

日英機械翻訳前編集における日本語文分割処理について

木村真理子 野村浩一 平川秀樹

(株) 東芝 研究開発センター

mari@isl.rdc.toshiba.co.jp, nomura@isl.rdc.toshiba.co.jp, hirakawa@isl.rdc.toshiba.co.jp

日英機械翻訳システムを用いた翻訳の質・作業効率を向上させる為には、前編集における長文分割が有効である。本報告では、実際のマニュアル文書より抽出した、長文の指摘と書き換え候補生成とに用いる知識について述べる。その知識を既開発の前編集支援ツールにインプリメントし実文のマニュアル文書で評価した結果、指摘に関しては適合率62%、再現率90%、書き換え候補生成に関しては、適合率63%、再現率100%を得ることができた。

キーワード: 機械翻訳, 前編集, 日本語文書き換え, 長文分割

Sentence Division at Preediting for Japanese to English Machine Translation

Mariko Kimura Kouichi Nomura Hideki Hirakawa

Toshiba Research and Development Center
1 Komukaitoshiba-cho, Saiwai-ku, Kawasaki 210, Japan

Dividing long and complex Japanese sentences at preediting is effective in improving translation quality and efficiency in Japanese to English machine translation system. Knowledge used in spotting long sentences requiring rewriting, and in generating substitutes for indicated long sentences is accumulated from actual technical manual sentences, and is implemented into the MT system with preediting support tool. An evaluation performed showed that the tool searched 90% of the desired sentences (with 62% of the total search being correct) and generated substitutes at all times when necessary (with 63% of the substitutes matching the expected expression).

Keywords: Machine Translation, Preediting, Japanese Rewriting, Sentence Division

1. はじめに

翻訳翻訳における前編集作業は、機械翻訳システムの翻訳率の向上を通して、後編集作業を含めた翻訳業務の総合的な効率向上を目的としている。この前編集作業の効率化のためには、機械翻訳システムにとって翻訳が困難な文章表現を効率良く検出し、翻訳結果が向上するような書き換え候補をユーザに提示できる機能が必要である。現在、我々は、この前編集支援のためのインタフェースならびに支援知識の開発を進めている[1]。

日本語文の前編集の対象となる表現には、「長文」、「係り受け関係の曖昧な表現」、「冗長な表現」など様々な項目がある[2],[3]。これらの内、「長文」に対する前編集は、翻訳処理結果に対する効果も大きく最も重要な項目である。長文の分割がシステムに及ぼす効果としては、分割部分の曖昧性の減少だけでなく、分割による局所化効果、すなわち、全体としての曖昧性の減少、部分的失敗による全体の失敗の回避、対訳編集における文対応の把握の容易性の向上、などがあげられる。実際、文分割の効果を示す実験として、ニュース文の翻訳において、形態素解析の情報をもとに長文を分割することにより、機械の構文解析率が60%から80%へ向上した例が報告されている[4]。このため、様々な翻訳前編集支援システムや推敲支援システムにも指摘項目として長文の指摘機能が搭載されている[5]。

従来の実用システムにおいては、(1)単純に文中の文字数のみを長文の条件とし長文を検出・指摘する、(2)指摘された長文の書き換え方法を具体的に指示できない(=書き換え候補を提示できない)など、指摘の適切性、書き換え支援能力において問題があり、十分な支援が行われているとはいえず難しかった。この原因の1つとして、従来システムは、入力文に対して文字面処理、あるいは辞書引き・形態素処理といった比較的表層に近い処理しか行わず、このため十分な支援知識を記述するための知識基盤がなかったという点があげられる。一方、研究レベルの実験システムにおいては、例えば、構文解析や意味解析レベルの言語情報を利用して文の書き換えを支援する研究[6],[7]が行われており、複文の分割位置の同定に関する言語学的解析・実験がなされている。しかしながら、「何をもって”長文”とするのか」、「書き換え候補をいかにして生成するのか」といった、実際に前編集支援システムを構築するために必要な知識に関しては、あまり議論や実験がなされていない。

