日本文書き替え処理における制御ルールの機能別構成

6P－4

白井　諭*　池原　悟*　河岡　司**　上田洋美***
(shirai@nttkb.ntt.jp)
*ＮＴＴ情報通信網研究所　**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

1　はじめに

日英翻訳システムの研究開発が進み，翻訳業務にも適用されるようになってきたが，さらなる訳文品質の向上のための新しい理論や方式の提案が期待されている[1]。筆者らは，先に原文自動書き替え型翻訳方式を提案し，日英機械翻訳システムにおける走行実験に基づいて，訳文品質の向上効果等について報告した[2,3]。この報告では，書き替えのタイプを日本語内書き替えと疑似的日本語への書き替えの２つに分類したが，両タイプの書き替えは１つの処理の中で実施した。

上記の両タイプに該当する表現が重なり合うと，複数のルールを合成したルールを作らざるを得ない場合がある。本稿では，制御ルールを機能別に分類し，これらを別ステップとして段階的に適用することにより，書き替えルールの汎用性を高める方法を提案する。また，日本文書き替え処理の持つ表現検定と表現生成の機能を応用して，解析誤りを補正したり複合的表現を抽出したりする方法を併せて提案する。

2　日本文書き替え処理の応用

先の報告[2,3]では，日本語内書き替え（縮約展開，冗長圧縮，構文組み換え）と疑似的日本語への書き替え（独立句，様相時制表現，接続表現）の２つのタイプを定義し，形態素解析および係り受け解析によって得られた解析情報に対し，単語の表記，品詞，意味属性ならびに文節間の係り受け属性により記述された書き替えルールを用いて処理を実施した。この際，書き替え後の表現を再度書き替え処理の適用対象とすると，予想外の書き替えに伴う副作用の恐れがあるため，１箇所への書き替え処理の適用は１回以内に制限していた。

ところで，次の文では２つの書き替えが必要である。

「本処理の結果を用いて変換および生成する。」

1. 「を用いて」は助詞相当語として固定的に捉え直すことにより英語の前置詞相当語句 "based on" に訳出する（疑似的日本語への書き替えの独立句）
2. 「変換および生成する」は名詞「変換」と動詞「生成する」の並列であるので「変換しそして生成する」に置換する（日本語内書き替えの縮約展開）

ここで，1. の書き替え前の「用いて→生成する」は述語相互の係り受けであったが，書き替え後の「～を用いて→生成する」は格関係に変化する。従って，1. のルールには「生成する」対応の文節を記述する必要が生じ，2. のルールの記述範囲と重なっている。

先の報告の段階では，このような場合は１つの専用ルールを作成した。しかし，性質の異なる書き替えは，２回に分けても副作用の恐れはほとんどなく，分けることによって各ルールの汎用性を高めることができる。そこで，本稿では書き替えルールを機能別に分類し，段階的に適用することを考える。

また，形態素解析や係り受け解析は，基本的には解析ルールにより処理が制御されるが，特定の単語の特殊な振舞いはルール化しづらく，解析失敗の原因となる。誤りの回復は，本来は解析処理に帰すべきであるが，本稿では，解析誤りをパターン化して捉え，正しいパターンに置換することによる解析誤りの回復を併せて考える。

3　日本文書き替えルールの機能分類

本来の書き替え処理は，形態素解析と係り受け解析で得られた解析結果に適用される。このため，誤りの回復は本来の書き替え処理に先立って行なうことにする。

日本語内書き替えは意味解析を可能とするための処理であるのに対して，疑似的日本語への書き替えは訳出のための制御である。従って，この両者には相互干渉の心配はないので，２つに分離する。

また，疑似的日本語への書き替えのうち，独立句は表現を固定的に捉え訳語も想定するのに対して，様相時制表現や接続表現は要素合成が不可能な表現を抽出するのが目的である。そこで，後者を分離して位置づける。

3．1　解析の後処理

本来の書き替え処理に先立って行なわれる形態素解析と係り受け解析の解析結果を対象として実施する。

①形態素補正

ＡＬＴ－Ｊ／Ｅの形態素解析は単語あたり99.8%の精度があるが[4]，主に次のような場合に失敗する。

a. 長いひらがな列：ひらがな書きの自立語に幻惑されて単語認定を誤ったと思われる。
　［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)

b. 接辞を伴う複合語：「畜産/物/価格/安定/法」のような場合の接辞認定を強化した副作用と思われる。
　［誤］現(接頭語)/代用(名詞)/語(接尾語)
　［正］現代(名詞)/用語(名詞)

これらの回復には，誤りパターンを条件部，正解パターンを生成部に持つルールを用いるのが直截的であるが，誤りの主因である語（a. の「いも」，b. の「代用」）を無効扱いにして形態素解析を部分的にやり直す方法も考えられる。いずれによるかはケースバイケースで選択する。

A Functional Restructuring of Rules for Automatic Rewriting of Japanese Sentences
Satoshi SHIRAI*, Satoru IKEHARA*, Tsukasa KAWAOKA**, and Hiromi UEDA***
*NTT Network Information Systems Laboratories (1-2356 Take, Yokosuka, 238-03), **NTT Communication Science Laboratories, and ***NTT Advanced Technology Corporation

