新聞記事日本文における書き替え対象表現の分布

白井 諭† 阿部 さつき‡ 矢部 孝幸‡ 久保 京子‡ 池原 悟† 横尾 昭男†

†NTTコミュニケーション科学研究所 ‡NTTアドバンステクノロジ(株)

1 はじめに

機械翻訳を翻訳業務に適用するには現状では訳文品質の低さが問題となる。それを克服する方策として，第1に対象文書に合う翻訳システムを開発する，第2に既存の翻訳システムに対象文書を合わせる，という2つの考え方がある。

第1の方策としては，解析ベースや用例ベースに分類される翻訳方式が多数提案されてきた。これらは，従来のような要素合成の考え方に基づくのではなく，原言語と目的言語とでなるべく大きな表現や構造のまとまりから順に対応付けていくことを目指している。しかし，翻訳品質を向上させるには，解析ベースの方法では辞書やルールの数と記述精度を，用例ベースの方法では用例の数と網羅性をそれぞれ向上させる必要があり，それらは現状の規模から考えても容易ではない。

第2の方策としては，制限言語の考え方や人手による前編集を挙げることができる。制限言語は執筆者の自由な発想を妨げかねないほか，機械翻訳の結果までは保証されないという問題がある。人手による前編集では，翻訳結果を見ながら前編集を再試行することが可能な反面，同じ表現が何度も出現すればその都度同じ作業を繰り返す必要がある。字面上は同じ表現であっても，書き換えてもよい場合と書き換えてはいけない場合があるため，前編集の自動化は困難であった。

これに対し，筆者らは，前編集の対象になる表現の分析に基づき，①単語の詳細な文法的，意味的属性を用いてルールを記述する，②原文の解析が進行し，ルールの適用条件の判定に必要な情報が揃った時点でルールを適用する，ことにより副作用の心配のない自動書き替えが実現できることを示し，日英翻訳システムALT-J/E[池原87]に適用して，訳文合格率（訳文だけで意味のわかる文の割合）が2割程度向上することを実験的に確認した[白井90,93,95]。しかし，どの程度のルールが適用可能かが不明であるため，チューンの工程が組みにくい。

そこで，本稿では，このような自動書き替えの対象となる表現が実際の文章にどれくらい含まれるかを明らかにする。具体的には，新聞記事および市況速報という2種類の文章を対象に作成した書き替えルールの使用情況を分析し，書き替えルールの必要量や対象とした文章の特徴を分析する。

2 書き替え処理の概要

書き替え処理の対象となる表現は次の条件を満たすと考えられる[白井93,95]。

①そのままでは適切な翻訳結果が得られない
②意味を変えない別の表現に書き替えられる
③その書き替えを行なえば翻訳可能となる
④既存の翻訳機能に対し悪い副作用を生じない

これらのうち①～③は人手による前編集の場合と同じであるが，④は異なる。すなわち，人手による前編集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，書き替え処理の場合，登録した書き替えルールは該当する表現すべてに適用されるため，書き替えてはならないものを書き替えてしまう恐れがある。このため，書き替えルールは，Ⓐ適用条件を精密に記述すること，Ⓑ適用条件が判定できる情報が得られた段階で適用すること，が必要である。また，②に関して，人手による前編集では原言語内に意味を変えない別表現が存在しなければ書き直すことができない。これに対して，書き替え処理の場合，翻訳システムの内部で書き替えを行なうことから，原言語内に別の表現がなくても，目的言語に適切な表現があれば，それを直接指示することで救済できる。

以上から，書き替え処理は，翻訳システムの中では，形態素解析と構文解析を終えた後，意味解析の前に置くことにした。また，初期は前編集相当の書き替え（日本語内書き替え）と目的言語を

指示する書き替え（疑似日本語書き替え）の2種類を設定したが，その後，後処理として形態素解析や係り受け解析の誤り補正に利用したり，主体的表現を捉えるのに応用したりできることがわかったので，現状は図1の構成となっている。また，書き替えルールは表1に示すように4種類（14細分類）に体系化されている[白井94a, b]。