本論文では、実際のマニュアル文書を対象として得られた、長文の指摘のための知識および書き換え候補生成知識について述べ、その評価について報告する。2章では、長文処理に入る前に、前編集支援システムの概要、支援知識記述の枠組などについて簡単に触れ、3章以降で、長文処理の方式とその実験結果、考察について述べる。

2. 前編集支援システム

2.1 前編集支援へのアプローチ

前編集に作業における一般的問題として以下の2つがあげられる[1]。

- (1)前編集知識習得のバリア
- (2)書き換え操作による効果の不可知性

最初の問題は、前編集作業を行う際にどの様な基準で文をリライトするのかという知識の習得が困難である点である。自然言語の文のリライト基準を全て外延的に記載することは事実上不可能であり、どうしても「口語表現を使用しない」など、内包的記述とならざるを得ない。また、内包的記述を行ったとしても、一般にかなりの数のリライト基準が存在し、この習得はかなりの労力を要する。また、内包的基準では、実際の文に直面した場合、具体的書き換えまで持つて行くのはユーザにとって困難である場合があり、書き換え操作の効果(翻訳の変化)を事前に知るの是一般に困難である。さらに、ある問題を解消しようとして文のリライトを行った場合、そのリライト操作は、別の問題の解消や別の問題の生起を伴うことがある。すなわち、前編集作業は動的なプロセスであると言える。このような問題を解消するため、以下のような指針の基に前編集支援システムの開発を行っている。

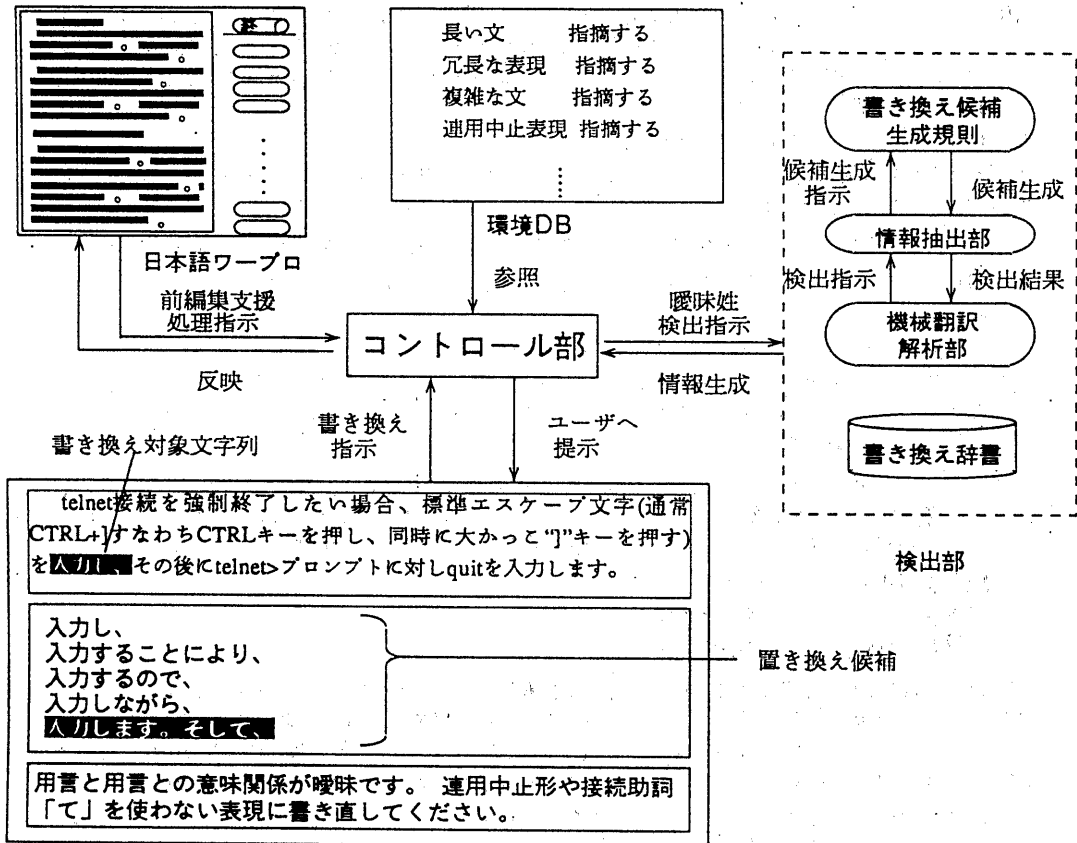
- (1)対話的処理をベースとする
- (2)前編集支援知識は、引き続き機械処理(翻訳処理)の情報に基づき、できる限り具体的書き換え候補を生成する

