジナルデータをそのまま活かすことができるので組版編集のコストダウンが可能になる。かつての版下作りは鋏と糊の手作業であったが、機械翻訳では短時間に対応できる。ただ日本語の場合フォントの種類が限られているので、例えばイタリック体の場合はどうするのかといった、事前に取り決めておくべき問題は残る。ただスピード、コストを追求する顧客にとってはスバルの技術が受け入れられると考えている。

3．新しい翻訳者層の創出

● リライタの生産性

スバルでは世間で言うポストエディタを敢えてリライタと呼んでいるが、今後は新しい職業になるだろうと考えている。リライタの生産性は中堅クラスになると 1日50枚はこなせる。機械翻訳の結果の中にはリライトせずに使えるものもあるので、どちらかと言えばエディタを使いこなす技術が必要だ。白紙状態からワープロに向かって翻訳して行くことと機械翻訳で下訳したものをリライトする場合を比べれば、肉体的、精神的にも疲れ方が違うとあるリライタは言っている。リライト作業は慣れれば楽だと言う人も多いと聞く。リライタの生産性をあげるためにユーザ辞書をきちんと作り、辞書メンテナンスを誠実に行えばリライタの習熟度も比例して向上し、生産性も上がる。これが合理的な方法であり、リライタにもコストダウンを要求できるのである。現在のリライタへの発注単価は将来的にはさらに安くできると考えている。継続性があれば、リライタにとっては以前よりリライトし易くなり、単位時間当たりの翻訳量も増えるからである。翻訳者の生産性のピークは年数で10年位しか続かないと言われている。翻訳という作業は本人の知力、気力、体力に左右される。生産性原理を翻訳に当てはめることはできない

と言う人もいる。生産性を追求できないという事は結局安く速くできないという事である。翻訳作業に生産性を求めること自体間違いだという人もいるが、逆に人的要素にしか生産性向上を求められないとすれば、人手翻訳の世界では永遠に安く速くできないことになる。スバルは翻訳会社として生き残るために、機械翻訳システムを上手に利用し、辞書を的確に作り、リライタビリティが上がるような環境を作って、リライタの生産性をあげていき、生産性が向上した部分をお客様に還元していく戦略を採用している。同時にリライタも収入も増やせるようになる。スバルはこのような考え方に立って事業展開をしていくし、日本の産業の下支えをして来た翻訳者にも、このような呼びかけをしているのである。スバルのリライトのほうが人手翻訳よりも収入がより得られると言っている。

● 翻訳者の未来

今後、翻訳者は人手翻訳にこだわる人と機械翻訳のリライトで一人立ちしていく人と、この両方をやるという 3つのタイプに分かれるのではないか。スバルは人手翻訳から撤退すると言ったが、従来から協力していただいた翻訳者に「今後は機械翻訳のリライトしか出せないけれどもどうか」と提案したら「自分は両方やる」といった人もいるし、「リライトでは」と縁を切っていった人もいる。翻訳者にとっては全く新しい文化であり、戸惑いがあるだろうが、産業翻訳業界はかかる方向をたどるものと思う。

4．「産業翻訳」革命

● 旧体制への挑戦

「機械翻訳は使えない」というのはユーザ側の声であるが、ユーザで一番声が大きいのは翻訳会社である。このような声は検証もせずに出されている。また問題意識がなく、現状に安住しているような人達、さらにメーカのドキュメント担当者など、機械翻訳に懲りた人の声が大きい。私はこれらの声に対してあえて挑戦している。できない、やれない、人間にしかできないと言う人に対して「机上の理論ではなくビジネスとしてやれるのだ」と実証したい。翻訳業界は発注側と受注側の密接な関係があって、中々新規参入ができないのが実情であろうが、このような体制にも挑戦して行きたい。このような姿勢や気概がなければ成功はしない。「技術があるから成功できるのではないか」という評論家的な意見もあろうが、「技術があっても成功するものではない」ことは歴史が証明している。自ら生み出した技術と共に敗れるならば悔いもないが、過去10年間機械翻訳に携わってきた者として、ビジネスとして成功できるということを見せなければ、翻訳業界のみならず機械翻訳システムメーカも、大きな打撃を受けるのではないかと思っている。結局は世間に受け入れられないものを作って来たと言われているのと同じである。私はコンピュータが自然言語を扱うという作業に性急に結論を出すべきでないと考えているが、ただ一般にビジネスは結果でしか評価されないのも知っているつもりだ。このビジネスは「不可能だ」といわれるようなことにも挑戦しなければ成功しないだろう。その意味で守旧派と革新派に分かれるが、守旧派からすればこのような無謀なことをする必要はないと言うだろう。スバルは旧体制に挑戦もするし、ビジネスとしても成功したいと思っている。

● 人手翻訳と機械翻訳の棲み分け

先に述べたように、コスト、納期、品質が翻訳サービスという製品を構成する大きな要素でなので、いつかは受け入れられると確信している。だからと言って、なんでも機械翻訳でやるべきだと言うのではなく、人間でしかできないものもある。例えば読み易くするためなるべく原文と離れた翻訳も構わないとの依頼もある。これなどはむしろライティングの世界で、スバルが目指す翻訳とは別である。例えば、シドニーシェルダンの翻訳など原文から離れ、面白おかしく翻訳している。このような翻訳とスバルが目指す産業翻訳は全く違った世界である。スバルが扱うものは大量で、コストとスピードが要求される産業翻訳なので、人間でなければできない領域を侵そうとは思わない。スバルはあくまでも磁気データしか扱わない。ハードコピーで渡され、OCRで磁気入力して欲しいといわれても断っている。OCR は認識精度にまだ問題があり、間違いもある。原文のレイアウト情報が活かせないのが致命的だ。

