Automatic Rewriting Method for Internal Expressions in Japanese to English MT and It's Effects

Satoshi Shirai [+], Satoru Ikehara [++], Akio Yokoo [+++], Yoshifumi Oyama [+]

[+] NTT Comunication Science Laboratories
Hikarinooka−1, Yokosuka−shi, 239 Japan 〈{shirai, oyama}@cslab.kecl.ntt.co.jp〉
[++] Faculty of Engineering, Tottori University
Minami 4−101, Tottori−shi, 680 Japan 〈ikehara@ike.tottori−u.ac.jp〉
[+++] ATR Interpreting Telecommunication Laboratories,
Hikaridai2−2, Seika−chou, Kyouto−fu, 619−02 Japan 〈ayokoo@itl.atr.co.jp〉

Abstract

近年多くの機械翻訳システムが研究開発されてきたが、訳文の品質は必ずしも満足できるレベルには達していない。そこで、機械翻訳の現場では、翻訳システムの能力を引き出す方法の一つとして、原文の前編集が広く行われてきた。これは原文を人手でによって翻訳し易い文に書き替えるものであり、機械翻訳の省力化効果を損なうものである。自動化が望まれていたが、書き替えてよいか否かを自動的に判断することは困難であった。これに対して、すでに我々は、「最近開発された意味解析技術を用いれば、副作用の無い自動書き替えを実現することができる」こと、また、それだけでなく、「原文前編集ではできなかったような種類の書き替えまでが可能である」ことを示し、「原文自動書き替え型」の翻訳方式を提案した。

本論文では、この方法を更に発展させ、より汎用的な「内部表現自動書き替え型」の翻訳方式を提案した。すなわち、翻訳処理の途中結果を検証しながら必要に応じて割り込んで解析結果を書き替え、正しい翻訳結果を導くための仕組みとして実現する方法を示した。

具体的には、まず、「原文自動書き替え方式」の書き替え規則を拡張し、機能的な再分類を行った。次に、書き替え規則が７種類の基本機能の組合わせで実行できることを示し、翻訳の処理フェーズに合わせて書き替え規則を配置する方法を示した。最後に、２種類の新聞記事翻訳に適用し、必要な書き替え規則の数とその述べ使用回数などを実験的に評価した。その結果によれば、新聞記事の翻訳では、書き替えに必要な規則は、分野によっては、比較的少数（数百文オーダ）の原文を解析することによって網羅的に収集できること、書き替え規則は２文に一回以上の割合で適用され、訳文合格率を文平均 20 %向上させる効果を持つことなどが明らかとなった。

1. Introduction

日本では、ここ 10 数年間にわたって多くの機械翻訳システムが開発され商品化されてきた（文献）。その中で、英日翻訳は、ほぼ実用的な品質が達成されてきたのに対して、

日英翻訳は，まだ実用的な翻訳品質は達成されていない。これにはいくつかの原因が考えられる。その一つは、表現体系の違いである。すなわち、英語は表現構造が表現内容と比較的独立であるため、文法的知識を用いた解析が効果的であるのに対して、日本語では表現構造と表現内容の相互依存性が高いため、文法的知識だけで精度よい解析を実現することが困難なことである。

また、第２の原因としては、日本語と英語の間の文化的、歴史的事情の違いが考えられる。自然言語は、話者の概念や思想を聞き手に伝えるための手段であるが、同時に、我々人間の思考の手段でもある。言語の違いは文化の違いでもあり、文化の違いによる思考方法の違いは言語表現の違いでもある。特に、言語族の異なる言語間での発想法の違いは、機械翻訳を困難にしている。日本語（膠着言語）と英語（屈折言語）を考えた場合、日本語は、明治時代以降、欧米の文化や思想を取り入れる過程で、従来の日本語に無い概念や表現の仕組みを発達させながら、欧米言語から日本語への翻訳技術を確立してきた。これに対して、英語では、日本語が持つ固有の概念や表現法の多くは必ずしも取り入れられてきたとは言えない。このことも、英日機械翻訳に比べて日英機械翻訳を困難にしている原因の一つと考えられる。

ところで、これらの問題を解決するには、第１に、翻訳対象文書の表現の性質に翻訳システムの機能を合わせる方法、第２に、既存の翻訳システムそれぞれの性質に翻訳すべき文章を合わせる方法が考えられる。

第１の方法としては、用例翻訳の研究（文献）や知識ベース型の翻訳（文献）がある。用例翻訳は、従来の要素合成法の限界を超えることを狙い、対訳コーパスを用いて原言語と目的言語の表現をより大きな単位で対応づけるものであるが、表現の網羅性を保証することが困難である。また、知識ベース型翻訳は、常識知識や世界知識など膨大な量の知識を必要とすると考えられ、その収集とデータベース化に問題がある。

第２の方法としては、古くから、制限言語（文献）や原文前編集（文献）の方法が試みられている。このうち、制限言語は、文章作成の段階に適用されるもので、自然言語の最大の特徴である発想の自由が阻害されるため、利用者に受け入れられてこなかった。これに対して、原文前編集は、翻訳結果を確認しながら編集方法が習得でき、慣れるに従って機械翻訳の能力を引き出せるようになるため、翻訳の現場でもある程度受け入れられてきた。しかし、原文前編集は、多くの人手コストを必要とするため、その自動化が望まれてきた。

