自然言語処理 85−13
（1991. 9. 20）

機械翻訳使用のための統合的環境

西野文人　伊吹潤　中村直人　内田裕士

富士通研究所

これまで一文一文の翻訳に対しての方式については多く論じられてきた．しかし，文書全体をどのように翻訳し，翻訳結果をどのように利用するかといった統合的なシステムとしての議論は少ない．手軽に翻訳システムが使えるようになった現在，これまでの大量翻訳の工数削減だけではなく，情報獲得やコミュニケーションの支援としての翻訳システムの使い方に注目する必要がある．本稿では文書翻訳のための統合的機械翻訳環境として，文書の型・構造の解析，翻訳の制御，文の書き換え処理，文書の整形の機能について述べ，さらに本来の翻訳処理部のあるべき姿を検討した．

Integrated Environments for Machine Translation Utilization

Fumihito NISHINO　Jun IBUKI　Naohito NAKAMURA　Hiroshi UCHIDA

Fujitsu Laboratories Ltd.

1015, Kamikodanaka Nakahara-ku, Kawasaki 211, Japan

The emergence of low-cost machine translation systems has substantially changed their forms with regards to the user. They are no longer kept exclusively for large-scale users, but are open to a variety of personal users, each with different needs and therefore, with different system requirements. In order to respond to these demands, we need to evaluate the system with a new point of view such as how we treat the whole document, and for what purpose we use the system. In this paper we will describe the integrated translation environments we have created, including document-format analysis, the rewriting process, and the reformatting of the translated result, and estimate several requirements of the core machine translation system.

1 はじめに

これまで機械翻訳システムは大量の科学技術文書の翻訳作業の支援を行なうことによって翻訳作業の工数を削減させることを主たる目的として使用されてきた．そして，翻訳の品質を上げることが開発の優先とされてきた．しかし，一般の人も手軽に機械翻訳を利用することができるような環境が整ってきて，使われ方も当然変わってきている．単にそれぞれの文の翻訳の品質ばかり見ていると，真の目標を見失いかねない．我々は統合的な文書処理の立場から機械翻訳システムについて見直し，その利用方法について検討した．

2 機械翻訳の使われ方

単なるテストやデモとしての翻訳でもない限り，単に文を翻訳したいということはほとんどない．実際的に使用する場合は，ある文書を翻訳して原文と異なる言語で文書を完成させたかったり，母国語でない言語で書かれた文書から情報を得たかったり，言語の異なる者どうしでコミュニケーションを図りたかったりという目的があって，それらの作業の中の一作業として翻訳作業がある．

これまでの機械翻訳システムの研究開発は，翻訳処理（与えられた文をどのように翻訳するかという部分）に対しての研究開発に主眼が置かれており，実際に機械翻訳システムをどう利用し，翻訳結果をどうするのかといった翻訳の目的（外国語文書を作成するのか，外国語文書から情報を獲得するのか，外国人とコミュニケーションをするのか）から考えた文書全体の処理，統合的な環境での機械翻訳システムのあり方についての議論があまりなされなかった．

2.1 大量翻訳の支援としての機械翻訳

これまで機械翻訳システムは大型汎用機あるいはスタンドアロンなワークステーション上で動作していた．このような環境下では一般人が気軽に機械翻訳を利用することはできなかった．したがって，機械翻訳の利用者は翻訳を日常の業務とする者に限られていた．このような状況から，これまで機械翻訳システムの典型的な利用方法として「多量文章の一括翻訳」や「対訳エディタを使っての翻訳・編集」が考えられていた [1]．そして機械翻訳の目標としては，「翻訳者に比べて大量の翻訳をはるかに短時間で行なうこと」，「分野によっては数百万語に達する用語の適切な選択を可能とすること」などが言われてきた [2]．したがって利用の仕方としては，あらかじめ専門用語辞書の作成などの前準備に相当な時間をかけ，翻訳を行なうべき文書は定型的なフォーマットを持ち，さらに前編集・後編集を人間が行なうというものであった．

2.2 読解支援やコミュニケーション支援としての機械翻訳

近年，ワークステーション上で，あるいはパソコン通信などのネットワークを介して，一般の人も手軽に機械翻訳を利用することができる環境が整ってきた．このような状況では，論文や手紙などの文書作成の他に外国語で書かれたオンラインマニュアルや電子ニュースの記事などの文書読解（情報獲得）の支援として，翻訳システムが多く利用されるようになってきた．また，日本人と外国人との間での電子会議のコミュニケーション支援としても利用されようとしている．