● 上げ潮で迎えるのはどちらか

「機械翻訳は使えない」と思い込んでいる翻訳業

者と、様々な困難に必死で取り組んでいる業者のいずれかが21世紀を上げ潮で迎えられるかと言えば、やはり顧客の要求に応えられる業者だと思っている。その時には翻訳業界の再編成が行われよう。機械翻訳を避けてきた会社、顧客のニーズに応えられない会社はビジネスフィールドには生き残って行けないだろうと思う。ただ個人翻訳者は残るだろう。また企業をリタイヤし、ドキュメント制作の経験のある人が定年前後に元の会社の援助を受け、別な形でドキュメント専門の会社を作るというケースが最近目立ってきている。品質というニーズには応えられるけれども、スピード・コストという点では問題が残るのではないかと思う。それに企業としての発展性に欠ける。私は今世紀中に再び機械翻訳のブームが来ると確信している。その時に世間が機械翻訳の方向を採用するのか、それとも問題解決の方向が見い出せない人手翻訳を使って行くのか、はっきり分かれるのではないかと考えている。

（質疑応答）

Q．周辺技術開発の目の付け所は？

A．やはり現場の声に十分耳を傾けることから始めなければならない。仕事をする場合、自分たちがもっと楽したいという気持ちがあるだろう。言語処理技術というものは極めて難しいが、テキスト処理技術でできるものは結構多い。例えば形式的なチェックをやるツールを作るだけでも、時間短縮には効率を上げられるし、納期の短縮はそれだけコストダウンに繋がる。つまり生産性が上がるということで、これらの積み重ねである。これらは現場の声として上がって来るものなのである。

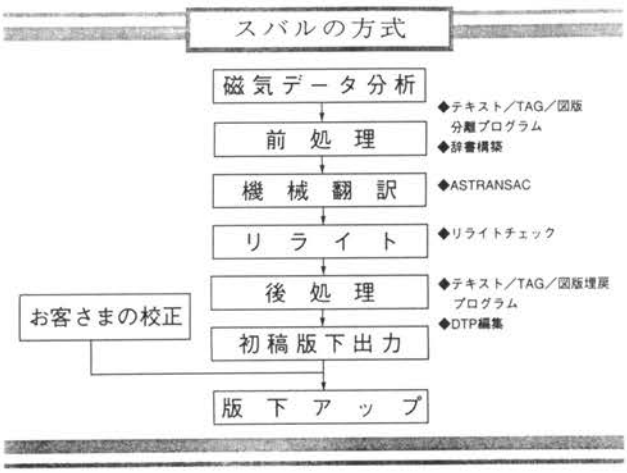
スバルは英日翻訳が主体で、前編集は一切やっていない。テキストとレイアウトを分離する前処理はやっているが…。さらにコード違いなど、日本語に再現できないものはコードの修正作業などをやるが原文に手を入れるということはやっていない。

Q．後編集のノウハウをソフト化するというテーマがあるが…

A．それは極めて難しい問題である。翻訳学校では「このような構文だとこう訳す」という教え方をしているが、これも難しいし、システムにそれを期待してはいけない。機械翻訳には速い、休まない、忘れないということしか期待していない。しかし辞書をきちんと作れば、もてる能力を100%出せるものは魅力である。ユーザにできるのはそれだけである。それ以外は人間がやらねばならない仕事である。

Q．英文でも長い文章は分割せずに機械にかけるのか

A．前編集はやらない。かつては前編集にウエイトをかけるか、後編集にウエイトをかけるか議論の分かれる所であったが、現状のＭＴの品質ではさほど差がないと思う。構文解析で失敗するというのはシステム側の問題である。スバルはリライトのみ外部に委託している。リライタから上がってきたものの品質をどうして上げていくかというテーマで、形態的なものも含めてテキスト処理技術を駆使している。翻訳の中身については人間しかできないと思うし、



人間がやれば習熟度は上がるものであり、習熟度と共に生産性を上げていく事を主眼としている。同一プロダクトでバージョンアップ品についてはコスト・納期半減をお客様に約束している。

Q．東芝ASTRANSACを採用した理由は

A.一番ライセンスが多かったこと、SUNマシンを8台も導入していたことである。ただ機械翻訳のレベルで頭抜けたものはないと考えている。我々のやろうとしていることはテキスト技術処理は何とかなっても、言語処理技術が多くからんでくるし、この言語処理技術はおいそれとやれるものではないので、その面での支援が東芝から得られるだろうということも理由の 1つである。現在パーソナルユースの翻訳ソフトも多く出回って来ているが、やはり業務用と言うかビジネスとしてやっていくという観点からすればプロ用としてスバルが開発した技術は欠かせないと言えるだろう。

Q．辞書の管理は

A.我々なりに管理方法は考えている。 スバル独自のマスター辞書を作っているし、これをどのように顧客別に構築するか検討している。スバルがやろうとしていることは工場生産的なイメージであり、工程が明確で、辞書の工程は何日で上げねばならないといった発想でやっている。やはり方法論が確立していなければ計画的に作業が進められない。そのパート、パートでまたどれだけ生産性が上げられるかという検討も行っている。

Q．パソコンソフトは使用しているのか。

A.昔は使用していたようだ。昔のものは UNIXとの比較ではスピードが遅い。それにディスクの制限やメモリーの制限もあり、使いづらい。機械翻訳周辺のツールはやはりUNIXマシンを使用して開発している。