ところで、前編集では、字面上は同じ表現でも前後関係によって、書き替えてよい場合と書き替えてはいけない場合がある。それを無視して書き替えを行うと、書き替えによってむしろ訳文品質は低下するようになる。このように、書き替え可否の判断は、前後の意味的な関係に依存するため、前編集作業を自動化することは困難であった。

これに対して、筆者等は、「最近開発された意味解析技術（文献）を用いれば、副作用の無い自動書き替えを実現することができ、原文前編集ではできなかったような種類の書き替えまでが可能であること」、また、「それによって、大幅な訳文品質向上効果が得られること」を示し、「原文自動書き替え型翻訳方式」を提案した（文献）。ところで、この方法は、入力原文に対する構文解析結果を書き替えるものであるが、解析の様々な段階で必要に応じて適用できるようにすれば、原文の解析精度の向上などにも効果が期待でき

ると考えられる。また、そのためには、書き替え機能を基本機能に分割し、翻訳処理のフェーズと独立に書き替え規則の定義を行うことなどが必要と考えられる。そこで、これらに着目して、本論文では、約 1,000 文の新聞記事分析から得られた 940 種類の書き替え規則を分析し、様々な書き替え規則が 7 種類の基本機能の組合わせで表現できること、また、それにより、個々の書き替え規則と独立した処理システムが容易に構成できることなどを示す。また、その結果に基づいて、「内部表現自動書き替え型翻訳方式」を提案する。

2. Outline of the Automatic Rewriting Method

本章では、すでに我々が提案した「原文自動書き替え方式」の基本的な枠組みについて述べる。

2.1 Framework of the Automatic Rewriting Method

原文書き替えを自動化しようとする場合、書き替えの悪い副作用をいかに防止すえるかが最大の問題となる。人手による原文書き替えでは、書き替えられる文は特定されており、他の文への副作用はない。これに対して、自動書き替えの場合は、登録した書き替え規則は該当する表現の全てに適用されるため、書き替えてはならないものまで書き替えてしまう可能性がある。