(1)は、動的プロセス性への対処である。(2)は、翻訳処理の内部情報に基づいた効果的書き換えを実現し、具体的かつ外延的な支援の実現への対処である。

2.2 前編集支援システムの概要

前編集支援システムの構成を図1に示す。前編集支援システムは、前編集用の画面(原文、書き換え候補、ガイダンスなどを表示)を持ち、日本語ワープロから文を受取り、編集結果を返すというサイクルを繰り返して

動作する。システムコンポーネントとしては、文を解析して書き換え候補やガイダンスなど(書き換え情報)を作成する検出部、ユーザーが実際に文の修正を行うインタフェース部、さらに各部を総合的に制御するコントロール部がある。検出部は機械翻訳の解析部の情報を用いる情報抽出部と、ユーザー定義が可能な置換辞書を用いる辞書制御部からなる。情報抽出部は機械翻訳システムの解析処理結果から翻訳前編集支援に必要な形態情報や構文構造を抽出し、これらの情報を基にして書き換え対象の同定や書き換え候補の生成を行うとともに、文の書き換えのために有効なガイダンス情報を作成する。インタフェース部は、解析部より得られた書き換え情報を用いて、図中に示されるような画面構成でガイダンスを提示する。複数の指摘項目が検出された場合には、あらかじめ与えられた優先度に応じて順にユーザーに提示する。コントロール部は、曖昧性検出部とインタフェース部との間のデータ変換などを行う。環境DBには、前編集支援項目の内、どの項目に関して曖昧性を検出するかに関する情報が記憶されており、これにより前編集支援項目のカスタマイズが可能である。



編集画面

図1 前編集支援機能構成図

2.3 前編集支援知識の記述の枠組み

前編集支援知識は、図1に示した情報抽出ツールの規則として記述する。情報抽出ツールは、翻訳処理の各サブプロセスの終了時点において起動される各種言語規則処理部分と、その言語規則から駆動されるCライクなプログラムの処理部分とから構成されている。ここで言う言語規則とは、翻訳の言語解析処理で使用される種々のマッチング規則(単語系列パターンを認識する規則、文の解析木構造を認識する規則など)であり、翻訳処理の中間構造(例えば、辞書引き・形態素処理の後では単語の系列、構文解析処理の後では構文解析木)に関して適用される。これにより、特定の言語構造などの認識が可能となる。プログラムの部分は、認識された言語構造から必要な情報を出力したりするために使用される。このプログラムには、基本データタイプとして翻訳処理途中の各種構造(単語リスト、構文解析木、意味解析木など)が取り入れられており、各種解析情報への容易なアクセスが可能となっている。このため、前編集支援知識の記述の枠組みとしては、表層情報、形態素解析情報、構文解析情報を始めとして、目的言語の構造に関する知識をも利用することが可能である。

ある言語現象に対して長文の書き換え対象同定規則を記述したい場合、それを形態素解析レベルの規則によって行うか、意味解析レベルの規則によって行うかは、通常の翻訳規則と同様、「翻訳処理のどの段階で得られる情報を規則の適用条件として用いるか」によって決定される。例えば、付属語や句読点など構文解析処理以降ではノード属性に還元されてしまう文字列を規則のキーとして用いたい場合は、単語系列レベルの規則で記述する。また、意味解析後に初めて検出される、意味修飾関係の曖昧性の有無を判断する場合は、意味解析木が作成された後に意味解析木に適用される規則で対処する。

