英字新聞記事見出し翻訳の自動前編集による改良

吉見　毅彦[†]　佐田いち子[†]

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見出しの翻訳の品質はあまり高くない．この問題に対して本研究では，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る解決を目指している．見出しを通常の表現形式に書き換えれば，より品質の高い翻訳が，システムの既存部分にほとんど変更を加えることなく得られる．例えば “Sales up sharply in June” という見出しは通常のシステムには受理されない可能性が高いが，“Sales *were* up sharply in June” のように be 動詞 “were” を補えば従来のシステムでも適切な翻訳が得られるようになる．本稿では，見出し特有表現の典型例の一つである be 動詞の省略現象を対象とし，be 動詞が省略されている見出しに be 動詞を正しく補うための書き換え規則を，形態素解析と粗い構文解析によって得られる情報に基づいて記述する．この方法を，我々が開発している英日翻訳支援システム Power E/J に組み込み，未知データの見出し 312 件を対象として実験を行なったところ，再現率 81.2%，適合率 92.0%の精度が得られた．

キーワード:　機械翻訳，自動前編集，原文書き換え，新聞記事見出し

Improvement of Translation Quality of English Newspaper Headlines by Automatic Preediting

Takehiko Yoshimi[†] and Ichiko Sata[†]

Since the headlines of English news articles have a characteristic style, different from the styles which prevail in ordinary sentences, it is difficult for MT systems to generate high quality translations for headlines. We try to solve this problem by adding to an existing system a preediting module which rewrites the headlines to ordinary expressions. Rewriting of headlines makes it possible to generate better translations which would not otherwise be generated, with little or no changes to the existing parts of the system. While most MT systems would not probably accept, for example, the headline “Sales up sharply in June”, they would be able to generate a satisfactory translation of the expression “Sales *were* up sharply in June” where the verb “were” has been inserted. Focusing on a conspicuous phenomenon, the absence of a form of the verb of ‘be’, we have described rewriting rules for putting properly the verb ‘be’ into the headlines, based on information obtained by morpholexical and rough syntactic analysis. We have incorporated the proposed method into our English-to-Japanese MT system *Power E/J*, and carried out an experiment with 312 headlines as unknown data. Our method has satisfactorily marked 81.2% recall and 92.0% precision.

KeyWords:　*Machine Translation, Automatic Preediting, Rewriting, Headline*

† シャープ (株) ソフト事業推進センター, Software Business Development Center, SHARP Corp.

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[1]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一層重要である．

英字新聞記事の見出しは，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や，読者の注意を引くために，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場合が多い．その原因は主に，見出し特有表現の構文解析を適切に行なうための構文解析規則が，様々な種類や分野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では記述されていないことにあると考えられる．

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原言語の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が考えられる．後者のアプローチとして，長い文の構文解析が失敗しやすいという問題に，長文を複数の短文に分割することによって対処する方法 (金淵培 江原暉将 1994) や，書き換えを行なうべきかどうかの判定精度を高めるために，完全な構文情報が得られる構文解析終了後にまで書き換え規則の適用を遅らせる方法[2](白井諭, 池原悟, 河岡司, 中村行宏 1995) などがこれまでに示されている．

実際に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，原言語の表現を書き換える前編集系を新たに開発する方が，書き換え結果が既存の構文解析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や，規則の汎用性を維持することができるという点でシステムの開発，維持上望ましい．

本研究では，従来の機械翻訳システムによる新聞記事見出し翻訳の品質が低いという問題に対して自動前編集モジュールを設けるアプローチを採り，浅いレベルの手がかりに基づいて原言語の表現を書き換えることによってこの問題を解決することを目指している．自動前編集に

1 テキストから重要な文を選択するテキスト抄録システムにおいて，見出しを最も重要な文であるとみなす考え方 (仲尾由雄 1997; 吉見毅彦, 奥西稔幸, 山路孝浩, 福持陽士 1999) がある．

2 この方法は，日英間の構造的な差異を調整し，より自然な翻訳を生成するために構文構造を書き換える方法 (長尾眞 辻井潤一 1985) に近いと考えられる．

よる見出し翻訳の品質改善の一例として本稿では，見出し特有表現のうち比較的高い頻度[3]で見られる be 動詞の省略現象に対象を絞り，be 動詞が省略されている見出しに be 動詞を正しく補うための書き換え規則を，形態素解析と粗い構文解析[4]によって得られる情報に基づいて記述し，これらの書き換え規則によって適切な書き換えが行なえることを示す．