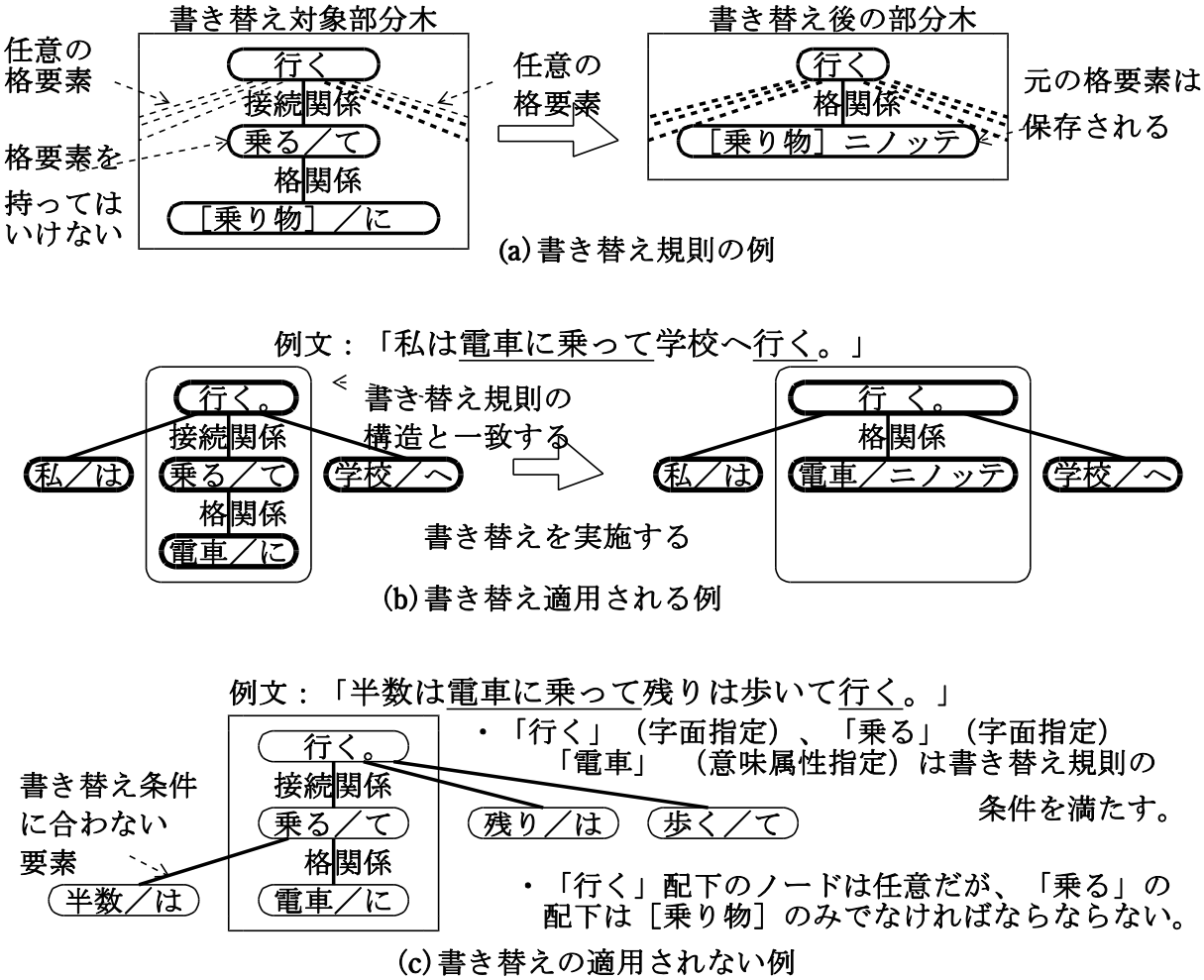


Fig.1 Conditions for Applying Rewriting Rules

特に、原文の段階での書き替えでは、書き替え対象は字面表記で指定されることになるため、字面が一致した表現はすべて書き替えられてしまう。このような書き替えの副作用は、書き替え規則の整備が進むにつれて増大し、全体の訳文品質は逆に低下するようになる。

この問題を解決するため、我々は、書き替え規則の適用条件の記述では、「字面、単語や表現の文法的属性のほか、単語の意味的属性として ALT−J/E （Automatic Language Translator−Japanese to English）の単語意味属性（3,000 種）を使用すること」、また、「単語と構文解釈の候補が出そろった時点で、書き替え規則を適用する」ことを提案した。

本方式による書き替えの例を図１に示す。この例では、書き替え規則は(a)のような構文木で示される。書き替えが適用される条件は、まず、①「乗り物（意味属性指定）」が「乗る（字面指定）」に対して格関係にあること、②「乗る」が「行く（字面指定）」に対して接続関係にあることであるが、同時に、③「行く」に対しては、任意の数の要素との係り受け関係を持ってもよいが、④「乗る」に対しては、「乗り物」以外の係り受けを持ってはならないこともが条件となっている。これによって、「〜に乗って〜行く」の表現でも、(b)、(c)に示すように、書き替えてよい場合と書き替えてはならない場合が識別される。

2.2 Expressions to be Rewritten

書き替えの対象となる表現は、原言語の表現に直接対応する目的言語の表現がないため、話者の意図を判断して、言い直さなければならないもの、慣習の違いなどにより、訳す必要の無いものなど、高度な意訳等が必要で現状では翻訳困難である表現である。また同時に、既存の翻訳技術で翻訳できる範囲であるが、たまたま使用しているシステムでは能力が足りない表現も書き替えの対象となる。

このように書き替えの対象となる表現は、従来の人手による前編集の対象表現と同等と見えるが、機械翻訳の場合は、人手の前編集と異なる点がある。すなわち、人手による前編集の場合は、原言語内に意味を変えない別の表現が存在しなければ、書き替えはできないのに対して、翻訳システム内部で表現を書き替える場合は、原言語内で適切な言い換えができなくても、目的言語に意味的に対応する適切な表現があれば、それにダイレクトに置き換えてしまうことができる。

このことに着目して、我々は、書き替え対象表現を以下の２つのタイプに分け、合計６種類の表現の書き替えを提案した。

(a) Rewriting within the Source Language

着目する表現に対して、当システムで翻訳可能な別の原言語表現のある場合に書き替えを行うもので、①縮約展開型の書き替え、②冗長除去型の書き替え、③構文組み替え型の書き替えの３種類の書き替えを提案した。これらは、原言語内での書き替えであるため、書き替え後の文は、原言語の文としても意味の分かる文となる。ただし、人手による前編集と同様、翻訳システムに合わせた書き替えであり、必ずしも原言語の表現として適切になるとは保証されない。

(b) Rewriting into Pseudo Source Language

別の原言語表現はないが、部分的に対応する目的言語表現のある場合に書き替えを行うもので、①独立句的表現の書き替え、②様相・時制表現の書き替え、③接続表現の書き替えを提案した。これらは、目的言語固有の表現に対応づける書き替えであり、書き替えた後の文は、必ずしも原言語の文として意味が通じる必要はない。

3. Classification of the Expressions to be Rewritten

3.1 Extention of Automatic Rewriting Method

前章で述べた書き替え方式では、一つの表現部分に２つ以上の書き替え規則が重複して適用されると予想外の書き替えが行われ副作用が生じる危険性があるため、１つの表現に適用される書き替え規則は１つに制限されていた。このため、複数の書き替え規則が適用されるような表現にたいしては、複数の規則を合成した別の規則を作成することが必要であった。

これに対して、ここでは、規則の汎用性を向上させるため、複数の書き替え規則が適用される場合も段階的に一つづつを適用していけばよいように、書き替え規則を機能別に分類する。また、書き替え処理が書き替え対象表現を検定する処理と新しい表現生成の処理から構成されることに着目して、適用範囲を原文解析結果の誤り修正や複合表現の抽出にまで拡張する。

(1) Phased Application of Rewriting Rules

まず、一つの表現に複数の書き替え規則が適用される例として、下記の文を考える。

例文：「本処理の結果を用いて変換および生成する」

この例文には、以下の２つの書き替え規則が適用される。

①「を用いて」の句を助詞相当語として、英語の"based　on"に相当する擬似日本語に置き換える。
②「変換および生成する」は２つの動詞の並びの一方が縮約され名詞となった表現であるので､縮約を展開し、「変換し､そして生成する」に書き替える。