Q．今後どのような開発テーマを検討しているのか

A．イメージデータからテキストデータを取り出す手法を検討中である。またMAC上で稼動するDTPの処理系も開発したい。翻訳レベルでそれほどコストダウンできなくとも、ほとんどが翻訳から版下まで作るという作業形態なので、翻訳で多少コストがかかっても版下レベルでコストダウンできる場合も多い。

また顧客側の協力も欠かせない。顧客によっては無用とも思える校正をしたがる人もある。校正は原則 1回だけとしている。顧客側にもマニュアル文化が確立していない。例えばセクションが違えば訳語や表記も違うというケースも多い。

Q．他の翻訳会社ではどのような対策を講じているのだろうか

A.ドキュメント部門にプログラム技術を持ち込むのは最近の話で、これまでは翻訳企業にはこのようなソフト開発者はいなかった。ドキュメントのこともよく知り、また翻訳のことも知っていなければソフト開発はできないのではないか。情報処理の技術がある人でも、テキスト処理に興味を持ち、翻訳現場の現状を改善しようという気持ちがなければ周辺技術に取り組まないだろう。プログラムが組めるというだけでは成功しない

Q．OCRを使用しない理由は何か

A.将来ユーザの要望が出てくればやるかもしれない。現在OCRはもってはいるが、タグが生かされないという問題もある。フォントの問題なども今の OCRではサポートできていない。だから再度原文にかえってゴシックを使うとかクーリエなど使うとすればそこに人手が必要となる。フォント情報をもっているのでそのままテーブル換算して日本語側に渡す方が良いという判断である。

Q．リライタの生産性と評価について

A.リライタを採用する時も試験をやっているし、依頼する時もリライト基準書を渡して、それに則って作業をして貰い、成果物を社内で評価するので、我々が期待した品質のものでなければ必然的に縁が遠くなる。

安定している人は今まで、手翻訳をやっている時もそれほど外部のチェックで赤が入らない。しかし初めて翻訳業界に参入する人はリライトに賭けている。今後スバルと付き合って人手翻訳では得られない仕事があるとか、生産性の問題で初めて確信を持たれるのではないのだろうか。人手でやると大量の場合やはり人数で分担するので、どうしても品質にムラがでる、その点機械翻訳でやると、個人のバラツ

キが目立たない。

リライトの基準としては機械でのアウトプットを生かして貰うことを前提にしている。テニヲハなんかは問題でなく個人の主観でこうあるべきだという主張もある。これがやはりバラツキを発生させる問題となる。現在の機械翻訳の内容をみると、やはり日本語を無理矢理生成した形跡がある。たとえば慣用表現のレベルで十分な生成が行われていない。したがって日本語の表現で、あり得ないものが矢張り出てくる。1 語で言うべき所を冗長な表現になっているところもある。これは現在のシステムでは避けられないものである。生成の段階でエバリュエーションなんかやって、こなれた日本語にしてあるものなどは現在ないのではないかと思う。解析では何段階かでやっているものはあるが、生成ではないのではないか。これはルールというより単語単位の問題だと思う。この辺りをリライターはこなれた対応をやっている。

Q. 翻訳業の今後のとるべき戦略は

A. 翻訳業のある者は付加価値の高い翻訳を目指していくだろう。例えば原子力とかバイオとかを狙って行く等が1つの方法だと思うが、これらは定常的に注文が有るものでもない。企業としては毎月一定量の安定受注が望ましいのである。それだけに競争も激しいのが実状である。バイオなど特注だけで 1年間過ごせるならば手翻訳でもできるかも知れないが現在時点では定量的な収入源といえばコンピュータ関連であり、これに期待するのは当然であろう。この分野では人手翻訳ではもう追いつかなくなるであろうというのが実感である。機械翻訳をやれば絶対的に速いが、その前と後が大変なのである。細かい話になるが先ず辞書を固めてから試訳してみるというステップをとる。所定の効果出ればこれを本番化するという手順である。ここまで行くまでにも種々の問題を発見できるのである。

Q 翻訳業界の変遷は

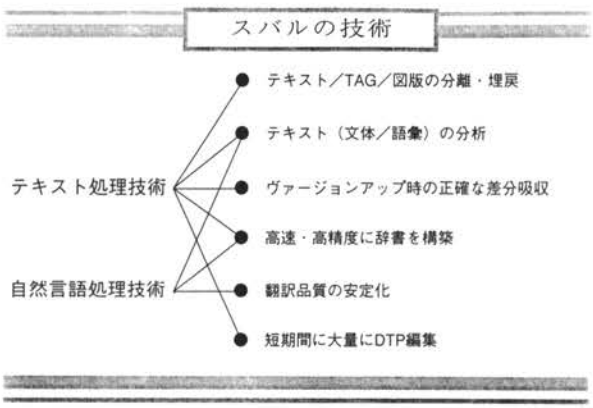
A 翻訳業界はすでに30年の歴史をもつが、未だに上場企業がない。業界大手企業でもかつて最高50億円/年しか売上がない。後発のソフト会社になると1部上場企業も出現しているのを見れば残念である。その原因はやはり大量翻訳ができないからである。結局人的要素が商品の価値の殆どを占める業界は生産性に問題が生じてくる。工場生産では設備投資を行い、研究開発なども積極的に行なっているが、このような発想をもたないとドキュメント業界の将来も暗雲が晴れないというように思える。

Q リライトのやりかたは?

