英字新聞記事見出し翻訳の自動前編集による改善

吉見毅彦　佐田いち子
シャープ　ソフト事業推進センター
{yoshi,sata}@isl.nara.sharp.co.jp

要旨

英字新聞記事の見出しは通常の文の表現形式とは異なる特有の形式をしているため，従来の英日機械翻訳システムによる翻訳の質はあまり高くない．この問題に対して本稿では，見出しを通常の表現形式に書き換える自動前編集系を既存のシステムに追加することによって翻訳の質の改善を図る．書き換えの精度は，見出し特有表現の典型例であるbe動詞の省略現象を含む見出しを対象とした場合，再現率81.2%，適合率92.0%であった．

1　はじめに

近年，WWWを通じて英字新聞記事に接する機会が増えてきたことに伴い，より正確に英文記事を日本語に翻訳する必要性が高まってきている．新聞記事は見出しと本文から構成されるが，見出しは記事の最も重要な情報を伝える表現であるため，見出しを正確に翻訳することは他の表現の翻訳に比べてより重要である．見出しは，できるだけ少ない文字数でできるだけ多くの情報を伝えるために，通常の文の表現形式とは異なる特有の形式をしている．このため，従来の英日機械翻訳システムでは適切に翻訳できないことが多い．その原因は主に，見出し特有の表現形式の構文解析が適切に行なえないことにある．様々な種類のテキストを扱うことを前提に開発された機械翻訳システムの構文解析規則は，標準的な表現に対応することを目的に記述されているからである．

既存の構文解析規則で適切に扱えない表現への対応策の選択肢としては，特殊な表現形式が扱えるように構文解析規則を拡張するアプローチと，既存の構文解析規則は変更せず，既存の規則でも適切に処理できるように入力表現を書き換える新たなモジュールを設けるアプローチ[1, 2]が考えられる．実務で運用されている機械翻訳システムでは構文解析規則の規模は非常に大きくなっているため，既存の規則との整合性を保ちながら新たな規則を追加することは容易ではない．また，特殊な表現を扱うための規則を追加すると，規則の汎用性が損なわれる恐れがある．これに対して，既存の規則には手を加えず，入力表現を書き換える前編集系を開発する方が，構文解析規則の汎用性を維持することができるという点と，書き換え結果が既存の規則で正しく解析できるかどうかを人手で判断することは比較的容易であるという点で望ましい．

汎用的な用途の機械翻訳システムでは適切に処理することが難しい新聞記事見出しの翻訳の質を改善するために本稿では，主に語彙情報に基づいて入力表現を書き換えるアプローチを採り，比較的簡単な書き換え規則を記述するだけでも妥当な書き換えが行なえることを示す．なお，本研究の前編集系(英々変換系)は，英字新聞記事見出しの書き換え専用に設計したものではなく，通常の表現も対象とした一般的な枠組である．実際，見出し以外の表現に対する書き換え規則として約160規則を実装している．

2　英々変換系

本節で述べる英々変換系を組み込んだ機械翻訳システムにおける解析処理の流れを図1に示す．このシステムでは，形態素解析終了後に英々変換を実行して英語表現を書き換えた後，書き換えた部分の形態素解析を行ない，入力全体の形態素解析結果を構文解析系に送る．一度目の書き換え結果に対する構文解析が失敗した場合[1]，処理の制御は英々変換に戻る．再度英々変換を行なうときには，各書き換え規則に記述されている規則の信頼度(後述)に従って，一度目の英々変換では用いなかった規則を新たに適用したり，逆に一度目の処理で行なった書き換えを取り消したりする[2]．英々変換系での処理は，形態素解析結果に対して先頭から順に書き換え規則の適用条件との照合を行なっていき，適用条件が満たされる部分を順次書き換えていく．