本稿の対象は英字新聞記事見出しという限定されたものであるが，英字新聞記事は英日機械翻訳システムの一般利用者が日々接することが多いテキストの一つであるため，実用的なシステムにおける見出し解析の重要性は高い．また，本稿の目的は be 動詞を補うことによって見出し解析の精度を向上させることにあり，書き換えた見出しの翻訳が日本語新聞記事の見出しの文体に照らし合わせて適切であるかどうかは本稿の対象外である．

2　英々変換系

2.1　英々変換の枠組

本節で述べる自動前編集系 (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，表現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に失敗した場合[5]，処理の制御は英々変換に戻る．再度英々変換を行なう場合には，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[6]．

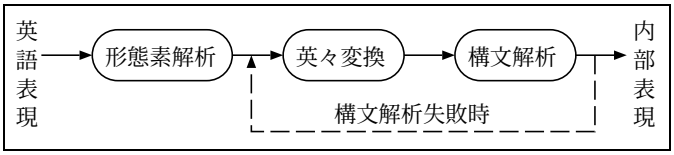

図 1　解析の流れ

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．この英々変換系は，新

3 284 件の見出しを対象とした我々の調査で確認された見出し特有の表現 (上野田守 布施敏夫 1978) は，be 動詞の省略を含むものが 73 件 (25.7%)，等位接続詞のコンマでの代用を含むものが 25 件 (8.8%)，“say” のコロンでの代用を含むものが 4 件 (1.4%) などである．ただし，現在形で過去の事象を表す表現や冠詞の省略などは今回の調査では考慮しなかった．

4 具体的には，4.1 節で述べる手続きによる処理を指す．

5 本稿では，入力表現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．

6 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．

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，挿入語句を識別する規則や長い表現を分割する規則などが記述されている．

2.2　書き換え規則の形式

書き換え規則には，次に示すように，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．

(識別番号，適用条件，書き換え操作，適用抑制規則集合，信頼度)

書き換え対象候補が適用条件を満たすかどうかの判定は，書き換え対象候補の形態素語彙属性や構文属性を調べる手続きを用いて行なう．

書き換え操作には，英語表現を追加，削除，置換する操作と，システム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．編集記号の付加によって解釈の曖昧性が減るため，解析の精度と速度の向上が期待できる．

ある規則 R に与えられている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場合には適用されない．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翻訳品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る．各規則には，その信頼性に応じて A，B，C のいずれかの信頼度を与える．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) 構文構造から生成される翻訳よりも高い品質の翻訳が生成される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場合，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場合に初めて適用する．

簡単な書き換え規則の例を図 2 に示す．この規則は新聞記事見出しの書き換え用ではないが，倒置文の構文解析が失敗することに対処するためのものである．この規則は，現在着目している語が入力文の先頭語であり (`p == 1`)，着目語の (細分類) 品詞候補として過去分詞の可能

性があるが名詞の可能性がなく，さらに着目語の直後の語が “is” であるときに適用される．この適用条件が満たされると，着目語の先頭文字を小文字に変換し，“What is” という語句を着目語の直前に挿入する．この処理によって，例えば “Affiliated is the parent company of Globe Newspaper Co.” という文が “What is affiliated is the parent company of Globe Newspaper Co.” に書き換えられる．

```
(301,
 (p == 1 &&
  word_class(p, past_participle) == TRUE &&
  word_class(p, noun) == FALSE &&
  word(p+1, "is") == TRUE),
 (to_lower(p), insert(p-1, "What is")),
 (),
 A)
```

図 2　書き換え規則の例

3　英字新聞記事見出しの調査

英字新聞記事の見出しでは，述語の時制や態などに関する情報の省略や，冠詞の省略，略語の使用，等位接続詞のコンマでの代用など文字数を節約するための様々な工夫がなされている (上野田守・布施敏夫 1978)．本研究では，これら見出し特有の現象のうち時制情報などの省略に関連する be 動詞の省略現象を扱うことにし，ロイター記事 (Lewis 1997) の見出し 284 件を対象として次の四項目の調査を行なった．