A お客様からあがってきた校正はチェックして、辞書に反映できるあるものについてはマスター辞書の訂正を行う。校正結果でシステムに問題があると判断されるものはユーザには対処できないのでリライタに頼るしかない。例えばある英文のパターンが出来た場合には箇条書きにして貰うとか文体の変更をお願いするが、できるだけシステムに取り込むための努力はしている。

Q 現在は英日翻訳だけか

A 近い将来は日英もやりたい。現在の技術が完成すればそれを日英に応用して行きたい。ただ日英の場合、リライタの問題がある。チェッカーがネイティブになるので、日本人が評価するのが難しい。



ＮＴＴコミュニケーション科学研究所 池原研究グループ

ＮＴＴの研究開発本部には３つの総合研究所あり、合計１３の個別の研究所がそのいずれかに所属しています。研究者は約4,000名で、電気通信に関連する研究実用化に、日夜、取り組んでいます。

機械翻訳の研究は、現在、基礎技術総合研究所に所属するコミュニケーション科学研究所で行われていますが、本研究は自然言語処理研究の一つの柱として実施してきた経緯がありますので、その辺の概要を含めてご紹介します。

１．自然言語処理研究の経緯

ＮＴＴの自然言語処理の研究は、従来、基礎研究所のグループ（最近、武蔵野地区から厚木地区へ移転）と情報通信研究所（横須賀地区）のグループで進めてきました。基礎研究所では、理論的な研究が中心であったのに対して、情報通信研究所の研究は、実践的で実用化可能な技術確立を狙いとしてきました。

機械翻訳関連の研究は、平成６年４月、コミュニケーション科学研究所に移管されました。この研究所の本部は京阪奈地区にありますが、機械翻訳のグループは、横須賀地区内にとどまっております。

横須賀地区（情報通信研究所）における自然言語処理の研究は、1980年の開始以来、おおよそ以下のような経緯を辿ってきました。

まず形態素解析技術の確立を狙い、それを応用した日本文音声出力システムを実現しました。1985年から、本システムは、三鷹地区をモデル地区として実施されたＩＮＳモデルサービスに組み込まれ、電話を介した新聞記事朗読サービスに使用されました。

引き続き、ここで開発された技術をさらに発展させて、キーワード自動抽出システム、日本文訂正支援システム（ＲＥＶＩＳＥ）、音声認識用言語処理など様々な研究開発を行いました。キーワード自動抽出システムはパッケージとして、また、日本文訂正支援システムは、日本文校閲システムＶＯＩＣＥ－ＴＷＩＮに組み込まれて、商品化されました。日経新聞社では、1987年以来、新聞記事原稿（現在一日約70万字）の校正にこのシステムが使用されています。

その後、応用研究として、自然言語でアクセスできる「テキストベース内容検出・検索技術」、学習型の「日本語文書検索分類技術」、聞きやすい文への変換を含む「日本語読み上げ技術」、固有名詞や専門用語抽出を狙った「用語抽出技術」、電報受け付け業務用の「姓名漢字入力支援技術」など、さまざまな研究開発が行われてきました。

２．機械翻訳研究の経緯

機械翻訳の研究では、1985年、（イ）自然言語処理基盤技術の確立、（ロ）翻訳通信サービスに必要な機械翻訳技術の確立の２つを目標に研究を開始しました。

ＡＴＲが音声会話文の翻訳を手がけているのに対して、記述文の翻訳を対象としています。日英翻訳の研究開始に続いて、英日翻訳の研究を開始しましたが、技術的応用範囲の広い日英翻訳を優先するため、英日翻訳は、現在、中断しています。

従来の日英機械翻訳では、原文の前編集が必要と言われていますが、翻訳通信への応用を考えたとき前編集作業は大きな障害になります。そこで、従来の機械翻訳の限界を越えることを狙い、日本の伝統的な言語理論の一つである、言語過程説（時枝文法）の言語思想に着目して、多段翻訳方式を提案し、日英機械翻訳システムＡＬＴ－Ｊ／Ｅ*1を試作してきました。

この方式は、解析、変換過程において、原文の意味を失わないようにするため、表現の構造も意味の一部だと考え、意味的に見た（厳密には、意味ではなく言語上の約束から見た）表現の単位を抽出し、何段階化に抽象化する方式です。抽象度の低い規則から先に適用して、翻訳は実行されます。対象を客

体化した表現（客体的表現）と話者自身の客体化されない（直接的な）表現（主体的表現）を分ける観点からも解析と変換過程が構成されています。

多段翻訳方式を構成する技術は、大きく見て、意味解析技術と意味理解技術の２つに分けられます。そこで、段階的にそれぞれの技術を実現する方針をとり、全体の研究を２段階に分けて実行してきました。現在、意味解析技術を対象とする前半の段階をおおよそ終え、後半の意味理解型の機械翻訳方式の研究を開始したところです。

3．研究開発された翻訳技術の概要

既に多くの技術開発を行いました。それらの成果は、上述した研究の狙い（イ）に従って、多くの関連研究に応用されています。以下では、研究中の項目を含む、主な技術項目についてご紹介します。

（１）日本語解析技術の研究

形態素解析技術、構文解析技術、意味解析技術、文脈処理技術などの日本語解析基盤技術を実現しました。形態素解析技術では、解析精度 99.6～99.8%（単語単位）を達成しました。構文解析技術では、従来、言語解析の大問題とされてきた長文解析の問題に挑戦して、従属節の係り受け解析、名詞並列句解析の新しい方式を考案し、大幅な解析精度向上の見通しを得ました。