このように文書読解やコミュニケーションの支援として機械翻訳を使おうとすると，翻訳工数削減という使い方と比べて，機械翻訳に対して要求されるものは当然のことながら変わってくる．文書の種類は様々であり，これらを前編集なしで翻訳しなければならない．しかし，一文一文の翻訳品質は文書作成のときほど重要なものではなくなる．キーとなる部分が訳されていて，文書全体として意図が通じるか，読みやすいかなどが問題になってくる．そのためには，翻訳失敗とだけ言って何も翻訳結果を出力しないというようなものは困る．大意を伝えるための工夫が必要になる．

3 統合的翻訳システム

これまで製品化されている多くの機械翻訳システムは，一文一文に対しての翻訳処理あるいは特定の文書処理システムに依存した処理に重点がおかれてきた．しかし，機械翻訳システムを実際的に利用するには，翻訳処理以外の様々な作業が任意の文書に対して必要になる．

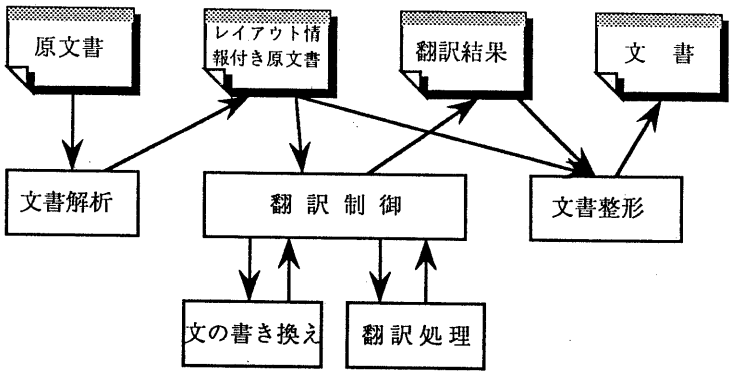

図 1: システムの構成

これらのすべてを機械翻訳本体で行なうのではなく，システムを次の５つのパートに分けた（図 1）．

文書解析 文書型の判定や文書構造の解析を行ない，文や記号列の切りだしを行なう．

翻訳制御 文の書き換え処理の呼びだし，翻訳を行なう行なわないの制御，翻訳失敗時の処理などを行なう．

文の書き換え 文の明確化（文の書き換え，長文分割，係り受け制御，省略語補いなど），翻訳失敗時の文の書き換え，翻訳結果の後編集などを行なう．

翻訳 単にある言語の文を指定された言語の文に変換する．

文書整形 原文，原文書のレイアウト構造と翻訳結果をもとに，適当な表示・印刷形式をつくる．

このように，機械翻訳本体で多くのことを行なわずに，機能を分割して機械翻訳の外側で多くのことを行なうことは，例えば文の明確化処理などはあきらかに解析の二度手間になるというような無駄もあるのだが，以下のようなメリットの方が大きいと考えている．

機械翻訳本体の文法が簡素化される

現在の機械翻訳の文法規則はモジュール化されていても非常に複雑になっていて，その管理が非常に困難になっている．現在の文法規則には，各言語の構文規則に則って処理する部分と，意味的・語用的な優先規則を示しているヒューリスティックス部分がある．このヒューリスティックス部分は各分野や文体に依存している．ヒューリスティックス部と言語の構造を記述した変換部とははっきりと分離することによって，翻訳本体の文法規則の拡張性・管理性を良くすべきであると考えた．

使用目的・対象に応じた処理

機械翻訳をどのような目的に利用するかによって，機械翻訳に対する要求は当然のことながら変わる．例えば，機械翻訳を翻訳家の人が大量の文書の下訳に使うならば，苦し紛れの翻訳結果を出力するよりは，単にエラーと言って何も翻訳結果が出力されなくても構わないかもしれない．しかし，情報獲得やコミュニケーションのための機械翻訳では，とにかく何らかの情報が伝わってくれなければ困る．そのためには，失敗時処理のようなものが必要になるのである．このように使用目的によって機械翻訳に対する要求は変わってくる．これらは翻訳の呼び出しを制御したり，文書整形の仕方を変えたりで対処すべきであると考えた．