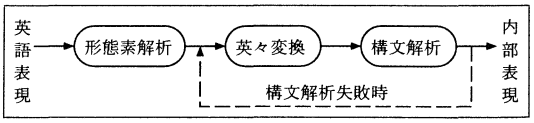

図 1: 解析処理の流れ

書き換え規則には，適用条件と書き換え操作の他，制御情報として適用抑制規則集合と信頼度を記述することができる．

[1]本稿では，入力全体を覆う構文構造が生成できないことを構文解析の失敗と呼ぶ．

[2]二度目の構文解析に失敗した場合には，断片的な構文構造を内部表現とする．

適用条件としては，書き換え対象候補が存在することを示す基本的な手がかりとなるキー表現の語彙属性と，その前後に存在する表現の属性を指定する．単なる属性の比較としては記述できない条件の照合は関数呼び出しによって行なう．

書き換え操作には，英語表現を追加，削除する操作と，システムに固有の編集記号を付加する操作がある．実験に用いたシステムでは，利用可能な編集記号として，多品詞語の品詞を指定する記号や，節や句の範囲や従属先を指定する記号など54種類が定義されている．これらの編集記号を付加することによって解釈の曖昧性が減るので，解析の精度と速度の向上が期待できる．

ある規則Rの適用抑制規則集合は，Rの適用を抑える他の規則に関するメタ条件を表し，規則Rは，その適用抑制規則集合に記述されている識別番号の規則が既に適用されているときには適用されない．規則Rの適用抑制規則集合には，Rの書き換え対象と重複する部分を書き換えようとする規則だけでなく，書き換え対象がRのものと重複しない規則を含めてもよい．

規則には，その信頼性が高く，規則の適用によって解釈の質が向上することがほぼ確実な規則もあれば，信頼性があまり高くない規則もある．信頼度は，このようなことを考慮して，信頼性があまり高くない規則による悪影響を抑えるために設定したものであり，規則の信頼性に応じてA，B，Cのいずれかを記述する．信頼度Aの規則は最初の構文解析の前に適用し，構文解析に失敗してもこの規則による書き換えは取り消さない．規則に信頼度Aを与えるのは，この規則を適用しないと構文解析に失敗することがほぼ確実であり，たとえこの規則によって書き換えた表現の構文解析に失敗して断片的な構文構造しか得られなかったとしても，この規則を適用しない場合の(断片的な)構文構造よりも質が高いと期待される場合である．信頼度Bの規則は最初の構文解析の前に適用するが，最初の構文解析に失敗したとき，この規則による書き換えは取り消す．信頼度Cの規則は最初の構文解析の前には適用せず，最初の構文解析に失敗したときに初めて適用する．

3 新聞記事見出しの調査

新聞記事の見出しでは文字数を節約するために，時制や態などに関する情報の省略(典型的にはbe動詞の省略)，冠詞の省略，縮約語の使用，等位接続詞のコンマでの代用など様々な工夫がなされている[4]．このような工夫が実際の見出しでどの程度見られるのかをつかむために284件の見出しを調査した．その結果，見出し特有の表現形式のうちbe動詞の省略が284件中73件と最も頻繁に見られた．be動詞と組み合わせることによって初めて完全な形式の定形述語となる表現を準述語と呼ぶことにし，be動詞が明示されていない見出しの件数を準述語別に集計した結果を表1に示す．

これら73件の見出しを我々の実験システムで翻訳した場合にどのような問題が生じるのかを調べた．その結果，文と解釈すべき見出しが名詞句とみなされることが63件(86.3%)と最も多かった．例えば次の見出しH1は，準述語"to inspect"を"Agency"に従属させた名詞句の解釈が生成されたが，H1'のように"to inspect"を"is to inspect"とみなして，予定を表す文と解釈しなければならない．