②形態素多義絞り込み

ＡＬＴ－Ｊ／Ｅでは“内部表記”という概念を用い，例えば「なる」の内部情報を「成る｜鳴る｜生る」の３つに展開し，意味的に整合するものが最終的に選択される。しかし，意味解析の負荷につながるため，直前・直後の文節（「なる」では直前の「～に」を伴う名詞）を手がかりにして不要なものをあらかじめ絞り込む。

具体的には，条件を満たす文節列に対する優先／非優先をルールに記述し，総合的に優先される解釈を残す。新聞記事1000文を用いた予備調査では，１文平均2.15の形態素多義を1.15まで絞り込めることがわかった。

③係り受け補正

現在，ルールベースによる係り受け解析処理を新規に作成中であるが[6]，特定の単語の特殊な振舞いがルールに書き切れない場合がありうる。誤りのパターン化が可能なら，正しい係り受け関係に置換するルールによる係り受け解析誤りの回復も考慮する。

３．２　日本語内の書き替え

先に提案した日本語内書き替えとして④⑤⑥の３つがある。本稿では，４つめとして，パターン化可能で英訳する上で問題になりそうな⑦敬語の標準化を追加する。

④縮約展開（例）

［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

⑤冗長圧縮（例）

［書替前］男/も//いれ/ば//女/も//いる
［書替後］男/も//女/も//いる

⑥構文組み換え（例）

［書替前］二/機種//合わせ/て//月/百/台//生産する
［書替後］二/機種/の//合計/月/産/は//百/台/だ

⑦敬語の標準化

次の例では，述語動詞は書き替え前の「なる」から書き替え後の「読む」に変化し，意味解析が容易になる。

［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)

３．３　疑似的日本語への書き替え

先に提案した疑似的日本語への書き替えのうち，訳語の指定という観点から，独立句的表現のみを本項目として扱うこととし，扱う範囲を発展させる。

⑧助詞相当語

「～＋助詞＋動詞＋て」（バスに乗って学校に行く→ by ）や「～＋の＋名詞＋助詞」（駅の前に花屋がある→ in front of ）など，英語の前置詞（句）に訳出すべきものを捉え直す。形態素解析の段階で機械的に実施すると副作用を招くことは先の報告[2, 3]で指摘した通りであり（例えば，「数人がバスに乗って残りが電車に乗る」では助詞相当語にしてはならない），本稿のように係り受け解析の後で文の構造に応じて実施すべきである。

⑨副詞相当語

独立不定詞（言うまでもなく → needless to say ）や非人称独立分詞（一般的に言えば → generally speaking）などへの訳出を意識して捉え直す。このタイプも形態素解析で行なえば副作用の恐れがある。

⑩連体詞相当語

日本語の句（複数文節）と対応づけられる英語の単語のうち，名詞の限定修飾に用いられるものを固定的に捉え直す。例えば，「印象に残る → impressive 」，「喜びにあふれた → joyful 」などが該当する。

⑪フレーズ

助詞相当語に似ているが，可変部分に入る語やそれの訳出に一定の条件を伴うものをルール化する。例えば，「驚いたことには → to one's surprise 」「失望したことには → to one's disappointment 」では，日本語は「動詞＋たことには」，英語も「 to one's ＋動詞訳の名詞派生形」とパターン化できる（ただし，使用できる動詞には制限がある）。なお，この例では英文生成時に one's を主語に応じて変形する必要がある。

３．４　主体的表現の固定化

話者の主観や判断等を表す表現は，要素合成的な手法で訳出することは不可能である。パターン的に捉えることにより，訳出を制御する方法を考える。

⑫接続様相時制表現

英語の仮定法などの特殊構文に訳出すべき日本語表現はかなりパターン化可能であると思われる。

［前］～する(動詞)/なら//～する(動詞)/のに。
［後］～する(動詞+仮定法過去)//～する(動詞 → "would V")。

⑬様相時制表現

英語の完了不定詞などに訳出すべき日本語表現もかなりパターン化可能であると思われる。

［前］～し/た/ようだっ/た。
［後］～する(動詞 → "seemed to have V[過去分詞]")。

４　おわりに

本稿では，先に提案した日本文書き替え処理を整理発展させ，汎用性と適用範囲の拡大を図る方法を提案した。現在，先の報告の940ルールを分類中であり，その結果については，処理系の構築と併せて，別途報告する。

＜謝辞＞
本稿をまとめるに当たり，ご討論くださった小見佳恵氏と阿部さつき氏（ルール記述の観点），土倉三津恵氏と中村三紀氏（処理の実現方法）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＞

[1] TMI-92 Proceedings of the Coference, Montreal, Canada, 1992
[2] 白井, 池原, 河岡：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，情処研報 NL-95-12 または信学技報 NLC-93-12, 1993
[3] 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, 1993
[4] S.Ikehara, M.Miyazaki, S.Shirai and A.Yokoo : An Evaluation Method for MT System and Its Application to ALT-J/E, ＡＩ学会誌 Vol.7, No.6, 1992
[5] 白井, 横尾, 木村, 小見：日本語従属節の依存構造に着目した係り受け解析，47情処全大 3M-1, 1993