ここで、書き替え前の「を用いて→生成する」は述語間の係り受け関係を持つが、擬似日本語の"based　on"への書き替え後は、格関係に変化する。従って、①の規則の適用条件の記述は、「生成する」に対応する文節の範囲までを含むことになり、②のルールの適用範囲と重なっている。しかし、このような性質の異なる書き替えは、２回以上に別けても副作用の心配は殆ど無い。むしろ、分けることによって規則の汎用性を高めることができる。

(2) Application to the Result of Source Text Analysis

形態素解析や構文解析では、通常、それぞれの解析のために用意された規則集が使用されるが、言語では、一般的な解析規則が適用できないような様々な例外的な現象が存在す

る。特に、特定の語がもつ特殊な振舞いの規則化は容易でない。実際の言語解析では、そのような表現が解析に失敗する部分の大半を占めており、解析精度の限界の主要な原因となっている。

しかし、このように解析に失敗する表現を集めてみると、解析誤りには一定の傾向と特徴があり、誤りが一定のパターンとして捉えられることが分かる。このことに着目して、ここでは、形態素解析、構文解析の中に含まれる誤りのパターンを発見してそれを正しい結果に書き替えることを提案する。

3.2 Classification of Elemental Functions included in Rewriting Rules

前節の考え方にしたがって、適用順序を考えながら書き替え規則を機能的に分類する。まず、解析誤りの発見と修正のための書き替えは、処理の性質上、形態素解析と構文解析の終了直後に実施するととする。引き続き、２章で述べた書き替えを実行する。ところで、２章では、書き替え規則を「日本語内の書き替え」と「擬似日本語への書き替え」の２つに分けた。前者は、日本語の意味解析を助けるための書き替えと見なされるのに対して、後者は、目的言語への変換を助けるための書き替えと見なすことができる。また、後者のうち、「独立句の書き替え」は、訳語を想定して原文表現を固定的に捉えることを狙っているのに対して、「様相時制表現」や「接続表現」の書き替えは、要素合成法が不可能な表現を抽出するものであると言える。これらは処理の内容が独立しており、相互干渉の心配はないと見られるので、それぞれ独立に考える。

以上から、書き替え規則を大きく４つのグループに分類し、その内部をさらに合計１４種類に分類する。以下にその詳細を示す。

(1) Rewriting as Post Processing for Analysis

①形態素補正

形態素解析では，下記の例に示すように、長いひらがな列や接辞を伴う複合語の解析で失敗することが多い。下記の２番目の例は、複雑な複合語解析に備えて接辞処理を強化したときに生じる副作用の一つである。これらの誤りをパターンとして登録し、修正する。

＜誤＞～し(動詞)/た(助動詞)/いも(名詞)/の(助詞)/だ(助動詞)

＜正＞～し(動詞)/たい(助動詞)/もの(抽象名詞)/だ(助動詞)

＜誤＞現(接頭辞)/代用(名詞)/語(接尾辞)

＜正＞現代(名詞)/用語(名詞)

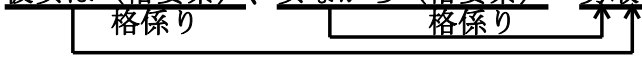
②形態素多義絞り込み

例えば、形態素解析では、かな書きされた動詞「なる」に対して「成る」、「鳴る」、「生る」などの解釈が生成されるが、どの解釈が正しいか判定できないまま出力され、選択は後の意味解析に任せられることがある。このような時、隣接する文節の情報を手がかりに決定できる範囲での候補絞り込みを行い、意味解析の負担を軽減する。今迄の実験では、この種の書き替えによって、形態素解析で残された単語当たりの多義は、平均 2.15 から平均 1.15 まで絞り込めることが分かった。

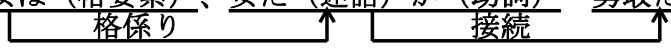
③係り受け補正

例えば、下記の例では「名詞＋ながら」の名詞句は述語の働きをするが、解析では名詞句として解釈され解釈されるため、係り受け解釈が正しく実行されない。このような場合、単語の解釈を変えるだけでなく、係り受け関係も修正する。