表1: be動詞の省略件数

準述語	件数
過去分詞	24
to不定詞	17
現在分詞	12
形容詞(叙述用法)	11
前置詞句	6
多語動詞の一部	3
合計	73

(H1) Agency to inspect health of 17 banks

(H1') Agency *is* to inspect health of 17 banks

また，次の見出しH2のように準述語に過去形か過去分詞かの曖昧性がある場合に過去形と誤って解釈される見出しが4件(5.5%)あった．

(H2) Japan's hope finally dashed

(H2') Japan's hope *was* finally dashed

次の見出しH3のように全体を名詞句と解釈することも文と解釈することもできず，構文解析に失敗する見出しは6件(8.2%)存在した．

(H3) Japanese wonder if Jamaica on the ball

(H3') Japanese wonder if Jamaica *is* on the ball

省略箇所に人手でbe動詞を補った[3]見出しを同実験システムで翻訳して書き換え前後の翻訳結果を比較し，どの程度の改善が期待できるのかを調べた．その結果，73件の見出しのうち69件が書き換え前の見出しの翻訳の質に比べて同等かより高かった[4]．このことから，be動詞が明示されていない見出しを英々変換系で書き換えることによって翻訳の質が向上すると期待される．

4 見出しの書き換え規則

見出しにbe動詞を補うために，準述語候補をキー表現とし，その前後に存在する表現に対する経験的な制約を適用条件とする書き換え規則を記述した．適用条件には，be動詞を補うべき見出し全般に見られる特徴を反映した共通条件と，各準述語候補によって異なる個別条件がある．ある準述語候補に関する書き換えは，共通条件のすべてとその準述語候補の個別条件のすべてが満たされるときに実行される．

[3]準述語がto不定詞の場合は，be動詞を補う代わりにtoをwillに書き換えた．

[4]残り4件の翻訳が悪化した原因は，辞書または構文解析規則の不備であり，書き換え自体は妥当であった．

4.1 各準述語に共通な適用条件

be 動詞が明示されていない見出しに be 動詞を補う処理は，be 動詞と準述語候補を組み合わせて完全な形式の定形述語を復元し，見出しのある部分を構文解析で一つの節と解釈するための処理である．そのような解釈が可能になるためには，復元された定形述語の主語になる名詞句が見出し中に出現していなければならない．主語候補の名詞句は，準述語候補の直前に現れるか，準述語候補の直前に副詞が存在し，その副詞の直前に現れると仮定し，条件 1 を置く．

条件 1 準述語候補の前方に主語候補が存在する．

見出しにはそれほど複雑な名詞句は現れないと考えられる[5]ことから，本研究では関係節などを支配しない単純な構造を持つ名詞句だけを主語候補とした．

次の条件 2 は，例えば準述語候補が to 不定詞の場合，“face to face” や “option to DO” のように実験に用いたシステムの辞書に to 不定詞と主語候補が連語として登録されている場合，主語候補と to 不定詞の結び付きを優先し，書き換えを行なわないようにするための条件である．

条件 2 準述語候補と主語候補が連語を構成しない．

この条件によって，“Effective option to treat ear infections” などに対する書き換えが抑えられる．

見出しに be 動詞を補うと，それまでは節と解釈できなかった部分が節と解釈できるようになる．そのような節をここでは潜在節と呼ぶ．条件 3 は，見出しのある部分の解釈として潜在節と通常の節が可能であるとき，後者を優先し，書き換えを抑えるためのものである．

条件 3 潜在節と構文的に競合する節が存在しない．

例えば “Navy hopes to lure recruits with incentives” という見出しに対しては，be 動詞を補い “are to lure” を主辞とする潜在節としての解釈が構文的には可能であるが，既存の定形述語 “hopes” を主辞とする通常の節としての解釈が存在するので書き換えは行なわない．また “Agency to inspect health of 17 banks goes into action” では，“Agency” の直後に be 動詞を挿入することは構文的に不可能であり，“goes” を主辞とする通常の節としての解釈しか許されない．

条件 3 は，見出し中に節が存在しても，それが潜在節と構文的に競合しない場合には満たされる．例えば 3 節で挙げた見出し H3 には “wonder” を主辞とする節が存在するが，この節と潜在節 “Jamaica on the ball” とは節境界を示す接続詞 “if” によって分離されているので競合しない．