3. 「長文」の分割支援

3.1 「長文」の判定

1章でも触れたように、前編集の対象となる「長文」を適切に指摘するためには、文字数といった単純な判定基準では不十分であるということは、経験的には知られている。しかしながら、それではどのような基準によりこれを判定すべきかについては、ほとんど報告がなされていない。このため、我々は、実際の技術文書を対象として、文書の性質を分析するとともに、人手による前編集(長文の処理)を施し、その結果を基に長文判定のための知識を抽出・規則化するというアプローチを取った。

3.1.1 対象データと人手による長文処理

知識抽出の対象とした文書は、実際に市販されている複数の日本語マニュアルから抽出した500文からなる文書であり、以後、「文書1」と呼ぶ。

実験データ「文書1」 総文数: 500文 総単語数: 12634語 平均単語数: $12634/500=25.3$ 語

「文書1」の全ての文に対して、文を分割するか否かという観点からチェックを行い、全ての文を次の3種類に分類した。

- (1) 書き換えが不要だと思われる文(「書き換え不要文」)
- (2) 書き換えが必要であり、さらにその書き換えが可能である文(「書き換え可能文」)
- (3) 書き換えは必要と判断するが、その方法が文分割以外であるか、または、書き換えの方策が立たない文(「書き換え不可能文」)

ここで(1)に該当するのは、翻訳が適切に行われると期待される文である。(2)は、複文などであり、文分割を行った方が翻訳結果の向上が期待できるような文である。(3)は、そのままでは翻訳は成功しそうな復雑な文であり、分割点が存在しない文である。この作業の結果、それぞれの出現頻度は、以下のようになった。

「書き換え不要文」: $373/500=75\%$

「書き換え可能文」: $95/500=19\%$

「書き換え不可能文」: $32/500=6\%$

我々は、この人手による前編集作業によって「書き換え可能文」と分類された文を長文指摘のターゲットとすることとした(以後このターゲットを目標文と呼ぶ)。この目標設定以外に、翻訳結果が向上する文をターゲットとするという方針も考えられる。これは、機械翻訳前編集として理想的な書き換え対象文は、翻訳結果の向上が望める文であるという考え方に基づく。我々がこの方針を取らなかった理由は、以下の通りである。

翻訳結果の向上は翻訳システムの能力に依存しており、これを基準にして規則の構築を行った場合、翻訳システムのある時点での能力にチューンした支援規則となってしまう可能性がある。これを避けるためには、固定的な支援知識ではなく、システムの成長に伴って変化する支援知識とする必要がある。すなわち、翻訳結果の向上をベースとした評価体系は、動的に翻訳結果を判定する自己診断能力に基礎を置くべきものである。

ただし、機械翻訳の品質の向上は、本テーマの最大の課題の1つであり、4章では、翻訳品質の向上をターゲットとした場合の支援知識の評価も並行して行う。

3.1.2 単語数に基づく長文判定

前編集すべき「長文」を検出するには、その特徴を明確化しなければならない。文の複雑さは、様々な要因によって決定されるが、一番単純な判定基準は「文字数」である。文字数は、判定基準として計算量が非常に少ない(すなわち、辞書引きや形態素処理を必要としない)という望ましい性質をもっている。しかし、機械翻訳システムなどでは、いくら長い語句でも辞書に登録されていたら1語として扱われる為、構文解析・意味解析における文の長さの基準は、文字数ではなく単語数であると考えるのが自然である。このため、単語数を基本的なパラメタと考え、単語数が長文の判定基準としてどの程度有用であるか調査するため、「文書1」に対する分析を行った。

まず、「文書1」の各文の構成単語数を情報抽出ツールを用いてカウントし、単語数をパラメタとした文の

頻度分布を求めた。さらに、この分布に目標文であるかないかの情報を重ね合わせ、単語数をパラメタとした評価を行った。

ある基準で長文と判定された文を、その基準における指摘文と呼び、評価項目である適合率と再現率は以下のように定義される。

適合率(%) = 全指摘文中の目標文の数 / 指摘文の総数 x 100

再現率(%) = 全指摘文中の目標文の数 / 目標文の総数 x 100

また、これとともに、長文候補として1回指摘されるのに要する平均期待文数を指摘頻度として計算した。この値は、前編集システムを使用した場合のインタラクションの頻度の目安となる。単語数に着目した場合の適合率、再現率、指摘文総数、指摘頻度は以下ようになった。