ユーザチューナビリティの向上

これまで機械翻訳の文法や辞書は機械翻訳を提供するメーカ自身が開発・管理してきた．それは，機械翻

訳のための文法開発が少なからずプログラミング的素養を必要としていたからである．しかし，メーカーから提供されるシステムは利用者にとって完全なものではない．利用者が翻訳システムを使っていく上で，何回も同じ単純な誤りをされると頭にきて使う気がなくなる．利用者がシステムに対して，誤りの修正や，利用者の持つノウハウを教えることができ，翻訳結果を自由にカスタマイズできることが必要である．

利用者が機械翻訳の文法や辞書をカスタマイズする方法として，実際の機械翻訳の文法を修正させたり追加させたりする方法をとることは，管理を非常に困難にする．直接文法規則をいじらないでもすむように，あらかじめ文法の中にいろいろな変数を用意しておき，利用者にはその変数の値を変更させる方法もあるが，利用者からの様々な要求をすべて変数化しておくことは文法規則も複雑になるし，対処しきれるものでもない．

我々は利用者に対して，機械翻訳の外側で原文を書き換えたり，翻訳結果を書き換えたり，あるいは原文を直接翻訳結果に書き換えたりできるシステムを用意した．この方式では，インタフェースが文そのものなので内部ロジックを知らない利用者でも容易にチューニング可能である．

翻訳システムに対する独立性

機能が分割されているので，これらの機能を道具として利用可能である．そして，他の機械翻訳や，他の自然言語処理システムにも適用が可能である．

4 周辺処理システム

ここでは実際の周辺処理システムがどのようになっているかについて述べる．

4.1 文書解析

我々は翻訳作業を文単位の翻訳ではなく，文書全体の翻訳であると捉えた．その文書の種類（手紙なのか，論文なのか，特別の形式のものなのか）によって，翻訳の仕方を変えたり，それを元にした情報を利用する（例えば，ある部分には必ず人名が来るなど）ことによって誤訳を防ぐためである．そのためには，まず文書がどのような種類の文書であるのかを知り，文書の種類に応じた翻訳処理を行なうことが重要である．例えば，電子ニュース型の文書であることが指定あるいは推測された場合には，最初のヘッダ部の処理，記事中の引用の処理，signatureの判定処理，電子メールアドレス記号列処理などが起動さ れなければならない．

文書種別判定

システムはまず与えられた文書に対して，その文書の言語，文書型（マニュアル，手紙，論文，特許など）や全体的なフォーマット情報（右端位置，行数など）を利用者の指定あるいはシステムによる推定によって得る．

文書構造解析

文書の型，レイアウト情報に基づいて，文書の構造（タイトル，段落，箇条書きなど）を解析する．そして，文書中の各パートの役割を決定し，さらに文の切りだしを行なう．この結果得られる 文書は，機械翻訳で処理しやすい1行1文の構造であるとともに，その文の文書中での役割（タイトル，箇条書きなど）および，文書のレイアウトを復元するためのレイアウト情報（センタリング，右寄せなど）を合わせ持ったものである．

4.2 翻訳制御

文書の中から翻訳必要部分を取りだし，形態素解析を行ない，文の明確化，翻訳，翻訳失敗時処理，後編集が行なわれる．これらの処理は翻訳の目的に応じて適当に組み合わせられる．

翻訳不要部分の制御 例えばレイアウト情報や，文書中のプログラム，数式，住所・電話などは翻訳を必要としない．このような部分はそのまま出力する．

直接翻訳 あいさつ文やことわざなどはそれらの文の意味を分析するのではなく，決まりきった翻訳結果を出力すればよい．すなわち，あらかじめ与えられたある特定のパターンの文に対しては，

文を分析しての翻訳を行なわずに，与えられた訳文を直接出力する．

機械翻訳システム依存処理 文の明確化は特定の機械翻訳システムに依存しないように規則が作成されるべきである．しかし，すべてがこのような書き換えだけでうまくいくわけではない．システムに依存した処理もある程度必要であり，そのために機械翻訳システムによっては特定の情報（係り受け先や，品詞，訳語など）を指示するための特殊な入力インタフェースを持っているものも少なくない．文の明確化は機械翻訳に依存しない標準フォーマットで行なわれ，その後，機械翻訳に依存して再編集あるいはその機械翻訳本体がたまたまうまく処理できない部分の修正処理が行なわれる．