節境界は接続詞や関係詞やコンマなどの節境界標識によって明示されている場合もあれば明示されていない場合もあるが，本研究では接続詞で明示されている場合のみを扱った．さらに，見出しは高々二つの節から構成され，かつ一方が他方の中央埋め込み節ではないものと仮定した．条件 3 が満たされるかどうかを厳密に判定するためには構文解析を行なう必要があるが，ここでは次のような単純な手順で行なう．

(1) 見出し中に節境界標識の接続詞が存在し，それによって見出しが二分される場合，そのうち潜在節を含む部分を手順 (2) の処理対象とする．節境界標識が存在しない場合，見出し全体を手順 (2) の処理対象とする．

(2) 処理対象の先頭から順に，述語になり得る定形動詞を探していく．もし見つかれば，その述語候補と人称，数が一致する名詞を主辞とする名詞句がその前方に存在するかどうかを調べる[6]．もしそのような名詞句が存在すれば，それを主語とみなし，条件 3 が満たされないものとする．ただし，準述語候補が過去分詞の場合，過去分詞形と現在形または過去形が同形であっても準述語候補を述語候補とはしない．

4.2 to 不定詞に関する適用条件

準述語候補ごとに異なる適用条件の例として，to 不定詞に関する規則の適用条件を示す．to 不定詞に関しては，条件 1 ～ 3 の他に次の条件 4 が満たされるとき，書き換えを行なう．

条件 4 to 不定詞の前方に “from”，“for”，“too” が存在しない．

to 不定詞の前方に “for” などの語が存在する場合，“for ～ to ～” のような構文的枠組を与える表現である可能性が高いと考えられるため，条件 4 によってこの構文的枠組を維持する．

4.3 be 動詞の屈折形生成

適切な書き換えを行なうためには，be 動詞を主語候補の直後に挿入すべきかどうかを判定するだけでなく，挿入する場合には be 動詞の屈折形を決定する必要がある．屈折形は，人称，数，時制，相情報などに基づいて決めなければならないが，ここでは，時制は現在とし，主語候補の主辞の人称と数に従う区別だけを行なうことにし “am”，“are”，“is” のいずれかとする．新聞記事見出しでは過去の事柄が現在形で表される場合も少なくない [3, 4] ので，現在時制とすることはそれほど不自然ではないと考えられる．

4.4 規則の制御情報

書き換え規則を記述するために調査した 284 件の見出しでは，一つの見出しに対して二箇所以上に be 動詞を補う必要がある例はなかった．このため，ある準述語候補に関する書き換えが行なわれた場合，他の書き換えを行なわないようにした．すなわち，2 節で述べた，ある準述語候補に関する規則の適用抑制規則集合の要素は，その規則以外のすべての準述語候補に関する規則の識別番号とした．

規則の信頼度については，すべての見出し書き換え規則について B とし，書き換えられた見出しの構文解析に失敗したときには書き換えを取り消して元の表現に戻すようにした．

[5] 調査した 284 件の見出しのうち関係節を支配する名詞句を含むものは 2 件であった．

[6] 名詞句の検索は条件 1 での主語候補検索と同じ手続きを用いて行なう．

表 2: 実験結果

準述語	訓練データ		試験データ	
	再現率	適合率	再現率	適合率
過去分詞	87.5%(21/24)	100%(21/21)	87.8%(36/41)	94.7%(36/38)
to 不定詞	100%(17/17)	100%(17/17)	88.2%(15/17)	88.2%(15/17)
現在分詞	91.7%(11/12)	100% (9/9)	62.5% (5/8)	100% (5/5)
形容詞 (叙述用法)	81.8% (9/11)	90.0% (9/10)	69.2% (9/13)	90.0% (9/10)
前置詞句	83.3% (5/6)	83.3% (5/6)	66.7% (2/3)	66.7% (2/3)
多語動詞の一部	66.7% (2/3)	100% (2/2)	66.7% (2/3)	100% (2/2)
合計	89.0%(65/73)	96.9%(63/65)	81.2%(69/85)	92.0%(69/75)

5 実験結果

規則記述のために調査した見出し 284 件 (訓練データ) を対象として行なった実験の結果と，この訓練データとは異なる 312 件の見出し (試験データ) を対象として行なった実験の結果を表 2 に示す.