単語数	適合率(%)	再現率(%)	指摘文総数	指摘頻度
15以上	24	100	396	1.26
20以上	33	99	285	1.75
25以上	43	90	198	2.52
30以上	58	79	129	3.87
35以上	71	61	81	6.17
40以上	81	48	56	8.92
45以上	83	30	34	14.7
50以上	79	20	24	20.83
55以上	100	13	12	41.66

適合率は、「単語数15以上」位から向上し、「単語数55以上」では100%になる。再現率は、「単語数20以上」付近から減少しはじめ、「単語数55以上」付近で0になる。

3.1.3 言語表現に基づく長文判定

文の複雑さを計るもう一つの手段として、文の辞書・形態情報・構文情報を用いることが考えられる。一般的に、重文や複文は、短文に比較して多くの意味内容を含んでいることが多く、表層的には長い文となりやすい。複文といった、文の構文的特徴は、その文を解析することにより得ることが可能である。武石他は、複文を分割する分割点を認識するために、接続表現(「ので」「が」「ながら」など)の分類を行い、それに基づく優先度情報と各文要素のスコープ解析を利用している[6]。このように、言語表現は、長文判定のための判定基準としての利用可能であるとともに、分割点や書き換え候補生成のキーとしての働きを持つ。

我々は、「文書1」から得られた目標文および目標文に対して適用された人手による書き換え結果を解析し、目標文に含まれる、対象となった言語表現とその表現に適用された書き換え操作を抽出した。目標文の認識、すなわち書き換え対象同定の観点からすれば、ここで抽出された表現は、表層的であり、長文の認識規則というには単純となったが、これらの表現に対しては書き換え候補生成が可能であり、前編集支援としては望ましい性質を持っている。ここで抽出された書き換え対象同定表現の例を図2にあげる。現在、約100表現がインプリメントされている。

書き換え対象同定規則に着目した場合の適合率、再現率、指摘頻度は以下の通りである。

適合率=87/244=36% 再現率=87/95=92% 指摘総文数: 242 指摘頻度: 2.06

この結果から言語表現に基づいた書き換え対象同定規則のみによる対象同定は、再現率は92%と高いが、適合率は36%にとどまっている。3.1.2で示したように、単語数25以上を判定基準とした場合に、適合率43%、再現率90%、指摘頻度2.52が実現できるので、単純に比較すると単語のみを判断基準とした方が効率的である。

連用中止表現(動詞、助動詞)	「～すると」などの表現
接続助詞「ので」を含む表現	「～なければならず」などの表現
接続助詞「が」を含む表現	「～した後」などの表現
接続助詞「とともに」を含む表現	「～した為」などの表現
「であれば」などの表現	「～とした場合」などの表現

図2 「書き換え対象同定表現」の例

3.1.4 単語数と言語表現を考慮した長文判定

本節では、単語数と言語表現を組み合わせて、目標文の同定を行う方法について考察・実験評価を行う。前編集支援においては、できるだけ漏れなく適切に書き換え対象を指摘したい。このためには次の2つの方策を考えよう。

- (1) 判定基準により排除された目標文を正しく対象であると認識するよう基準を変更する。
- (2) 判定基準に合致した非目標文を対象であると認識しないよう基準を変更する。

言語表現による判定を考えると、上記(1)は、「書き換え可能な表現が文中に存在するが文全体としては書き換えの必要が無い文」に対応し、(2)は、「書き換えの必要があったが、文中に書き換え可能な表現が存在しなかった文」に対応する。この「書き換えの必要性」の部分に単語数による判定条件を加味する方向でデータ上の検討を行った。この結果、次の2つのヒューリスティクスが得られた。