失敗時処理 情報獲得やコミュニケーション支援では，情報を漏らさず正しく伝えるのが重要である．そのためには，機械翻訳システム本体が翻訳処理に失敗した時には，何らかの形で翻訳リトライを行ない，情報を伝達することを努めることが重要である．このような失敗処理を翻訳本体内で行なうのでは融通性がない．そこで，翻訳の外側で文を書き換えることによって，翻訳リトライをすることにした．例えば，翻訳が失敗した場合には，文を分割したり，修飾語句の削除などによる圧縮処理を繰り返し行ない，最終的には単語レベルの翻訳を行なってでも情報を伝達するというようなことを行なう．

4.3 文の書き換え

現在の機械翻訳レベルではどんな文でも正しく翻訳できるわけでもない．あいまい性の少ない明確な文を与えなければ正しい翻訳結果は望めない．そこで一般的には機械翻訳処理を行なう前に人間によって文の明確化を行なう前編集作業が行なわれる．しかし，機械にとってはあいまいな文でも，専門知識や状況などから人間にとってはあいまいでないということも多くある．このようなあいまい性をすべて人間が解決しなければならないとするとこれは苦痛である．また，情報獲得のような使い方では，人間による前編集は望めない．そこで，機械が自動的に自分の解釈に従って前編集を行なうことにした．

文の明確化のための前編集項目としては以下のようなものを行なっている．

いい回しの変換 分野や文体に特定な言い回しは，標準的な文に書き直す．

- ～に追われていました ⇒ ～で忙しかった
- ～に至っておりません ⇒ まだ～しておりません

省略語の補い 主語の補い，共有化されている述語の補いなどを行なう．

- 先ほどの質問にお答えします． ⇒ 私は先ほどの質問にお答えします．
- 彼は仙台へ，彼女は名古屋へ行きます． ⇒ 彼は仙台へ行きます．そして，彼女は名古屋へ行きます．

長文分割 長文は適当なところで区切って接続詞でつなげる．あるいは箇条書化する．

- すぐに返事するつもりでしたが，論文の執筆に追われていたため，返事が遅くなってしまいました． ⇒ すぐに返事するつもりでした．しかしながら，論文の執筆に追われていた．そのため，返事が遅くなってしまいました．

係り受け制御 制御用の括弧で括る．

- 事故のため，東名を走行中の車に若干の遅れが出ていま ⇒ 事故のため，[東名を走行中の車に若干の遅れが出ています]．

訳語指定 指定された特定の訳語を与える．

このような処理を行なうための文の書き換えシステムは次のようなものである．

パターンマッチングに基づく文書き換えシステム

本システムはパターンマッチングと書き換えを行なう規則を記述する構造変換部と，各構造変換部の実行を制御する手続き部の二重構造になっていて，記述された規則にしたがって，入力の形態素リストが変換される．

本システムからは機械翻訳実行ルーチンを呼び出せるようにしてあり，文の明確化処理（長文分割，省略語処理，係り受け解析など）などの文の書き換えを行なうとともに，機械翻訳の呼び出し，翻訳失敗時処理，機械翻訳のバイパスなどの翻訳制御にも利用している．

4.4 文書整形

これまでの多くの機械翻訳システムでは，原文と翻訳結果が一文単位で縦あるいは横に対応づけられて表示している．これは文を一文ずつ編集するには都合が良い．しかし，翻訳結果の利用目的（大意をつかむのか，後編集をするのか），翻訳の品質（原文をどれくらい参照しなければならないかに関わる），利用者の能力（原文・翻訳文の言語に対してどのくらいの能力があるか）によって，とるべき表示形式は異なるはずである．

例えば，現在の機械翻訳の品質で，文書全体の大意をつかむという使い方からすれば，原文とのリンクがある程度保存されたまま元のレイアウトに準拠して表示するのがよい．複数のウィンドウを利用するなどのハードウェアやソフトウェアに依存した方法もあるが，パラグラフ単位に元のレイアウトを意識した図2のような表示形式もある．