＜書替前＞彼女は（格要素）、女ながら（格要素）　勇敢だ（述部）。



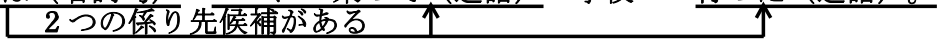
＜書替後＞彼女は（格要素）、女だ（述語）が（助詞）　勇敢だ（述部）。



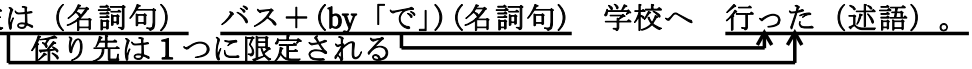
④係り受け多義絞り込み

例に示すように、係り受け先の候補の数が減少するような書き替えを行う。

＜書替前＞彼は（名詞句）　バスに　乗って（述語）　学校へ　行った（述語）。



＜書替後＞彼は（名詞句）　バス＋(by「で」)(名詞句)　学校へ　行った（述語）。



(2) Rewriting within Source Language

⑤縮約表現の展開

活用語尾の省略や並列名詞句の一部省略などを元の省略のない表現に書き替える。

＜書替前＞変換//および//生成する

＜書替後＞変換し//そして//生成する

⑥冗長表現の圧縮

英語に訳出不能と見られるニュアンスなどは取り除き、より平坦な表現に書き替える。

＜書替前＞男/も//いれ/ば//女/も//いる

＜書替後＞男/も//女/も//いる

⑦構文組み替え

日本語独特の構造の表現で、直訳困難な表現の構造を書き替える。

＜書替前＞二/機種//合わせ/て//月/百/台/生産する

＜書替後＞二/機種/の//月産/は//百台/だ

⑧敬語の標準化

日本語の敬語表現は複雑で、それをそのまま訳すことは適切でない場合が多い。そのような場合、フラットな表現に書き替えた上で、表現に丁寧な英文を生成するための印を付与する。

＜書替前＞お(接頭語)/読み(連用形名詞)/に(助詞)//なる（動詞）

＜書替後＞読む(動詞＋尊敬)

(3) Rewritin into Pseudo Source Language

⑨助詞相当語書き替え

第２章の図１の例のように、英語に訳したとき助詞１語に置き換えられるような表現を対象とする。

⑩副詞相当語書き替え

独立不定詞句（「言うまでもなく→ nwwdless to say」）や非人称独立分詞（「一般的に言えば」→ generally speaking）など直訳すれば節になるような表現を句に書き替える。

⑪連体詞相当語書き替え

「印象に残る」→ impressive、「喜びにあふれた」→ joyful などのように、日本語の修飾句を英語形容詞 1 単語に置き換える。

⑫フレーズ書き替え

上記の書き替えと類似であるが、使用条件によって表現の一部が変化する場合の書き替え規則が該当する。例えば、「驚いたことには」→ to one's surprize では、one's の部分が使用条件によって変化する

(4) Rewriting for Freezing of Subjective Expressions

⑬接続様相時制表現書き替え

接続、様相、時制が結合した表現を英語側から見た意味に書き替える。

＜書替前＞～する(動詞)/なら//～する(動詞)/のに

＜書替後＞～する(動詞＋仮定法)//～する(動詞→ would V)

⑭様相時制表現書き替え

様相、時制が結合した表現を英語側から見た意味に書き替える。

＜書替前＞～し/た/ようだっ/た

＜書替後＞～する(動詞→ seemed tohave V(過去分詞))

4. Structures of Rewriting Functions and their Compornents

約 1,000 文の新聞記事の翻訳実験に基づき、940 種類の書き替え規則を作成し、それを前章で述べた基準にしたがって分類し、各書き替え規則も持つべき基本機能を分析した。その結果、従来の方法は、書き替え規則の適用条件記述方式に問題のあることが分かった。また、書き替えのための機能は 7 つの基本機能から構成されることなどが分かった。そこで、本章では、規則適用条件の記述方式の改良と書き替え基本機能の構成に着目した処理プログラムの構成法について提案する。

4.1 Augmentation of Description Ability for Rewriting Conditions

書き替え規則は、基本的に条件指定部と書き替え実行部から構成される。このうち、条件指定部では、単語の字面、品詞・活用形などの文法的属性、意味属性（3,000 種）のいずれをも用いて書き替え対象表現を記述することができるが、従来の方法では、以下の問題が指摘できる。

①日本語は、漢字書き、かな書き、送り仮名などの揺らぎが多い。これに対して、書き替え規則では、適用条件指定を厳密にし、副作用を防止するため、原則として、書き替え対象となる表現の前後に存在する単語や文節の条件も合わせて規定することが必要である。従って、適用条件の指定が表現の揺らぎに弱いことが問題となる。

②いくつかの複合した条件指定を行うとき、従来の条件指定では、and 条件と or 条件の解釈が固定されており、また、適用例外条件の指定法が一様でない。

③指定可能な単語属性の制約と指定方法に不統一性がある。

そこで、①の問題を解決するため、ルール記述のための標準表記を設け、表現の揺らぎは、標準表記に変換した後に、適用条件判定を行うこととする。また、②の問題を解決するため、AND/OR 条件指定を自由化するとともに、例外条件指定における OR 条件指定の簡素化する。また、③では、記述可能な単語属性を拡大し、体系化を図ること、記述の書式をルール種別間で共通となるよう統一を図った。

4.2 Elemental Functions for Execution of Rewrting

すでに作成した 940 種類の書き替え規則の分析結果によれば、書き替え実行機能は、以下の 7 種の基本機能に分解できる。

①連結：複数の単語を 1 単語にする。
②変更：各種の単語属性を変更する。又は新たに付与する。
③削除：単語、文節などを削除する。
④補完：単語、文節などを挿入する。
⑤分割：１単語を複数単語に分割する。
⑥入換：単語や文節の順序を変更する。
⑦評価：条件を満たす表現の優劣を判定する。