(1)　be 動詞が省略されているのはどのような場合か．

(2)　be 動詞が省略されている見出しをそのまま我々の実験システムで翻訳した場合の翻訳品質はどの程度か．

(3)　be 動詞が省略されている見出しに be 動詞が適切に補われた場合，項目 (2) の翻訳に比べてどの程度品質が改善されるか．

(4)　形態素語彙，構文上のどのような現象が，be 動詞が省略されている見出しとそうでない見出しを区別する手がかりとなるか．

本節では項目 (1)，(2)，(3) についての調査結果を示し，項目 (4) については 4.1 節で述べる．

3.1　キーの種類

be 動詞の省略は調査対象の見出し 284 件のうち 73 件において見られた．一般に be 動詞の省略は一つの見出しにおいて複数箇所で行なわれうるが，これら 73 件の見出しでは一箇所での省略しか行なわれていなかった．通常の表現形式では be 動詞と結び付けられ全体で定形述語と解

釈される表現をここではキーと呼ぶ．73 件の見出しに出現したキーは，受動態用法の過去分詞，to 不定詞，現在分詞，叙述用法の形容詞，前置詞句，複合動詞の構成素の六種類であった．ここで複合動詞の構成素とは，be 動詞と結合して複合動詞となる語句を意味し，例えば “be up” における “up” などである．各キーごとに，それが出現した見出しの例 (上段) と，省略箇所に人手で be 動詞を補った表現 (下段)，さらに出現件数を表 1 に示す．表 1 では，キーに下線を付し，人手で補った be 動詞を斜字体で示している．

表 1　be 動詞が省略されている見出しの例と件数

キー	例	件数
過去分詞	(H1)　Calabrian bank taken over by commissioners (H1')　Calabrian bank *was* taken over by commissioners	24
to 不定詞	(H2)　U.S. official to visit Japan as trade row grows (H2')　U.S. official *is* to visit Japan as trade row grows	17
現在分詞	(H3)　Senate preparing for new U.S. budget battle (H3')　Senate *is* preparing for new U.S. budget battle	12
形容詞	(H4)　Early gulf cash soybeans slightly firmer (H4')　Early gulf cash soybeans *are* slightly firmer	11
前置詞句	(H5)　No prospect in sight of EC budget accord (H5')　No prospect *is* in sight of EC budget accord	6
複合動詞の構成素	(H6)　Pan Am February load factor up (H6')　Pan Am February load factor *was* up	3
合計		73

3.2　従来システムによる見出し翻訳の品質

従来システムによる見出し翻訳の問題点を明らかにしておくために，be 動詞が省略されている 73 件の見出しをそのまま我々の実験システムで処理し，その結果を評価した．評価の際に翻訳のどの部分を対象とするかに関して，見出し全体を対象とすることと，キーに直接関連がある部分だけを対象とすることが考えられる．ここでは be 動詞の省略が翻訳品質に及ぼす影響に関心があるため，後者の局所的な評価を行なった．評価値は合格か不合格かの二値とした．合否判定は，翻訳が文法的であるかという観点と，文法的な翻訳の場合，翻訳の意味が元の見出しの意味と一致しているかという観点から行なった．翻訳の文体が新聞記事見出しとして適切であるかどうかは考慮しなかった．合格と認める翻訳は文法的であり意味的に等価なものである．be 動詞が省略されいることが原因で文法的でないか意味的に等価でない翻訳が生成された

場合は不合格とした.

表 2　be 動詞が省略されている見出しの翻訳品質

キー	合格	不合格
過去分詞	16	8
to 不定詞	1	16
現在分詞	10	2
形容詞	6	5
前置詞句	2	4
複合動詞の構成素	1	2
合計	36	37

評価結果を表 2 に示す．キー全体では，合格と不合格の件数はそれぞれ 36 件と 37 件でほぼ同じであるが，キー別に見ると，現在分詞の場合には 12 件中 10 件が合格したのに対して to 不定詞の場合には 17 件中 16 件とほとんどが不合格となった.