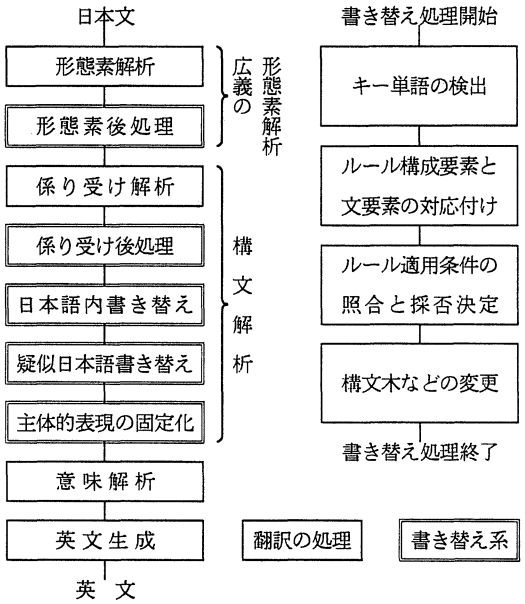

図1 書き替え処理の位置づけと構成

表1 書き替えルールの体系化

分類	細分類	書き替えルールの適用対象
解析後処理	形態素補正	単語の分割誤り等を狙い撃ち救済
	形態素多義絞り込み	単語並びにより不要な解釈を除去
	係り受け補正	係り受けの誤解釈を狙い撃ち救済
	係り受け多義絞り込み	多項関係により不要な解釈を除去
日本語内書き替え	縮約展開	語尾の省略や短絡的な用法を回復
	冗長圧縮	訳出不要な表現を除去して簡潔化
	構文組み換え	情況的な表現を断定文などに変換
	敬語の標準化	訳出不要な表現を除去して簡潔化
疑似書き替え日本語	助詞相当語	英語の前置詞句に直接対応付ける
	副詞相当語	英語の副詞表現に直接対応付ける
	連体詞相当語	英語の形容詞句に直接対応付ける
	フレーズ	英語の定型表現へ直接対応付ける
主体表現	接続様相時制表現	文の接続の構造を固定的に捉える
	様相時制表現	文末の表現を英語対応に固定する

3 書き替え対象表現の分布

書き替え対象表現の特徴を明らかにするため，本稿では2つの分野を選定して，それぞれの分野に適用された書き替えルールの比較を行なうことにした。選定した分野は，1つは代表的な書き言葉である新聞記事文，もう1つは専門的な表現を多く含む市況速報文である。ともに日本経済新聞社のデータベースであるテレコンBIZに収録されており，ダウンロードして使用することが許容されている。

これらの文に対して走行実験と書き替えルールの作成を数回繰り返した後，書き替えルールの適用情況を集計した。なお，書き替えルールによる訳文品質への影響に関する詳細は割愛するが，書き替えを行なわない場合に比べ，新聞記事文・市況速報文とも訳文合格率が20%程度向上した。

3.1 新聞記事の場合

本稿で対象とした新聞記事文は1994年8～9月の日本経済新聞から抽出した90記事874文である。記事の例を図2示す。

この文に適用された書き替えルールは異なり162ルールで，延べ463回適用された。内訳は，解析後処理42ルール（延べ224回），日本語内書き替え19ルール（34回），疑似日本語書き替

45. 都市ガス3社検討、ガス導管材料、海外調達――「国産限定」を転換。
94/ 9/ 4 日本経済新聞 朝刊 P 7 768字 FAX可

東京ガス、大阪ガス、東邦ガスの都市ガス三社はガス導管の材料を海外から調達する方向で検討を始めた。スチール管は韓国製鋼材を、ポリエチレン管の素材は欧州化学メーカー製を採用したい考え。円高差益を利用してガス導管への設備投資を軽減するのが狙い。ガス導管は品質について高い信頼性が要求されるためこれまで国産に限っており、海外からの本格的な調達はこれが初めて。

ポリエチレン管の素材を欧州の化学会社から輸入する方向で検討に入っており、スチール管に先だって導入が始まりそうだ。素材自体の価格は日本製に比べて二―三割は安いという。

三菱樹脂や積水化学工業などポリエチレンパイプを生産しているメーカーに素材の輸入を要請し、加工してもらう形をとる。年度内にも三社が共同で欧州メーカーに調査団を派遣し、購入先や購入量など詳細をつめる。

一方、スチール管についても安価な韓国製鋼材を試験的に使用する方針。これまで取引していた高炉メーカー以外のパイプメーカーを通じて韓国製鋼材を輸入、加工してもらう方法も浮上している。設備投資の根幹であるガス導管はこれまで新日本製鉄、住友金属工業、ＮＫＫなど高炉メーカーから調達している。導管への投資額（工事費除く）は九三年度で三社あわせて約三百億円。