意味解析技術では、従来例のない詳細な意味辞書が完成したことにより、それを使用した表現構造の多義解消の研究が大幅に進みました。

文脈処理技術においては、省略された主語と目的語の90～95%を正しく補って訳すことのできる、高精度な自動補完技術を実現しました。

また、解析精度と訳文品質の向上を狙って、原文書き替え型翻訳方式を実現しました。本方式は、解析に失敗した結果を狙い撃ちして、訂正する能力をも持つものであり、途中の解析失敗は回復できないとする自然言語処理の常識に挑戦するものだと期待されます。

その他の解析技術では、意味解析をベースに、様相・時制解析、名詞句解析（並列解析、同格解析など）、複合語解析（機能語結合型など）、特殊表現解析などで新しい技術を実現しました。

（２）変換生成技術の研究

慣用表現変換、結合価パターン変換、汎用的変換の３種の変換パスからなる多段変換方式を中心に、日英変換技術を確立しましたが、さらにこれらを拡充するため、名詞述語パターン変換方式や文レベルでのパターン翻訳方式などに取り組んでいます。

また、従来の直訳型の翻訳の限界を超える翻訳方式として、大域的変換方式を考案し、部分試作によって効果を確認中です。

英語生成方式では、副詞句生成方式、決定詞生成方式等の研究に取り組んでおります。決定詞の生成では、従来、困難とされてきた冠詞、数の決定、所有格代名詞付与などの問題に挑戦し、新方式の考案により、冠詞と数について、高い精度で生成できることを確認しました。

（３）翻訳辞書の整備

翻訳辞書として、日本語辞書（40万語）、単語意味辞書（40万語）、構文意味辞書（1.5万文型）、日英対照辞書（38万語ペア）、英語辞書（８万語）を作成しましたが、さらに、実験によって整合性を検証しながら、拡充を続けています。

単語意味辞書と構文意味辞書の作成に当たっては、名詞に対する 3,000種の意味属性体系を確立しましたが、これによって、一部の機能動詞を除き、動詞を意味によって訳し分ける技術が実現されました。

現在、利用者辞書登録の枠組みの整備等を実施中です。それらの作成支援技術としては、対訳コーパスを使用した翻訳不良語の自動抽出方式を試験中です。

また、利用者登録語の意味属性自動獲得技術については、既に翻訳実験で使用しています。

（４）その他

翻訳実験では、汎用機能の充足を図るため新聞記事を翻訳実験の主な対象としてきましたが、客観的な技術評価基準を求めて、日本語表現を分類体系化し、それに基づき機能試験文集（対訳コーパス：6,200文）を編集しました。その他、必要に応じて、種々の分野の文書を取り上げています。また、これらの経験に基づき、対象分野に適合するシステムの構築を支援するため分野適合型翻訳方式の研究を開始しました。

主幹研究員 池 原　悟
（池原研究グループリーダー）

新製品の紹介

ヨーロッパ言語と英語間の機械翻訳システム

Power Translator TM

株式会社 イリス インターナショナル

【Power Translatorの誕生から今日まで】

Power Translator の誕生から今日まで約４０年近い歳月が流れた。1950年半ば、米国連邦政府の基金援助により、Georgetown 大学とIBM は、ロシア語から英語に翻訳する機械翻訳システムの開発プロジェクトに着手した。この計画に国防総省、中央情報局（CIA）および原子力エネルギー委員会（AEC）が加わった。このプロジェクトに当時参加したのが、Power Translator の生みの親であるGeorgetown 大学のBedrich Chaloupka 助教授であり、英語、ロシア語、ドイツ語に堪能な社会学者であった。

開発の初期段階では、270 のロシア語の言語解析アルゴリズムを開発し類似構造の文章をなんとか翻訳できるレベルまでに到達した。しかし、米国政府が期待する高度の複雑な科学および軍事関係の文書の翻訳にはほとんど役に立たなかった。1960年代に入って、政府は更に150億円の予算を追加した。

言語学者が中心となり、言語解析アルゴリズムを文章解析に連結する作業が進められ、最終的には約650,000種類のアルゴリズムを開発した。その結果、大型システムの容量が膨大となり、逆に作業効率が低下し、国防総省が要求する緊急の翻訳作業に対処することができなかった。しかも、機械が解読できる前編集とかアウトプットの後編集とか厄介な問題までは到底対応できる段階に至らなかった。遂に、1966年、『ALPACレポート』により、米国における機械翻訳の歴史はここで一端中断することになった。

1970年、Chaloupka 助教授は、今までの努力を無にしないために、Georgetown 大学の仲間たち２名とソフトウエア開発会社であるGlobalink社(Domini A.Laiti社長)にこの機械翻訳のプロジェクトを持ち込み、今までの大型指向から一転して、数百倍も規模の小さい言語解析アルゴリズムを使って言語のもっとも簡単な構成要素に基づくシステムを開発した。Power Translatorの開発は、言語学者が中心となり、各言語の semantics の解析から言語学的アプローチをしたことがその大きな特徴と言える。

1970年代後半、パソコンが登場し始め、IBM PC/AT対応の機械翻訳システムの開発に専念した。1988年12月、プロトタイプの製品を、ラスベガスで開催されたハイテクコンベンションで、フランス語、スペイン語と英語の双方向翻訳システムを発表した。

1990年10月、Power Translator の前身である Globalink Translation System（GTS）Version 2.0 のリリースと同時に日本での発売を開始した。

Power Translator による PCネットワークによる機械翻訳サービスを開始したのは、日本では1992年10月からNIFTY-Serve の『イリス機械翻訳』として米国では1994年6月からGlobalink社によりインターネット（Internet Message Translation System = MTS）を通じて開始された。