キーが現在分詞である場合にほとんどが合格となるのは，キーをその前方に存在する名詞句に従属させた解釈が，be 動詞を補った場合の翻訳とほぼ等しい意味を伝えている場合，その解釈を合格としたためである．例えば表 1 の見出し (H3) は本来 be 動詞を補って (H3') のように解釈されるべきであるが，(H3) の翻訳「新しい米国の予算の戦いに備えて準備している上院」は，(H3') の翻訳「上院は，新しい米国の予算の戦いに備えて準備している」と意味的に等しいので合格とした．他のキーについてもこのような場合には合格とした.

キーが to 不定詞の場合には，キーをその前方の名詞句に従属させると，be 動詞を補った場合とは意味が大きく異なる翻訳が生成された．不合格となった 16 件はすべて，to 不定詞が「...するための」と訳され，本来伝えられるべき予定や運命などの意味に解釈することができなかった．例えば表 1 の見出し (H2) は予定を表す文と解釈しなければならないが，(H2) の翻訳「日本を訪問するための米国の職員」はそのように解釈できない.

元の意味と大きく異なる意味を伝える翻訳が生成されたもう一つの例は，過去分詞形と解釈されるべきキーが定形 (現在形または過去形) と解釈された場合である．例えば次の見出し (H7) では “sued” が過去形とみなされ，対象格と解釈されるべき “Three” が主格と解釈された.

(H7)　Three sued over ball valves for nine mile point

規則動詞や一部の不規則動詞の過去分詞形は定形と表記が同一であるため，このような誤りが生じる見出しの件数は少なくない.

不合格と判定された 37 件の見出しを正しく翻訳するためには，be 動詞を補わなければならない．これに対して，合格と認められた 36 件については，be 動詞を補った場合の翻訳とほぼ等しい意味を伝える翻訳が生成されるので，英日翻訳の見地からは be 動詞補完は可能ではあるが必要ではないという捉え方もできるかも知れない．しかし，これら 36 件の見出しも読者には

通常 be 動詞を補って理解されるので，本研究では見出しの構文的解釈の見地から be 動詞を補う対象に含める．従って本稿では，be 動詞が省略された見出しとは，be 動詞を補うべき見出しと be 動詞を補うことができる見出しを合わせたものを指している．

3.3　期待される改善度

be 動詞を補うことによってどの程度の品質改善が期待できるかをあらかじめ確認しておくために，73 件の見出しに人手で be 動詞で補った表現を実験システムで処理し，be 動詞が補われていない見出しの翻訳と比較した．評価値は，改善，同等，改悪の三値とした．3.2 節の評価で合格となった見出しの翻訳が改善されているとは，be 動詞を補うことによってキーとその前方の名詞句との構文的関係が改善されたことを意味する．例えば見出し (H3) の翻訳「新しい米国の予算の戦いに備えて準備している上院」と比較して，be 動詞を補った表現 (H3') の翻訳「上院は，新しい米国の予算の戦いに備えて準備している」はより適切であるとみなす．改善箇所と改悪箇所の両方が存在している場合，あるいは改善も改悪も見られない場合には同等とする．

表 3　be 動詞補完による翻訳品質の改善度

キー	合格			不合格		
	改善	同等	改悪	改善	同等	改悪
過去分詞	14	0	2	7	1	0
to 不定詞	1	0	0	16	0	0
現在分詞	9	0	1	1	1	0
形容詞	5	0	1	5	0	0
前置詞句	2	0	0	4	0	0
複合動詞の構成素	1	0	0	2	0	0
合計	32	0	4	35	2	0

評価結果を表 3 に示す．3.2 節の評価で合格となった見出し 36 件のうち 32 件と，不合格となった見出し 37 件のうち 35 件について，より適切な翻訳が得られている．このことから，英々変換によって be 動詞を正しく補うことができれば，システムの既存部分に変更を加えることなく見出し翻訳の品質が改善されると期待できる．なお，合格となった見出し 36 件のうち 4 件の翻訳品質が低下しているが，この原因は辞書または構文解析規則の不備であり，本稿の主要目的である be 動詞の補完とは直接の関係はない．

4　be 動詞補完規則の記述

be 動詞補完精度の評価指標には，補完漏れ件数の少なさを示す再現率と不要な補完件数の少なさを示す適合率を用いるが，規則の記述方針として，漏れを減らすことよりも不要な補完を抑えることを重視した．その理由は，不要な補完が行なわれた場合，構文構造と意味が大き