前述の 940 種類の書き替え規則の実行部がどのような基本機能から構成されるかを表 1 に示す。

表 1. 書き替え規則実行部の基本機能構成

分 類	基本機能 / 書き替え規則	連結	変更	削除	補完	分割	入換	評価	規則数
解析後処理	形態素補正	25	22	2	4	13			60
	形態素多義絞り込み		1					15	16
	係り受け補正		19		1				19
	係り受け多義絞り込み		1					13	14
日本語内書き替え	縮約展開		2	8	8				10
	冗長圧縮	21	17	33				1	55
	構文組み替え	10	20		8		12		41
	敬語の標準化			1					1
擬似的日本語への書き替え	助詞相当語	159	27	2	12	2			182
	副詞相当語	90	22	2	1				111
	連体詞相当語	9	7						15
	フレーズ	2							2
主体的表現の固定化	接続様相時制	12	11	1					22
	様相時制	74	39	5	2	3	1		114
その他（未分類）									278
合　計									940

4.3 Control of Rewriting

前節で書き替え実行部は７つの基本機能から構成されることを述べた。これに着目すれば、実行部のプログラムは、７つの独立した基本モジュールとそれを制御するプログラムで構成できることが分かる。

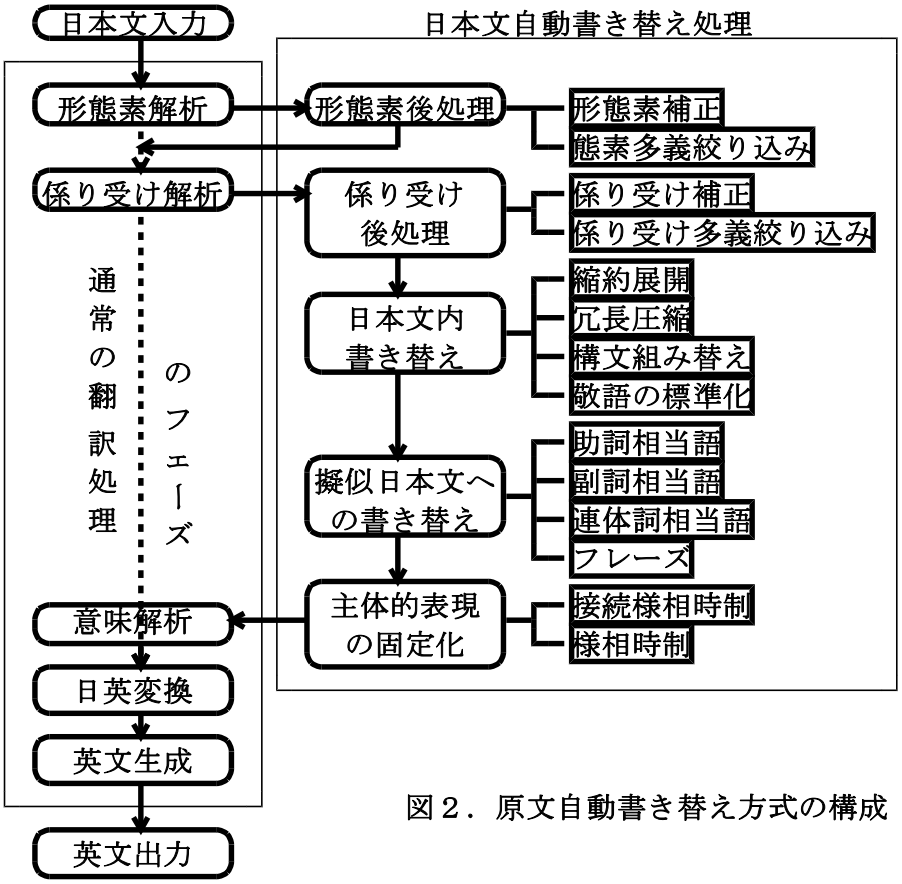


図２．原文自動書き替え方式の構成

5. Experimental Result of Newspaper Translations

5.1 Number of Times for Application of Writing Rules

(1) Conditon of Experiments

実際の機械翻訳での書き替え規則の適用状況を調べるため、機械翻訳ｼｽﾃﾑ ALT−J/E の書き替え処理を組み込み、以下の 2 種類の新聞記事を対象に翻訳実験を行った。

（１）新聞記事一般文

日本経済新聞（1994 年 8 ～ 9 月）の一般記事 874 文

（２）新聞記事市況速報文

日経新聞社テレコンＢＩＺ（1995 年 7 月）の 84 記事、546 文

実験では、翻訳実験と書き替え規則の作成を数回繰り返し、必要と見られる書き替え規則を整備した後、翻訳で使用された書き替え規則数の異なり数と累積数を求めた。

(2) Number of of Application Times for Rewriting Rules

翻訳実験で作成された書き替え規則の適用回数を表 2 に示す。この表から以下のことが

分かる。

1）一般記事へ適用された規則は、合計 162 ルールで、述べ適用回数は 463 回であるのに対して、市況速報への適用規則数は、43 ルール、述べ適用回数は 337 回である。一文当たりの適用回数は、前者は、0.53 回/文、後者は、0.62 回／文で、共に、2 文に 1 回以上の書き替えが行われている。

2）適用回数の多い規則は、形態素解析の後処理に関するもの、助詞相当語の書き替えであるが、係り受け補正や構文組み替えなど、構文を変更するものもかなり使われている。

3）一般記事に比べて市況速報では、少ない数のルールを何回も使う傾向が見られる。また、特に、市況速報では、主体的表現に関する書き替え規則がまったく使用されていない。これは、市況速報では、専門的で固定的な表現が多用されること、また、主観的な表現は余り使われていないことを意味するものと考えられる。