今日、３ヶ国語を公用語とするカナダでは、Power Translatorが公式の機械翻訳システムとして利用されている。日本では、大蔵省、宇宙開発事業団、動力炉核燃料事業団、ダイムラーベンツ、トヨタ自動車、その他の主要企業、大学研究室、特許事務所、さらには個人の翻訳専門家から語学研修者に至るまで幅広く利用されている。

【製品の概要】

Power Translator は、ドイツ語、フランス語、スペイン語と英語との双方向に翻訳する機械翻訳システムであり、米国 Globalink 社によって開発され、1993年に米国の "Discover '93" ソフトウエア優秀賞を受賞し、欧米で高い評価を得ている。

基本辞書として、約２５万語が装備されており、優れた翻訳精度、迅速な処理速度しかも操作が簡単であることが大きな特徴とされている。

現在、対応システムとしては、DOS（DOS/V）版, Macintosh 版、OS 版, UNIX 版 および Windows 版がある。Power Translator には、Professional（PTP）版と普及版（PT）との２種類があり、Pro-

fessional 版は、専門分野別のマイクロ辞書（オプション）の使用が可能であり、またユーザー辞書の作成が可能である。

国際ビジネスの現場だけで威力を発揮するばかりでなく、技術資料、文献、研究論文等の翻訳、さらにはヨーロッパ言語の学習にも快適なツールとして利用されている。

【特　徴】

1. 基本辞書 25 万語搭載
2. 翻訳速度 300-400 words/min.
3. 対話翻訳方式と一括翻訳方式をサポート
4. Professional 版は、専門分野別辞書(オプション)の使用とユーザー辞書の作成が可能
   なお、ユーザー辞書の登録は、1回の登録で双方向（例えば、英独・独英）に利用できる。
5. MS-WORD、WordPerfect 等のワープロ及びOCRの使用が可能
6. Macintosh 普及版は英語テキストの音声（男性・女性）出力が可能

【システム要件】

◎DOS版は、IBM PC/ATまたは互換機、DOS 5.0以上
（DOS/Vも可、ドイツ語はDOS 5.1以上）
メモリー 640K、HDD 20MB以上必要。

◎Windows版は、IBM PC/ATまたは互換機、DOS 5.0以上、Windows 3.1以上、メモリー1MB+EMS(4MB)
HDD 20 MB以上必要(ドイツ語は 25MB)。

◎MAC版は、System 7以上(漢字TALK 7.01も可)、機種 MAC IIシリーズ以上、メモリー2 MB以上、HDD 20MB以上(ドイツ語は22 MB) 必要。

◎OS/2版は、OS/2 (Ver.2.0以上)、HDD 20 MB 以上必要。

◎UNIX版は、motiy対応、SUN OS 4.1、Silicon Graphics WS、IBM RS-6000対応。

【価　格】

Professional 版

PTP(Ver.3.0)DOS	¥148,000
PTP(Ver.4.0)MAC	¥168,000
PTP(Ver.4.0)OS/2	¥180,000
PTP(Ver.4.0)UNIX	¥480,000

普 及 版

P T(Ver.2.0)DOS	¥38,000
P T(Ver.2.0)Wins	¥58,000
P T(Ver.2.0)MAC	¥58,000

価格は、各言語とも双方向 パッケージの価格

【専門分野別マイクロ辞書】

現在、言語別、対応システム別に次の専門分野別マイクロ辞書が用意されている。
Automotive, Business/Finance, Banking, Legal, Computer, Environment, Medical/Pharmaceutical, Telecom/Cables, Brewing, Aviation, Chemical, Petroleum/Mining, Infantry, Narcotics, School, etc.

・専門分野別マイクロ辞書(オプション)は言語別、1専門分野につき ¥ 28,000、
UNIX版は言語別(専門分野一括磁気テープ収録)
¥ 68,000

【NIFTY-Serve機械翻訳サービス】

現在、Power Translator Professional の Macintosh版および DOS 版を使用して、『イリス機械翻訳』として、ドイツ語、フランス語、スペイン語と英語間の機械翻訳サービスを提供致しております。

【その他】

1. 英語から中国語への片方向翻訳ソフト(DOS 対応版)が本年10月リリース、繁体語・簡体語サポート。価格は¥28, 000。来春、英語中国語双方向のシステムがリリースされる予定。
2. 将来の開発予定に、ポルトガル語、イタリア語が含まれている。

（お問い合わせ先）
株式会社　イリス　インターナショナル
〒104 東京都中央区銀座7-17-8 銀座松良ビル
TEL　(03)　3542-0950　(代表)

新製品紹介

日英翻訳支援ソフトウエア
LogoVista E to J Personal

ロゴヴィスタ株式会社

国際情報化社会の進展に伴い、ビジネスマンや研究者は英文情報の洪水の真只中に置かれ、そのタイムリーな内容把握と消化は日常のシーンの中では重要な業務となってきている。パソコンによる英日翻訳ソフトLogoVista E to Jは一昨年の発売以来、最先端のソフトとして、多忙な現代人が頼れるツール（パートナー）として、高い評価を受けている。

ワープロなどで作成された英文テキストを Logo Vista E to Jに読み込ませるだけで、鋭い英文解析力と的確な意味処理と、こなれた和文生成力を駆使して、充実した基本辞書と専門辞書を擁して、さらにはユーザーが構築するユーザー辞書を使って自動的に日本語訳を作り出すシステムである。