- ・単語数23以下の文は、「言語表現による判定」がマッチしても「指摘すべき文」である可能性が低い
- ・単語数40以上の文は、「言語表現による判定」がマッチしなくても「指摘すべき文」の可能性が高い

上記の結果を踏まえ、文の複雑さの指標として以下の指摘方法を考案した。

- ・単語数0以上24未満の文は、指摘しない
- ・単語数24以上40未満の文は、「言語表現による判定規則」を適用し、マッチしたもののみ指摘する
- ・単語数40以上の文は、全て指摘する

この書き換え対象同定方式を、複合判定方式と呼ぶ。これを用いて適合率と再現率を算出すると、以下のようになる。

適合率=91/174=52% 再現率=91/95=96% 総指摘文数: 175 指摘頻度: 2.85

複合判定方式においては、言語表現単独の場合に比べ、適合率で16%の改善があり、再現率で4%の改善が見られた。また、単語数(25以上)の場合と比較して、適合率で9%の改善があり、再現率で6%の改善が見られた。

3.2 書き換え候補生成

3.1.3「言語表現に基づく長文判定」で述べたように、「文書1」に対して施された人手による前編集結果から書き換え操作を抽出した。この人手による前編集操作を解析した結果、書き換え操作を、大きく4つに分類した。以下に4つのタイプを出現割合(%)と共に示す。

- (1) 単純分割 (60%): 分割点で2文に分割し、1文目の語尾を変化させる。2文目の最初に接続詞を付加する場合も含める。
- (2) 格要素補填 (27%): 分割後、2文目に主語などを補填する
- (3) 用言補填 (7%): 分割後、2文目に用言などを補填する
- (4) その他(格変化を伴う補填など)(6%)

単純分割が全体の60%を占めており、比較的簡単な処理によってかなりの部分の書き換え候補を生成できる見通しが立った。このため、まず、単純分割に対する書き換え候補生成規則を抽出し約100の規則を作成し、インプリメントした。書き換え候補生成の評価については、4章で述べる。

4. 評価実験

3章では、複数のマニュアル文書から取り出した500文からなる文書(文書1)をもとに書き換え同定規則および書き換え候補生成規則を作成し、規則抽出の元となった「文書1」に対する評価を行った。しかしながら、この評価実験では、別の文書に対する規則の有効性については必ずしも明白ではない。このため、新たな文書(「文書2」とする)を用いて規則の評価実験を行った。実験対象である「文書2」は、「文書1」と同様に、実際に商品と共に発売されているマニュアル文である。「文書2」の平均単語数は以下に示すように「文書1」に比べて10%程度多いが、同種の文書である。

総文数: 211文 総単語数: 5992語 平均単語数: 5992/211=28.4語

4.1 書き換え対象同定

書き換え対象同定規則に関しては、3章と同様の評価(すなわち人手による書き換え対象となった文をターゲットとした再現率、適合率)とともに前編集後の文書の翻訳結果に対する評価を行うこととした。後者の評価は、人手による書き換えが行われた文の内で、その書き換えによって翻訳の質が向上した文をターゲットとした場合の対象同定規則の再現率と適合率である。この評価は、実際の前編集作業者のスキルや翻訳システムの能力により、値が異なってくるが、翻訳結果への効果を見る上での一応の目安となると考えられる。

評価実験は、以下の手順で行った。

- (1) 評価用文書(文書2)の各文を手手で「書き換え不要文」、「書き換え可能文」、「書き換え不可能文」の3カテゴリに分類する。

結果: 「書き換え不要文」: 109/211=52%
「書き換え可能文」: 64/211=30%
「書き換え不可能文」: 38/211=18%

(2) (1)で「書き換え可能文」に分類された文に対して、分割箇所を決定し、3.2で述べた4つの書き換え選択肢より書き換え方法を選択し、文を分割する。分割された文の集合を、「文書2」とする。

(3) 「文書2」中の「書き換え可能文」と、「文書2」の文とを、機械翻訳システムを使って翻訳する。対応する文の翻訳結果を比較し、「翻訳向上」、「翻訳無変化」、「翻訳悪化」の3種類に分類する。