なお、この実験では、書き替え規則の整備により、訳文合格率は書き替え規則適用前に比べて、いずれも約 20 %向上した。

表２．適用されたルール数と適用回数

分 類	適用対象 / 書き替え規則	一般記事		市況速報記事	
		異なり	累積	異なり	累積
解析後処理	形態素補正	３２	６６	９	１０７
	形態素多義絞り込み	６	１２９	３	３９
	係り受け補正	２	２３	１	１
	係り受け多義絞り込み	２	６	０	０
	小　計	４２	２２４	１３	１４７
日本語内書き替え	縮約展開	１	１	１	１４
	冗長圧縮	５	５	１	１４
	構文組み替え	１３	２８	４	２２
	敬語の標準化	０	０	０	０
	小　計	１９	３４	６	５０
擬似的日本語への書き替え	助詞相当語	５８	１２５	１６	１２０
	副詞相当語	２５	３２	８	２０
	連体詞相当語	７	９	０	０
	フレーズ	０	０	０	０
	小　計	９０	１６６	２４	１４０
主体的表現の固定化	接続様相時制	１	１	０	０
	様相時制	１０	３８	０	０
	小　計	１１	３９	０	０
合　計		１６２	４６３	４３	３３７

(3) Rewriting Rules with high freequency in use

上記に述べた２つの実験で、適用回数の多い規則の上位 10 件の内容を表２に示す。

この表から、使用頻度の高い規則は、一般記事と市況速報では際立った偏りがあることが分かる。しかし、偏っているといっても、一方の分野で使用頻度の高い規則は、他方の分野でも若干の適用回数があり、書き替え規則の充実は、機械翻訳の総合的な品質向上に役立つと期待される。

表２．適用頻度の高いルールの例

	書き替え規則の内容	一般	市況	合 計
1	「～して」の多義を絞る	８４	１５	９９
2	文頭の「反面、～」を副詞化する	1	８２	８３
3	副助詞「は」の係り先を修正する	２１	1	２２
4	「～振りに」を for the first time に変える	2	１６	１２
5	「～円台」「～%台」の台を削除する	1	１４	１５
6	「～にかけて」を toward にする	1	１２	１３
7	「～に対する」を１語化する	2	7	9
8	「やや」を条件に応じて slightly にする	1	6	7
9	「～に加え」を beside に変える	1	6	7
10	「～で」の格助詞解釈を排除する	6	1	7

5.2 Convergence of the Number of Rewriting Rules

書き替え規則の収束性を調べるため、一般記事と市況速報の双方について、適用されたルール数と異なりルール数の変化を調べた。その結果を図３に示す。この図から以下のことが分かる。

「適用されるルールの述べ数は、一般記事の場合と同様、市況速報の場合も翻訳文数に比例して増大するが、異なりルール数は、比較的少数の文で飽和傾向を示す。特に、市況速報では、400 文程度でほぼ飽和状態に達している。」

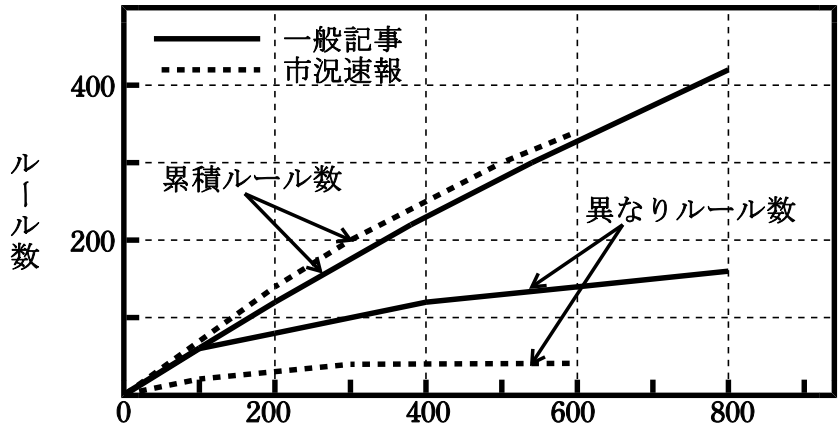


図３．書き替えに使用されたルール数

異なりルール数が速い収束傾向を持つことは、システム開発が容易であることを示している。今後、新聞記事以外の実験結果からも書き替え規則を作成していけば、汎用的な機械翻訳システムとしての訳文品質の向上が期待できると推測される。

6. Concluding Remarks

本論文では、品質の良い訳文結果が得られるよう翻訳処理過程を導くため、すでに提案されている「原文自動書き替え型翻訳方式」を発展させ、翻訳システムの処理過程に随時割り込んで解析結果を書き替える「内部表現自動書き替え型翻訳方式」を提案した。