く変化するため悪影響が出るのに対して，3.2 節で述べたように，be 動詞が省略されている見出し 73 件のうち 36 件については補完漏れが生じた場合でもある程度の品質の翻訳が得られることなどである．

4.1　適用条件

本研究で設定した適用条件は，be 動詞が省略されている見出しとそうでない見出しを区別する一般的な手がかりになりうる現象を 284 件の見出しにおいて分析した結果に基づいており，以下で説明する形態素語彙，構文上の四条件から主に構成されている．適用条件には，これら一般的な条件の他に，例えば “of” など特定の前置詞で導かれる前置詞句を処理対象外とする条件など，語彙に依存した個別条件も若干含まれる．

4.1.1　キー前方での名詞句の存在

be 動詞が省略されている見出しでは，キーの前方に名詞句が存在する．より具体的には名詞句は，表 1 の見出し (H1) などのようにキーの直前に現れるか，見出し (H4) のようにキーの直前に副詞が存在しその副詞の直前に現れる場合がほとんどであるので，次の条件 1 を設ける．

条件 1　キー候補の直前に，あるいはキー候補直前の副詞の直前に名詞句が存在する．

見出しに現れる名詞句は比較的単純な構造をしていることが多いので，次のような構造を持つ名詞句 NP を検出する手続きを記述した．

$$\begin{aligned} \mathrm{NP} &= \mathrm{NP0}\ (\mathrm{P}\ \mathrm{NP0})^{?} \\ \mathrm{NP0} &= (\mathrm{AV}^{?}\ \{\mathrm{AJ}|\mathrm{Ven}|\mathrm{Ving}\})^{?}\ \mathrm{N}^{+} \end{aligned}$$

ここで，P，AV，AJ，Ven，Ving，N はそれぞれ前置詞，副詞，形容詞，過去分詞，現在分詞，名詞を表し，上付き記号?と+はそれぞれ一回以下，一回以上の出現を意味する．

4.1.2　潜在節と競合する節の非存在

be 動詞とキー候補を組み合わせると定形述語が復元され，それまで通常の構文解析で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．このような節をここでは潜在節と呼ぶ．潜在節の主語になる名詞句は，前述の条件 1 を満たす名詞句である．例えば表 1 の見出し (H3) に be 動詞を補うと，(H3') のように定形述語 “*is* preparing” が復元され，見出し全体が名詞句 “Senate” を主語とする一つの節になる．

be 動詞補完の可否を決める手がかりの一つとして，潜在節と構文的に競合する節の有無に着目する．be 動詞が省略されている見出し (H1) ないし (H6) では潜在節と構文的に競合する節は存在しない．これに対して，次の見出し (H8) では潜在節と構文的に競合する節が存在する．

(H8)　Reagan hopes to lift Japan sanctions soon

この見出しにおける潜在節は “*are* to lift” を主辞とし “Reagan hopes” を主語とする節であるが，この解釈は既存の定形述語 “hopes” を主辞とし “Reagan” を主語とする通常の節としての解釈と構文的に競合する．このような場合には経験的に，通常の節としての解釈を優先することにする．

次の見出し (H9) では，“lost” の直前に be 動詞を挿入することは構文的に不可能であり，“was carrying” を主辞とする通常の節としての解釈しか許されない．

(H9)　Vessel lost in Pacific was carrying lead

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 be 動詞を補う．例えば表 1 の見出し (H2) には節 “trade row grows” が存在するが，この節と潜在節 “U.S. official *is* to visit Japan” とは節境界を示す接続詞 “as” によって分離されており競合しないので，(H2) は (H2') のように書き換える．

このような考察に基づき，潜在節と構文的に競合する節が存在しない場合に限り見出しに be 動詞を補うことにし，次の条件 2 を設ける．

条件 2　潜在節と構文的に競合する節が存在しない．

3.2 節の見出し (H7) では，“sued” を過去分詞形と解釈し be 動詞を補った潜在節 “Three *were* sued ···” と，“sued” を過去形と解釈した節 “Three sued ···” が構文的に競合する．このように，定形と同一表記の過去分詞がキー候補であり，このキー候補を定形と解釈した動詞を主辞とする節が潜在節と構文的に競合する場合には，条件 2 ではなく，後述する条件 3 に従うものとする．