結果: 「翻訳向上」: 34/64=53%

「翻訳無変化」: 30/64=47%

「翻訳悪化」: 0/64=0%

(4) 3.1.4で述べた書き換え対象同定規則と3.2で述べた書き換え候補生成規則を前編集支援機能を用いて「文書2」に適用し、前編集支援システムが書き換え対象として指摘する文と書き換え候補表現を求める。

結果: 「書き換え対象」: 93

(1)と(4)に共通に含まれる文は58文であった。この結果から、「文書2」に対する書き換え対象同定規則の評価結果は以下ようになる。

適合率=58/93=62% 再現率=58/64=90%

「文書1」に対する適合率、再現率と比較すると、再現率での低下は見られるものの、適合率は上昇しており、ほぼ同等の性能が得られたと言える。

次に、翻訳結果の向上をターゲットとして仮定した場合について述べる。(3)の結果、「書き換え可能文」中で翻訳結果が向上した文は、34文であり、また、(4)の結果、前編集支援システムが書き換え対象として指摘した文は93文であった。この内、共通に含まれる文は32文であった。この結果、翻訳結果の向上をターゲットとした場合の「文書2」に対する書き換え対象同定規則の評価結果は以下ようになる。

適合率=32/93=34% 再現率=32/34=94%

この結果から、今回開発した書き換え対象同定規則は、翻訳結果の向上が期待される文の大部分(94%)を書き換え対象として指摘しているが、翻訳結果の向上が期待できるのは3文に1文(34%)であるということがおおまかな指針として得られた。ただし、この値はあくまでも指針であり、これに関する検討は、5章で述べる。

4.2 書き換え候補生成

書き換え候補生成に関しては、人間が行った書き換え操作(書き換え後の表現)をターゲットとし、3章で述べた書き換え候補生成規則が生成する書き換え表現候補の再現率と適合率を求めた。実験の対象は、「文書2」の「書き換え可能文」の内、書き換え候補生成をシステムが現在既にサポートしているもの(=4.1(2)で(1)と分類された文)である。これらの文に対して、システムによって指摘された分割点と、それぞれの分割点に対応する書き換え候補の数とを、4.1の(2)で人間が行った分割と比較した。この結果は、以下の通りである。

実験対象文: 42文

システムによって提示された分割点の総数: 70箇所

人間が指定した分割点と一致した分割点の数: 44箇所

「文書2」に対して人手により「書き換え可能文」と判定された文は、64文であったので、文数の割合として、その内の66%(42/64)に対して書き換え候補を生成したことになる。1つの文の中に複数の書き換え候補(分割点)が提示されることがあるため、提示された分割点の総数は、対象文の数には一般に一致しない。今回の実験では、1対象文あたり1.67か所(70/42)の分割点が候補として提示された。今回の実験では、人手による分割点をすべて指摘していたため、分割点の指摘に関する適合率および再現率は以下ようになる。

適合率=44/70=63% 再現率=44/44=100%

人手により指摘された分割点44箇所において、前編集支援システムが生成した表現候補と実際に人間が行った書き換え候補を比較したところ、以下ようになった。

システムにより提示された書き換え候補の総数: 389

人間が行った分割に使われた表現を書き換え候補として含んだ分割点の数: 41箇所

1分割点あたりの平均書き換え候補数: 389/70=5.6

このため、書き換え候補生成規則の適合率および再現率は、以下ようになる。

適合率=41/389=10.5% 再現率=41/44=93%

以上から、新規文書に対する書き換え候補生成規則の分割点候補生成および書き換え表現候補生成ともに90%以上の再現率となり、高い確率で適切な候補を含んでいるといえる。書き換え候補生成規則の適合率については、かなり低い値となっている。しかし、1分割点あたりの書き換え候補数が5.6であり、人間とのインタラク