このような様々な表示形式の要求に対して，翻訳部は単に翻訳結果だけを出力する．次に，レイアウト情報のついた原文と翻訳結果がマージされて対訳文書が作成される．それから，指定された書式に変換される．

対訳文書作成

レイアウト構造情報付き原文と翻訳結果とから対訳文書を作成する．このとき，原文のレイアウト構造情報と原文と翻訳結果との対応関係がはっきりわかるような形式で保存される．このデータは次の印刷形式作成システムによって印刷用の形式に変換される．

印刷形式作成システム

翻訳結果だけを出力したり，対訳形式にしたり，パラグラフ単位出力にするなどの出力形式を整える．その結果は文書整形システムによって表示あるいは印刷される．

5 翻訳処理部分の姿

このような翻訳処理の周辺処理を充実させることを考えると，機械翻訳の本体としては以下のようにあるべきである．

5.1 簡素であれ

現在の翻訳処理の文法規則では，ユーザからの様々な要求に答えるべく，その中で様々な処理をしている．しかし，中には明らかにオーバデザインなものもある．翻訳処理本体の文法規則はもっと単純なものにして，一般的な文の精度の向上に努めるべきである．

以下に翻訳処理をもっと簡素化すべき例をあげる．

修飾記号の分離 実際の文書中の各文には，箇条書きのための記号（・，o，-，1.，a) など），罫線片，特殊な句読点（!!，:-) など），引用記号などがついていることがある．これらは，文の切りだし時に分離あるいは標準化し，翻訳本体には文そのものだけを与え，これらの記号類は翻訳結果に対して復元するというようにすべきである．

変換不要部分 文内に現れる人名，組織名などの固有名詞，日付，数量表現，記号列などは翻訳（変換）する必要はない．例えば，「10:30」を「10時半」と訳したり，「平成3年」を「1991」としたりすることは翻訳本体内ではすべきではない．

これらのものはその分野や文化に依存するものであって，これらを正しく翻訳するには背景となる知識が必要である．あらゆる背景知識を機械翻訳に与えることは不可能である．中途半端に解釈をした結果を出力することは，誤った情報を伝えてしまうかもしれない．翻訳本体としては，このようなものは何も処理をすべきでは

"FOURTH INTERNATIONAL SYMPOSIUM ON ARTIFICIAL INTELLIGENCE"

第4 人工の知能の国際的シンポジウム

November 13-15, 1991

1991年11月13-15日

The Symposium is sponsored by the ITESM (Instituto Tecnologico y de Estudios Superiores de Monterrey) and supported by the IJCAII, in cooperation with the AAAI,the Canadian Society for Computational Studies of Intelligence, the IAKE, the Sociedad Mexicana de IA and IBM of Mexico.

シンポジウムはAAAI(メキシコの知能とIAKEとSociedad Mexicana de IAとIBMのコンピュータの研究のためのカナダの社会)と共同でITESM(Instituto Tecnologico y de Estudios Superiores de Monterrey)によって主催されて,IJCAIIによって支持される.

図 2: ニュース表示

ない. 必要ならば, 前編集あるいは後編集として, 単位系の変換, 記号列の解釈, 固有名詞の読みや漢字への変換をおこなばよい.

予測可能な解釈 前編集を正しく行なうためには, 原文をどのように修正したら何が起こるかが予測可能でなくてはならない. ある文では正しく翻訳できたのに, 利用者からするとほとんど同じに見える文がほんのちょっとした辞書などの設定の違いにより, 翻訳結果が大きくことなるとすると, 何を根拠に前編集したらよいのかがわからなくなって, 非常に使いづらくなってしまう. 処理の原理が利用者にもよくわかるようにすべきである. 例えば, 係り受けに関しては強い手がかりもないならば一番近いところにかかるというような単純であるべきである. 分野に依存したヒューリスティックス規則は文の明確化処理として, 翻訳本体の外側で行なうべきであろう.

失敗処理 最初の翻訳処理が失敗したときに, なんとかして結果を出そうとするシステムもある. 確かに情報獲得のような場面では何らかの結果がでて欲しい. しかし, このような場合, 翻訳処理本体で無理矢理に解釈するより, 失敗処理は外側にだして処理すべきであろう.