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，接続詞で明示されている場合のみを扱う．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する．条件 2 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手順で行なう．

ステップ 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，そのうち着目しているキー候補を含む部分をステップ 2 の処理対象とする．節境界標識が存在しない場合，見出し全体をステップ 2 の処理対象とする．

ステップ 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[7]．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 2 が満たされないものとする．ただし，着目しているキー候補が定形と同一表記の過去分詞である場合，このキー候補を定形と解釈した動詞を述語候補とはしない．

7 名詞句の検索は条件 1 の判定で用いる手続きと同じ手続きを用いて行なう．

4.1.3　過去分詞に関する条件

キー候補に定形か過去分詞形かの曖昧性がある場合，キー候補を定形と解釈すれば，このキー候補を主辞とし潜在節と構文的に競合する節が存在することになるため，条件 2 に従うと，見出し (H7) などのように be 動詞を補うべき見出しに be 動詞が補われない．

この曖昧性の解消をここでは，キー候補直後の名詞句の有無と，キー候補の動詞型 (Hornby 1977) に基づいて行なう．キー候補を定形と解釈することは動詞の態を能動とみなすことであり，過去分詞形と解釈することはキー候補と be 動詞を組み合わせて受動態とみなすことであるが，キー候補が動詞型として SVOO 型も SVOC 型も持たない場合，キー候補の目的語が存在すれば，受動態と解釈することは構文的に不可能である．ここではキー候補直後の名詞句を目的語とみなし，キー候補の直後に名詞句が存在しなければ受動態と解釈して be 動詞を補う．

キー候補が動詞型として SVOO 型か SVOC 型を持つ場合は，キー候補の直後に名詞句が存在しても受動態と解釈できることがあるが，正確に判定するためには，キー候補直後の名詞句だけでなく，さらにその後方の名詞句の有無も認識する必要がある．定形か過去分詞形かの曖昧性に関しては，見出しではほとんどの場合後者と解釈していよいという経験則 (上野田守・布施敏夫 1978) があることと，粗い構文解析しか行なわない方針であることから，ここではキー候補が SVOO 型か SVOC 型を持つならば be 動詞を補うことにし，次の条件 3 を設ける．

条件 3　キー候補に定形か過去分詞形かの曖昧性がある場合，キー候補の直後に名詞句が存在しないか，キー候補が SVOO 型か SVOC 型を持つ動詞である．

この条件に従えば，見出し (H7) では “sued” の直後にその目的語となる名詞句が存在しないので，be 動詞が補われる．また，次の見出し (H10) では “offered” の直後に名詞句が存在するが “offered” は SVOO 型を持つので，be 動詞が補われる．

(H10)　U.K. money market offered early assistance

4.1.4　固定的表現の非存在

キー候補とその前方に存在する名詞句が連語や慣用句のように固定的な表現を構成する場合 be 動詞を補わない方がよいと考えられるので，次の条件 4 を設ける．

条件 4　キー候補が固定的表現の構成要素でない．

例えば次の見出し (H11) では，“need” と to 不定詞の間に結び付きがあると辞書に記述されているので，この結び付きを優先する．

(H11)　No need to state U.K. support for system — Lawson

ここでいう固定的表現とは，キー候補の辞書項目または条件 1 を満たす名詞句の主辞の辞書項目に記述されている表現だけでなく，“for ··· to ···” や “too ··· to ···” などのような相関語句も含む．従って，例えば to 不定詞がキー候補でありその前方に “for” や “too” などの語が存在する場合 be 動詞を補わない．

4.2　be 動詞の屈折形生成

適切な be 動詞補完を行なうためには，主語候補の直後すなわち条件 1 を満たす名詞句の直後に be 動詞を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ことも少なくない (白井諭, 大山芳史, 中尾嘉孝, 西垣万亀子, 上田洋美, 小見佳恵 1997; 上野田守・布施敏夫 1978) 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．