（以下2段落略）

図2 新聞記事の例

え 90 ルール（166 回），主体的表現 11 ルール（39 回）である。最も多く適用されたルールは「～して」型の連用中止の接続属性の多義を絞り込むための形態素多義絞り込みルール（84 回）で，このほか「～として」を助詞相当語として固定するためのルール（19 回），「～する見通しだ」を様相表現として固定するルール（9 回）などの適用回数が多い。

表 2 新聞記事文に適用されたルール

解析後処理	異なり	累積
形態素補正	32	66
形態素多義絞り込み	6	129
係り受け補正	2	23
係り受け多義絞り込み	2	6
小計	42	224

疑似日本語書き替え	異なり	累積
助詞相当語	58	125
副詞相当語	25	32
連体詞相当語	7	9
フレーズ	0	0
小計	90	166

日本語内書き替え	異なり	累積
縮約展開	1	1
冗長圧縮	5	5
構文組み換え	13	28
敬語の標準化	0	0
小計	19	34

主体的表現	異なり	累積
接続様相時制表現	1	1
様相時制表現	10	38
小計	11	39

合計	162	463

3.2 市況速報文の場合

本稿で対象とした市況速報文は 1995 年 7 月の 84 記事 546 文である。記事の例を図 3 に示す。

この文に適用された書き替えルールは異なり 43 ルールで，延べ 337 回適用された。内訳は，解析後処理 13 ルール（延べ 147 回），日本語内書き替え 6 ルール（50 回），疑似日本語書き替え 24 ルール（140 回）で，主体的表現は適用されなかった。最も多く適用されたルールは「半面、～」型の形態素補正ルール（82 回）で，このほか，「～を中心に～」を助詞相当語として固定するためのルール（24 回），「～日振りに」を "for the first time in ～ days" に訳出するためのルール（16 回）などの適用回数が多い。

95/07/04/11:21
大証前引け・反発も商い低調

大証修正は反発。小型の仕手系材料株などが個別に物色されたものの、景気の悪化懸念や参院選挙を控えていることから主力株中心に見送り気分が強く、商いは低調。森精機、栗本鉄、富士通ゼが買われ、ヤマトインタ、森田ポ、御幸毛、大真空も物色された。朝方、安値を更新したトーア紡も持ち直した。オートバクス、ワキタ、日精化、青山商も高い。半面、村田製、オムロン、任天堂、小野薬がさえず、参天薬、シマノも安い。

図 3 市況速報記事の例

表 3 市況速報文に適用されたルール

解析後処理	異なり	累積
形態素補正	9	107
形態素多義絞り込み	3	39
係り受け補正	1	1
係り受け多義絞り込み	0	0
小計	13	147

疑似日本語書き替え	異なり	累積
助詞相当語	16	120
副詞相当語	8	20
連体詞相当語	0	0
フレーズ	0	0
小計	24	140

日本語内書き替え	異なり	累積
縮約展開	1	14
冗長圧縮	1	14
構文組み換え	4	22
敬語の標準化	0	0
小計	6	50

主体的表現	異なり	累積
接続様相時制表現	0	0
様相時制表現	0	0
小計	0	0

合計	43	337

4 考察

まず始めに，書き替えルールの収束情況を調べるため，新聞記事文と市況速報文の双方について，横軸に文数，縦軸に使用累積ルール数と異なりルール数をプロットしてみた。結果を図 4 に示す。図 4 から，文数の増加に比例して使用されたルール数が増加するが，異なりルール数は飽和傾向にあることがわかる。新聞記事文で本稿で対象とした 874 文の範囲ではまだ飽和していないが，市況速報文では 400 文程度でほぼ飽和している。また，市況速報文は新聞記事文に比べ 1 文当たりのルール適用数がやや多いが異なりルール数では 4 割程度にとどまり，特定の表現が多用されていること