5.2 外部からの指示できるように

機械翻訳の外で解析して得られた情報, あるいはユーザが与えた情報は, 機械翻訳本体に的確に伝えることができるようになっている必要がある. 訳語の指定, 係り受けの制御, 形態素に対して品詞などの文法属性・意味属性などが指示できるべきである.

5.3 翻訳情報の提示

結果として欲しい情報は翻訳結果の文字列だけではない. 訳文の単語と原文の単語との対応関係であるとか, 単語が専門用語辞書であるかどうかといった情報が必要なことがある. これらの情報を含んだ形態素リストのまま出力を行なうことが必要である.

6 今後の機械翻訳に向けて

主に後編集を必要としない情報獲得やコミュニケーションの支援の立場から統合的な翻訳環境について見てきたが，後編集を必要とする環境についても若干考察する．

これまでの機械翻訳は，翻訳結果に対して後編集が行なわれることが多かったが，あまり後編集を効率良く行なえるようにはなっていない．機械翻訳を利用した翻訳作業を効率化するためには，後編集作業の効率化が緊急の課題である．

そのためには以下の点についての検討が必要である．

- 訳文チェックにかかる時間の減少

  どの部分の翻訳結果が確実であり，どこが不確実なのかをを評価・表示することが必要である．

  現在の機械翻訳の結果は正しい翻訳もあれば誤った翻訳を出力することもある．単語レベルで考えても，専門用語辞書から取り出された正しい単語の場合もあれば，専門用語辞書になくて，基本語を組み合わせたいい加減な単語になっている場合もある．この区別がつかないため，後編集者は，機械翻訳結果が正しいものか誤ったものであるかを判定しなければならない．翻訳結果に対する信頼性が低い場合には，原語のままを表示するとか，表示スタイル（色，フォントなど）を変えるという工夫が必要である．

- 翻訳結果の修正の手間の減少

  例えば，訳語の選択，翻訳結果の変形（2文の結合・分割，名詞の代名詞化など）などを支援することによって，翻訳結果の修正にかかる手間の削減の支援が必要である．このうち，ある種の統一的な翻訳結果の変換に関しては，本稿で述べた文の書き換えシステムが有効であろう．

後編集を前提とする翻訳システムを考えた時，正しい翻訳のみを与えるという観点から，現在のような分析型の機械翻訳システムとは別に，用例に基づいた機械翻訳システムも考える必要があるだろう [3]．現在，我々は定型的な文に対しては，機械翻訳のバイパス処理を導入している．この方式では，ユーザは非常に簡単に登録可能であり，一度教え込んだ文に対しては，文脈依存の翻訳結果をださなければならないというごくわずかの例外を除いて，正しく翻訳される．これをさらに進めて，登録された文のみでなく，その文から一部を変数化して得られる同種の文についても正しく翻訳することができるようにすることができる．すなわち，多くの用例を与えることによって翻訳を実現する用例に基づく機械翻訳システムに近づいてくる．用例に基づく方式だけで多くの自然言語現象をカバーしたシステムを作成するには，非常に多くの用例を用意しなければならず，現在の分析型のシステムにすぐにとって変わるというものではない．当面は，まず用例に基づいて翻訳を試みて，ある信頼値以上での翻訳結果が得られなかった場合には従来どおり分析型の機械翻訳を用いるという方式をとることで，両者が共存して使われていくことになるであろう．

7 おわりに

本稿では，機械翻訳を実際に利用するための我々のアプローチの仕方と，今後の展望について述べた．機械翻訳を実際に使いものになるようにしていくには，それぞれの応用に対して使い込んで，チューニングをほどこしていくことである．現在我々は，日々流れてくる実際の英文の電子ニュースの記事を日本語に翻訳するサービスとしての運用実験などを行なうことによって，改良を進めているところである．

参考文献

[1] 坂本，有賀：Muプロジェクトにおける総合システムの基本設計，自然言語処理研究会 46-6 (1984)

[2] 牧野，小関：統合自動翻訳システム PIVOT，bit 別冊 機械翻訳，pp. 184-190 (1988)

[3] Nagao, M., *A Framework of a Mechanical Translation between Japanese and English by Analogy Principle*, in ARTIFICIAL AND HUMAN INTELLIGENCE(A. Elithorn and R. Banerji, Eds.), Elsevier Science Publishers, pp173-180 (1984)