4.3　規則の制御情報

調査対象の 73 件の見出しでは複数箇所で be 動詞が省略されている例は存在しなかった．このため，形態素解析結果に対して先頭から順に適用条件との照合を行なっていき，あるキー候補に関して be 動詞補完が行なわれた場合，他のキー候補に関する補完を行なわないようにする．すなわち，2 節で述べた，あるキー候補に関する規則に与える適用抑制規則集合の要素は，その規則以外のすべてのキー候補に関する規則の識別番号とする．

規則の信頼度は，すべての be 動詞補完規則について B とし，be 動詞を補った見出しの構文解析に失敗した場合には補完を取り消して元の表現に戻す．

5　実験と考察

本節では，be 動詞補完規則作成のために調査した訓練データの見出し 284 件を対象として行なった実験の結果と，訓練データとは異なる試験データの見出し 312 件を対象として行なった実験の結果を示し，be 動詞補完が正しく行なえなかった見出しについてその原因を分析する．さらに，試験データにおいて正しく be 動詞が補えた見出しについて，その翻訳品質がどの程度改善されたかを検証する．4.2 節で述べたように，be 動詞の屈折形の決定は，時制などを考慮せず，主語候補の主辞の人称と数だけに基づいて行なっている．このため今回の評価では，システムが生成した be 動詞と人間が補った be 動詞とで，人称と数がそれぞれ一致していれば，時制などが適切でない場合でも正解とみなす．

5.1　実験結果

実験結果を表 4 に示す．表 4 によれば，訓練データで再現率 89.0%，適合率 97.0%の精度が得られ，試験データで再現率 81.2%，適合率 92.0%の精度が得られており，比較的簡単な規則でほぼ適切な補完が行なえている．

不要な補完は訓練データで 2 箇所，試験データで 6 箇所生じているが，これらは補完漏れ

表 4　実験結果

キー候補	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%(11/11)	62.5% (5/8)	100% (5/5)
形容詞	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)
複合動詞の構成素	66.7% (2/3)	100% (2/2)	66.7% (2/3)	100% (2/2)
合計	89.0%(65/73)	97.0%(65/67)	81.2%(69/85)	92.0%(69/75)

(訓練データで 8 箇所，試験データで 16 箇所) に比べて少なく，全体としては，不要な補完の抑制を優先するという 4 節で述べた規則記述における所期の目標が達成されている．キー別に見ると，訓練データにおいても試験データにおいても前置詞句の場合の適合率が最も低い．

5.2　失敗原因の分析

訓練データと試験データのそれぞれについて，補完漏れと不要な補完が生じた原因を調べた結果を表 5 に示す．

表 5　失敗原因の分析

原因	訓練データ		試験データ	
	補完漏れ	不要補完	補完漏れ	不要補完
形態素解析	1	0	3	3
条件 1	1	0	0	0
条件 2 (多品詞語)	2	1	2	0
条件 2 (節境界)	3	0	7	0
条件 3	0	0	0	1
条件 4	0	1	0	2
その他の条件	1	0	4	0
合計	8	2	16	6

5.2.1　補完漏れの原因

訓練データで生じた 8 箇所での補完漏れのうち 1 箇所は，キーになるべき語が辞書未登録語であったことによる形態素解析での問題であり，残りの 7 箇所での補完漏れが be 動詞補完規則の不備によるものであった．

7 箇所のうち 5 箇所は条件 2 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．その 5 箇所中 2 箇所は多品詞語の品詞解釈を誤ったことによるものであった．例えば次の見出

し (H12) では，この場合名詞と解釈すべき “imports” を動詞とみなし，“U.S. sugar” をその主語とみなす誤りが生じていたため，潜在節と競合する節が存在すると解釈された．

(H12)　U.S. sugar imports down in week — USDA

このような誤りに対しては品詞推定法 (竹田正幸 松尾文碩 1993; 竹田正幸, 須田淳一郎, 楠本典孝, 松尾文碩 1995) を導入することによって改善が可能であると考えられる．

5 箇所中残りの 3 箇所についての原因は節境界が正しく認識できないことにあった．条件 2 の判定で用いた節境界認識手続きでは一部の接続詞だけを節境界標識とみなしているために，次の見出し (H13) 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が一つの節から成ると誤解釈され，潜在節 “Africa *is* unable to pay its debts” と競合しない節 “OAU chief says” が競合すると判定されていた．