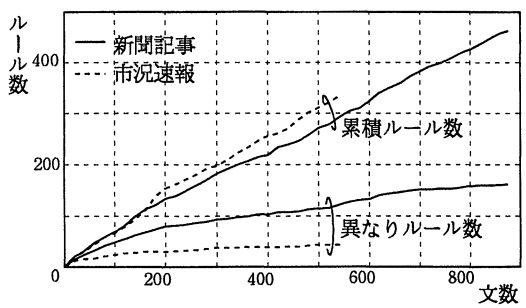

図 4 書き替え対象表現の分布とルール数

表4 共通ルールと使用回数

番号	書き替えルールの概要	使用度数		
		新聞	市況	合計
1	「～して」の接続属性の多義を絞る（形態素多義絞り込み）	84	15	99
2	文頭にある「半面、～」を副詞に解釈する（形態素補正）	1	82	83
3	副助詞「は」の係りを接続語に応じて調整（係り受け補正）	21	1	22
4	「～振りに」をfor the first time inに訳出（助詞相当語）	2	16	18
5	「～円台」「～%台」のような「台」を除去（冗長圧縮）	1	14	15
6	「～にかけて」を固定化してtowardに訳出（助詞相当語）	1	12	13
7	「～に対する」を助詞相当語として固定化（助詞相当語）	2	7	9
8	「やや～」を前後に対応してslightlyに訳出（副詞相当語）	1	6	7
9	「～に加え」を固定化してbesideに訳出（助詞相当語）	1	6	7
10	「～で、」の格助詞解釈を排除する（形態素多義絞り込み）	6	1	7
11	「一段と～」が副詞の場合を検出し固定化（副詞相当語）	4	1	5

がこの図からも読みとれる。

次に，新聞記事文と市況速報文とで共通に使用されたルールを抽出した。共通ルールを表4に示すように11ルールにとどまった。表4を見る限り，共通ルールとはいうものの使用度数の点では適用対象には偏りがあることがわかる。

なお，新聞記事文874文に適用された書き替えルール162件のうち1回しか使用されなかったルールは105ルールで65%を占めるのに対して，市況速報文546文では書き替えルール43件のうち14件で33%に過ぎない。このように，新聞記事文は表現に様々なバリエーションを含んでいることを考えれば，新聞記事文を対象にして作成されたルールはある意味で共通ルールとしての性質を備えているといえるかも知れない。

5 おわりに

本稿では，日英機械翻訳の訳文品質を向上させる上で効果のある書き替え処理はどれくらいの書き替えルールを具備すべきかを明らかにするため，代表的な書き言葉である新聞記事文と，専門的な表現が多用される市況速報文を対象にして，書き替えルールを作成し，書き替えルールの適用情況を考察した。その結果，新聞記事文では200～300ルールが必要で，そのルールを作成するには2,000文程度の走行試験を要すること，また，市況速報文では50ルール程度で必要ルール数は飽和し，その作成は400文程度でよいこと，この2種類の文で共通に使用されたルールは11ルールにとどまったこと，などがわかった。また，新聞記事で使用されたルールうちが1回しか使用されないものが2/3を占めるのに対し，市況速報文では1/3にとどまり，市況速報文は表現が限られていることも数字の上で裏付けられた。新聞記事文は様々な表現を含んでいるため，共通ルールとしての性質をある程度備えていると予想される。

今後は，市況速報文以外の専門分野に対して同様の考察を行ない，専門分野の特徴に関する検討を行なう予定である。

謝辞

書き替え処理の実現にご協力くださった松尾三津恵氏，中村三紀氏を始めとするＮＴＴアドバンステクノロジの各位に感謝する。

参考文献

[池原87] 池原，宮崎，白井，林：言語における話者の認識と多段翻訳方式，情報処理学会論文誌 Vol.28 No.12

[池原93] 池原，宮崎，横尾：日英機械翻訳のための意味解析用の知識とその分解能，情報処理学会論文誌 Vol.34 No.8

[白井90] 白井：日本文自動書き替えによる構文多義の解消，情報処理学会第41回全国大会 4S-6

[白井93] S. Shirai, S. Ikehara, and T. Kawaoka : Effects of automatic rewriting of source language within a Japanese to English MT system, TMI '93 Proceedings of the Conference, Kyoto

[白井94a] 白井，池原，阿部，松尾：日本文書き替え処理における制御ルールの類型情報の抽出，情報処理学会第49回全国大会 4K-10

[白井94b] 白井，池原，松尾，兵藤：日本文書き替え処理における制御機能の構成について，情報処理学会第49回全国大会 4K-11

[白井95] 白井，池原，河岡，中村：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，情報処理学会論文誌 Vol.36 No.1