具体的には、まず、すでに提案されている「原文書き替え規則」を解析結果を書き替え

る観点から見直すとともに拡張し、機能的な再分類を行った。次に、その結果に基づき、内部表現の書き替え規則が７種類の基本機能の組合わせで実行できることを示し、翻訳の処理フェーズに合わせて書き替え規則を配置する方法を示した。最後に、この方式を２種類の新聞記事翻訳に適用し、必要な書き替え規則の数とその述べ使用回数などを実験的に評価した。その結果によれば、新聞記事の翻訳では、書き替えに必要な規則は、分野によっては、比較的少数（数百文オーダ）の原文を解析することで網羅的に収集できること、書き替え規則は２文に一回以上の割合で適用され、訳文合格率を文平均 20 %向上させる効果を持つことなどが明らかとなった。

この仕組みを翻訳システムの一部として成長発展させるためには、書き替え規則を網羅的に収集登録することが必要である。しかし、この方式では内部表現を書き替えるため、書き替え規則の作成は、必ずしも容易でない。そこで、現在、素人でも容易に書き替え規則の作成ができるようルール作成支援システムを作成中である。ルール作成支援については、改めて報告したい。

Reference

Carbonell, J., et.al.(1992): JTEC Panel Report on "Machine Translation in Japan", Coodinated by Loyola College In Maryland (1992)

Furuse, O., Iida, H.(1992) : Cooperation between Transfer and Analysis in Example−Based Framework: COLING '92

池原,宮崎,白井,林(1987)：言語における話者の認識と多段翻訳方式, 情処論, Vol.28, No.12, pp.1269−1279

Ikehara, S.(1989) : Multi−Level Machine Translation Method, Future Computer Systems, Vol.2, No.3, pp.261−274

Ikehara, S.(1992) : Criteria for Evaluating the Linguistic Quality of Japanese to English MT System, MT Evaluation Workshop

Ikehara, S., Miyazaki, M., Yokoo, A.,(1993)：日英機械翻訳のための意味解析用の知識とその分解能, 情処論, Vol.34, No.8, pp.1692−1704

池原悟(1996)：機械翻訳の現状ー日英およびその他の言語ー, 情報の科学と技術, Vol.46, No.1, PP.26−33

加藤直人(1995)：定型パターンを含む文の機械翻訳手法,情報処理学会論文誌, Vol.36, No.9, pp.2081−2090

Matsuo, Y., Shirai, S., Yokoo, A. and Ikehara, S.(1994): Direct Parse Tree Translation in Cooperation with the Transfer Method, Proceedings of the International Conference on New Methods in Language Processing, pp.144−149 [NeMLaP]

長尾真(1983)「制限言語の試み」自然言語処理技術シンポジューム

Nagao(1984): A Framework of a Machine Translation between Japanese and English by Analogy Principle, in A. Elithorn & Banerji, ed. "Artificial and Intelligence", North Holland.

Nagao, M.(1985): 制限言語の提案:, 自然言語処理シンポジューム, 情報処理学会

長尾真(1985)「機械翻訳文の質の評価と言語の制限」, 情報処理, Vol.26. No.10, pp.1197−1202.

Nagao,M.(1992) : Some Rationales and Methodologies for Example−based Approach, Proc. of Workshop on Future Generation Natural Language Processing, UMIST, Manchester

長尾真(編)(1994)自然言語処理動向調査委員会:「自然言語処理これからの課題」

Nirenburg, S., Carbonell, J., Tomita, M. and Goodman, K.(1992) : Machine Translation; A Knowledge −Based Approach, Morgan Kaufmann Publishers

Nierenburg,. S (1993) : A Direction of MT Developments, Proceedings of the Fourth Machine Translation Summit, PP.189−193

野美山浩(1993)：事例の一般化による機械翻訳,情報処理学会論文誌, Vol.34, No.5, pp.

905−912
小倉，Bond，F.，池原(1995)：日英機械翻訳における副詞句翻訳の問題点について，言語処理学会第1回全大，PP.269−272
Rimon, M., McCord, M., Schwall, U. and Martinez, P.(1991) : Advances in Machine Translation Research in IBM, Proceedings of MT SUMMIT Ⅲ, pp.11−18
佐藤理史(1991)：MBT−2 実例に基づく翻訳における複数翻訳例の組み合わせ利用，人工知能学会誌, Vol.6, No.5, pp.861−871
佐藤理史(1992)：事例に基づく翻訳，情報処理, Vol.33, No.6, pp.673−681
Sirai, S., Ikehara, S., Kawaoka, T.(1993): Effects of Automatic Rewriting of Source Language within a Japanese to English MT System, Proceedings of the TMI−93, pp.226−239
白井，池原，河岡，中村 (1995a)：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，情報処理論文誌, Vol.36, No.1, pp.12−21
白井，池原，横尾，木村 (1995b)：階層的認識構造に着目した日本語従属節間の係り受け解析の方法とその精度，情報処理論文誌, Vol.36, No.10, pp.2353−2361
武田浩一，浦本直彦，那須川哲哉，荻野紫穂，堤泰次郎(1989)：知識ベースを利用した機械翻訳システム Shalt2,コンピュータソフトウエア, Vol.12,No.5,pp.22−32
宇津呂武仁，松本裕治，長尾真(1993)：二言語対訳コーパスからの動詞の格フレーム獲得，情報処理学会論文誌, Vol.34, No.5, pp.913−924
吉田将(1995)「日本語の規格化」情報処理学会自然言語処理研究会, NL51−4