特に従来の翻訳ソフトにありがちだった面倒な前編集を必要とせず、反対に翻訳結果に別解釈や品詞の設定などの経験をＡＩ機能により学習され次回以降の一次翻訳結果に反映されるという機能が好評を得ている。

このLogoVista E to Jで鍛えられた最新の機能をベースに、この１１月により多くのユーザーに使って貰うべく、またE to Jへのエントリー・バージョンとして、簡易版LogoVista EtoJを発売した。

以下、LogoVista E to J Personalの機能と特徴を紹介しよう。

●E to Jで鍛えられた翻訳エンジンはそのまゝスピードコントロール機能

翻訳支援システムに要求される要素は、まず第１に精度の高い翻訳結果を出すことで、次に翻訳にかかる時間が短いことにこしたことはないわけです。しかし、翻訳対象である英文が長くなればなる程、大きなメモリ空間を必要とし、処理時間もかかることになる。一方、パソコンの実装メモリはユーザによって異なることを考えて、E to Jでは実装メモリサイズに応じた能力を発揮できるようにユーザが選択できるようにしたのが「スピードコントロール機能」である。E to J Personalではこの機能をさらに進めて、ユーザの環境(メモリ容量)に応じて、翻訳スピードと精度の最適バランスが自動的に設定される。

勿論、この設定値は変更は可能になっている。１～１００まであるパラメーターのスピードを上げると、優先度の高い解析に絞り込む機能が働き始め、さらに自動的に文章を分割する機能も働く。しかもこの機能は長文程効果を発揮する。この最適値を E to J Personalが自動的に決めてくれる。

● 翻訳できる英文の長さ制限—３０ワード／文

E to J Personalでは一回に翻訳できる英文の長さが３０ワード(句読点も１語とする)までゞ、それ以上の場合は意味上分割できる箇所で区切って３０ワード以下になるようにして再翻訳すればＯＫ。
大抵長文の場合はセミコロンや関係詞などのところで分割できる。Personal Useとしては十分であり、その分“軽い”ソフトに仕上がっている。

● 別訳語を簡単に入れ替可能

英文の翻訳結果（一次翻訳結果）に対し、意図していた訳語が出て来なかった場合、別訳語機能を使って、より適切な訳語を見付けだして、本訳文の日本語と入れ替えることができる。また次候補に出てくる別訳語の何れも適切でないと思われたら、自分で訳語をユーザー辞書に簡単に登録もできる。
E to Jではシステムが最適と思われる日本語訳を作成して来るが、文法レベルまで溯って（例えば動詞を名詞と解釈したりして）、別な解釈も最大２０通り、画面上で覗くことができた。
E to J Personalではこの別解釈機能は備えていない。言い換えれば、それだけ一次翻訳結果にシステ

ムが自信を持っている裏付けとも云える。

● 一般辞書とユーザー辞書による翻訳

システムが具備している一般辞書の収録語数は E to J も E to J Personalも約１０万語と同じであり、ユーザー辞書もともに自由に構築できる。翻訳業務やプロフェッショナルな翻訳ツールを目指しているE to Jでは、より専門的な分野に対応するために、２１分野の専門辞書をオプションで用意している。しかし、一般のビジネスマンや学生などが使う場合は一般辞書と自分達の業務で使う言葉や、頻度の高い特別な用語はユーザー辞書に登録して使えば充分対応できるし、辞書は少なく軽くするにこしたことはない。

● ユーザーの英語力に応じて使える品詞設定機能

機能比較表

(LogoVista EtoJ Personal/LogoVista EtoJ)

	LogoVista EtoJ Personal Windows版 V1.5 Mac & PowerMac版 V1.5	LogoVista EtoJ Windows版 V2.1 Mac & PowerMac V2.1
(機能)		
翻訳速度	使用できるメモリ容量から、翻訳スピードと精度の最適バランスを自動的に設定(設定値は変更可能)	使用目的に合わせ、翻訳スピードと精度を設定可能
一括翻訳できる英文の長さ	テキストデータ8KB	テキストデータ80KB
翻訳できる英文の長さ	30ワードまで	121ワードまで
未知語検索機能	○	○
品詞設定機能	○	○
別訳語機能	○	○
別解釈尺機能	×	○
AI学習機能	×	○
(辞書)		
一般辞書	収録語数;英和約10万語	収録語数;英和約10万語
ユーザ辞書	複数個作成可能、ただし、翻訳実行時は1つのみ	複数のユーザ辞書が同時に使用可能
ユーザ辞書ユーティリティ	○	○
専門辞書	使用不可	オプションとして21分野の専門辞書使用可能
辞書の優先順位	1つのユーザ辞書と一般辞書の優先順位を設定可能	複数のユーザ辞書や複数の専門辞書が同時に使用でき、一般辞書とともに優先辞書の設定可能
(動作環境)		
本体	【Windows版】 Windows V3.1日本語版 NEC9800シリーズ IBM PC/AT互換機およびDOS/V機 【Mac and PowerMac版】 MacIIシリーズ以降、Powerbook140以上または PowerMac　OS;漢字Talk7	【Windows版】 同 左 【Mac and PowerMac版】 同 左
空きメモリ	【Windows版】 9MB以上 【Macintosh版】 5MB以上(7MB以上推奨) 【PowerMac版】 6MB以上(8MB以上推奨)	【Windows版】 9MB以上 【Macintosh版】&【Power Mac版】 6MB以上(8MB以上推奨)
ハードディスク空き容量	25MB以上	35MB以上
プロテクト	ソフトウェアプロテクト	ハードウェアプロテクト
(翻訳方法)	トランスファー方式	トランスファー方式
(価格)	¥79,800	¥194,000