ションを考えた場合、それほど問題にはならない数ではないかと考えられる。これについては、インタフェース操作を含めた今後のさらなる評価が必要である。

5. 考察

4章では、前編集支援システムのターゲットを人手による前編集結果とした場合と機械翻訳結果が向上した文とした場合との2種の評価結果をあげた。実際の翻訳業務の流れの中で前編集支援システムの性能を評価しようとする場合、様々な要因を考慮する必要がある。例えば、機械翻訳システムの出力結果を基にした評価の場合、前編集により翻訳結果が向上した文をターゲットとして評価を行ったが、文の分割操作の場合、翻訳結果が向上しなくても全体としての作業効率の向上が期待できる。この理由は、機械翻訳の後編集作業において、原文と訳文の対応関係を見ながら訳文の修正を行う際に、前編集により文分割がなされていれば、原文・訳文の1文あたりの長さが短くなり、その対応関係を把握するのが容易になり、後編集作業の効率向上が期待できるからである。また、前編集作業者と翻訳作業者との組み合わせ、すなわち運用体制によっても全体の効率を左右するであろうと予想される。例えば、一般に翻訳者の時間コストはかなり高いが、前編集作業を時間コストの比較的低い非翻訳者や原文作成者が行うといった工夫も考えられる。このように実運用においては、前編集作業を全体の流れのなかで評価する必要があり、今後の課題である。

本論文では、文の複雑さを決定する要因として単語数と言語表現(主として文の切れ目となるもの)をあげ、それらの組み合わせによる指摘方法を提案した。しかし、これら以外にも、「用言の数」、「有効語数(=数字、付属語などを除き、合成語を1語とした語数)」、「句点の数」、「接続詞の数」などが文の複雑さに関与していると考えられる。今後、さらなる性能改善が必要となった場合には、今回用いた2要因にその他の要因も加え、総合的判断による改良が1つの方向として考えられる。また、今回は文の単純分割にのみ対応して書き換え候補生成規則を作成したが、格要素補填などの分割方法にも対応することにより、支援効率を向上することが可能である。格要素補填を行うためには、文の構文・意味解釈から格要素の補填候補を取り出す必要があるが、本システムでは、2章で示したように、前編集支援のための規則から翻訳過程の情報を簡単に取り出すことができる。この機能を利用して、書き換え候補生成の性能向上を図っていく予定である。

6. 結論

本報告では、日英機械翻訳の前編集作業(日本語文リライト)において効果の高い長文分割処理を支援するための知識の抽出、前編集支援システムへのインプリメント、評価について述べた。長文分割支援においては、分割すべき文の適切な指摘、分割方法に関する適切な候補生成が重要な課題であり、実際のマニュアル文書(500文)からこれら課題に対する知識を抽出し、既開発である前編集支援システムに組み込んだ。実マニュアル(211文)に対して評価実験を行った結果、分割すべき文の指摘に関しては、適合率62%、再現率90%の性能が得られた。また、分割表現の候補生成に関しては、66%の文に対して書き換え候補の生成が可能であり、生成された書き換え候補に関しては適合率63%、再現率90-100%の性能が得られた。今後は、マニュアル以外のタイプの文書に対する実験・評価・改良を行うとともに、長文分割以外の前編集支援への知識拡張を行っていく予定である。

参考文献

- [1] 野村他:機械翻訳前編集支援ツールの開発,第45回情報処理全国大会(1991)
- [2] 吉田:機械翻訳と文法記述-日本語の規格化と制限日本語の設計,「日本語の特性と機械翻訳」,「大学と科学」公開シンポジウム資料(1987)
- [3] 内田:機械翻訳の効率化をめざす制限言語の仕組みと特徴,翻訳の世界,3月号(1992)
- [4] 金他:形態素解析情報に基づく長い日本語ニュース文の分割,第44回情報処理学会全国大会(1992)
- [5] 菅沼他:日本語文章推敲支援ツール「推敲」における字面解析手法とその評価,情報処理学会自然言語処理研究会研究報告,68-8,(1988)
- [6] 武石他:複文における分割点の特定について,第4回人工知能学会全国大会(1990)
- [7] 林他:日本文推敲支援システムにおける書き換え支援機能,第4回人工知能学会全国大会(1990)