書き換えられるべきであるのに書き換えられなかった見出しについて，その原因を分析した．訓練データでの 8 件の書き換え漏れのうち 1 件は辞書未登録語が存在したことによるものであり，残りの 7 件が書き換え規則の不備によるものであった．この 7 件のうち 5 件は，適用条件 3 が満たされるかどうかの判定を誤ったことによるものであった．5 件中 3 件は，“Typhoon kills two, 5 missing” のように節境界がコンマによって示される場合に，実際には二つの節から構成される見出しを一つの節から成ると誤解釈し，潜在節 (5 missing) と競合しない節 (Typhoon kills two) を競合するとみなしたことによるものであった.

試験データでの 16 件の書き換え漏れの原因の内訳は，辞書未登録語など形態素解析での問題によるものが 3 件，書き換え規則の不備によるものが 13 件であった．13 件中 9 件は条件 3 の判定誤りによるものであり，その 9 件のうち 7 件については節境界を正しく捉えられないことが原因であった.

書き換えられるべきでないのに誤って書き換えられた見出しは，訓練データで 2 件，試験データで 6 件と書き換え漏れの件数に比べて少なかった．書き換え誤りが生じた原因は，訓練データではすべて書き換え規則の不備によるものであり，試験データでは 3 件が形態素解析での問題によるもの，3 件が書き換え規則の不備によるものであった.

4.4 節で述べたように，書き換えは一見出しについて一箇所でしか行なっていない．訓練データには二箇所で書き換えを行なう必要がある見出しは存在しなかったが，試験データでは次の見出し H4 のように二箇所の書き換えが必要な見出しが 2 件あり，後方の準述語に対して be 動詞を補うことができなかった.

(H4) 10 killed, 6 hurt in Pa.

(H4') 10 *were* killed, 6 *were* hurt in Pa.

書き換え規則にはすべて信頼度 B を与えているため，書き換え結果に対する構文解析が失敗すると，一度行なった書き換えが取り消されるが，今回の実験では，書き換え後の構文解析に失敗する見出しは，訓練データ，試験データいずれにおいても存在しなかった.

6 おわりに

本稿では，標準的な表現を主な対象とした機械翻訳システムには適切な翻訳を行なうことが難しい英字新聞記事見出しを通常の表現に書き換えることによって翻訳の質を改善する方法を示した．見出し特有の表現形式のうち比較的高い頻度で見られる be 動詞の省略現象に対処するための書き換え規則を記述し実験を行なった結果，試験データに対して再現率 81.2%，適合率 92.0% の精度が得られた.

本稿では触れなかったが，コンマが等位接続詞として用いられてることが原因で適切な翻訳が得られないことも多い．今後，コンマに関する書き換え規則を記述するなど規則の拡張を行なう予定である.

謝辞 英々変換系の実装を行なって頂いたシャープ (株) ソフト事業推進センターの関谷正明さん (現在，同社設計技術開発センター) と，書き換え結果の評価などを担当して頂いた共栄ビジネス・サービス (株) の青木直子さんと，草稿に対して数多くの有益なコメントを頂いた Dr. Jiri Jelinek に感謝します.

参考文献

[1] 金淵培, 江原暉将. 日英機械翻訳のための日本語長文自動短文分割と主語の補完. 情報処理学会論文誌, Vol. 35, No. 6, pp. 1018–1028, 1994.

[2] 白井諭, 池原悟, 河岡司, 中村行宏. 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, pp. 12–21, 1995.

[3] 白井諭, 大山芳史, 中尾嘉孝, 西垣万亀子, 上田洋美, 小見佳恵. 英文記事ヘッドラインの特徴について. 第 54 回全国大会論文集, 情報処理会, 1997. 4B-1.

[4] 上野田守, 布施敏夫. 新聞英語. 朝日実務英語シリーズ. 朝日出版社, 1978.