英文を解釈する際の解釈の違いによって和文も当然変る。日常、私達が接する英文は必ずしも構文上、はっきりとしたものばかりではない。構文があいまいな文や複雑な文ではいく通りもの解釈が可能となる。このような時、あらかじめ英語の語句に特定の品詞やグループ化の設定を行って置くと、翻訳意図に沿った和文を生成していくことができる。

この品詞設定機能はE to J Personalにもあり、名詞(句)、動詞(句)、形容詞(句)、グループ化および英語のまゝにする等、ユーザーの英語力が充分発揮できる後編集機能のひとつと云える。

● システムが学習能力を備えたE to J

E to J では編集に使用された英語の語句の特定の意味と品詞の使用頻度を統計ファイルに保存するようになっている。これにより、ある意味で翻訳を行なえば行なうほど、利用されるユーザーにあった訳文が導き出されるように、「成長」していく。

E to J Personalではこの学習機能は備えていない。

以上、翻訳のプロフェッショナル・ユースを目指すLogoVista E to J と、手軽な個人向けに新たに発売した LogoVista E to J Personalの主な機能を紹介しょう。

価格的はE to Jが¥194,000に対しE to J Personalは¥79,800と、より手軽に提供されている。

そしてE to J Personalの購入者には¥126,000で、E to Jへのアップグレードの道が用意されている。

今や、ビジネスマンや研究者が自分のパソコンにLogoVista E to Jを搭載して、ワープロや計算ソフトと同じように、英日翻訳ソフトを手軽に、時には“秘密兵器”として、あるいは“文房具感覚”で使える時代に来ている。

この製品に関するお問い合わせは
ロゴヴィスタ株式会社（TEL 03-5690-8531）

新入会員の紹介

金 上 康 志　〒257　神奈川県秦野市北矢名1183-22　☎
高 橋 雅 仁　〒811-31 福岡県糟屋郡古賀町千鳥２丁目22-12　☎092-942-2398(九州松下内)

協会活動報告

第２回理事会	10月11日	（書面審査）①栗林理事の退任、長沢理事(富士通)選任について承認
拡大運営委員会	9月26日	①運営委員長の交替 ②協会活動の活性化対策 ③下期活動計画(委員、研究会)の方向付け④翻訳の日行事
運営委員会	10月14日	①運営委員会規定の見直し ②組織の見直し ③研究会活動方針確認 ④上期収支状況報告 ⑥ユーザ講習会開催準備状況
	11月29日	①委員会規定の見直し ②ユーザ講習会開催結果報告 ③技術委員会設置案 ④今後の活動予定
例文評価研究会	10月24日	①研究の方向付け ②ユーザ例文ネイティブチェックの評価研究
	11月28日	①ユーザ翻訳翻訳困難例文の言い換え例検討 ②ネイティブチェックの評価
	12月19日	①ユーザ翻訳困難例文の言い換え例の研究 ②今後の研究活動の方向付け
需要予測研究会	10月20日	①ヒアリング“産業翻訳革命”②既存調査資料の検討(ホワイトカラーの生産性)
	11月30日	①ヒアリング“ＯＡと事務生産性”②既存資料の検討(情報流通と業務の生産性)
利用技術研究会	12月16日	①ヒアリング“PC翻訳ソフト利用による技術翻訳実施結果”
講習会準備会議		②10月6日(講習会会場設備、スケジュール検討)③10月19日(講習カリキュラム検討)
□ ユーザ講習会	11月25日	名古屋商工会議所(機械翻訳レクチャー、デモンストレーション、翻訳実習)
# 翻訳フェア	10月21日	（東京YMCAホテル)(JTA主催)協賛事業への参加
# 翻訳の日事業	9月30日	（東京産経ビル)10月1日(大阪日経新聞社)協賛事業

MT-SUMMIT V

第５回機械翻訳国際会議

開催月日　平成７年７月１１日～１３日

開催場所　ＥＣ本部ヘミサイクル(ルクセンブルグ)

テーマ　＊ＭＴの経済側面　＊専門分野用語と電子化辞書

＊実業界へのインフラとＭＴの融和　＊オンラインＭＴサービス

＊ＳＭＥ対応のＰＣ版ＭＴ　＊ユーザ側からみたシステム評価基準

主催　ＩＡＭＴ(International Association for Machine Translation)

後援　ＥＣ　ＸⅢ部会

協賛　ＥＡＭＴ，ＡＡＭＴ，ＡＭＴＡ

組織委員長　Mr.Phillipe Van Den Daele(SEMA Groupe Benelux)

連絡先　Place du Champ de Mars 5,Bte 40,B-1050,Bruxelles

Tel.32-2-512-5317 / 32-2-508-5401　FAX.32-2-512-1076

[SEMAルクセンブルグ支社…SEMA Lux :Avenue de la Gare,65,L-1611 Luxembourg]

(Tel.352-40-08-11-5022　FAX 352-40-00-83)

ＡＡＭＴ
ジャーナル
No. 9 (Dec.1994)

発　行　アジア太平洋機械翻訳協会
所在地　〒103東京都中央区日本橋小舟町4-10 大宮ビル
☎ 03-3664-5637/5638　FAX 03-3664-1352
E-Mail (PC-VAN) JRD08020　(NIFTY-SERVE) KGE00013
編集委員　野村浩郷(九州工業大) 杉山健司、亀井真一郎
事務局　星野禎男　西郷容子