(H13)　Africa unable to pay its debts, OAU chief says

試験データで生じた 16 箇所での補完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 箇所，be 動詞補完規則の不備によるものが 13 箇所であった．13 箇所中 9 箇所は条件 2 の判定誤りによるものであり，その 9 箇所のうち 7 箇所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．

訓練データにおいても試験データにおいても，条件 2 の判定誤りが補完漏れの原因の半数以上を占めているので，この判定精度の向上に重点的に取り組んでいく必要がある．

5.2.2　不要な補完の原因

訓練データで生じた 2 箇所での不要な補完のうち 1 箇所は，多品詞語の品詞解釈を誤ったため，実際には潜在節と競合する節を検出することができなかったことによるものであった．残りの 1 箇所は，慣用句と解釈すべき表現をそのように解釈できなかったものである．

試験データにおいて be 動詞補完規則の不備が原因で生じた 3 箇所での不要な補完のうち 1 箇所は，定形か過去分詞形かの曖昧性がある場合過去分詞形と解釈するという経験則に反する例であった．残りの 2 箇所は慣用句の解釈を誤ったものである．

5.3　規則の制御情報について

4.3 節で述べたように，be 動詞補完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以上で be 動詞が省略されている見出しは含まれていなかったが，試験データには次の見出し (H14) のように二箇所で be 動詞が省略されている見出しが 2 件含まれており，後方のキーに対して be 動詞を補うことができなかった[8]．

(H14)　Swissair January traffic up, revenue down

8 これら 2 件の見出しでは節境界がコンマによって示されているため，複数箇所での補完ができるように適用抑制規則集合を変更しても，条件 2 の節境界の認識が正しく行なえない．このため，表 5 では「条件 2 (節境界)」に含めた．

be 動詞補完規則にはすべて信頼度 B を与えているため，補完結果に対する構文解析が失敗すると，一度行なった補完が取り消されるが，今回の実験では，取り消しが生じた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．

5.4　be 動詞補完による翻訳品質の改善度

be 動詞を補うことによって実際にどの程度の品質改善が達成されたかを確認するために，試験データにおいて正しく be 動詞が補えた 67 件[9]の見出しについて，be 動詞補完前と補完後の翻訳を比較した．3.3 節の評価基準と同じ基準で評価した結果を表 6 に示す．表 6 によれば，67 件のうち 61 件について翻訳品質が改善されており，be 動詞補完による新聞記事見出し翻訳の品質改善効果が確認された．なお，4 件の品質低下の原因は実験システムの既存部分の不備であり，be 動詞の補完とは無関係である．

表 6　試験データでの翻訳品質の改善度

キー	改善	同等	改悪
過去分詞	32	2	2
to 不定詞	15	0	0
現在分詞	3	0	1
形容詞	8	0	1
前置詞句	2	0	0
複合動詞の構成素	1	0	0
合計	61	2	4

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生成する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品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規則を記述し，小規模ではあるが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，提案した方法の有効性が確認できた．

今後取り組むべき課題として次のような点が挙げられる．

(1)　be 動詞の省略現象に次いで頻繁に見られる見出し特有の現象はコンマが等位接続詞として用いられることであり，これが原因で適切な翻訳が得られないことも多い．また，単に be 動詞を補うだけでは翻訳品質の向上が不十分であり，コンマを等位接続詞に書き換える処理も同時に行なって初めて適切な翻訳が得られる見出しも存在する．

9 見出し (H14) のように二箇所への補完が必要な 2 件を 69 件から除く．

従って，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必要がある．

(2) 提案した方法では，記事本文から得られる手がかりを利用せずに書き換えを行なっている．しかし，より高い精度の書き換えを実現するためには，記事の本文特に第一文から得られる手がかりに基づく処理を行なうことが有効であると考えられる．例えば本稿では適切に行なえていない時制や相の決定に必要な情報が本文中に明示されている可能性は高い．

(3) 本稿では，処理対象の表現は新聞記事の見出しであることを前提として書き換えを行なっているが，提案した方法を実際の機械翻訳システムに組み込んで利用する場合には，処理対象表現が新聞記事の見出しであるかどうかを判定する処理を実現する必要がある．

謝辞

英々変換系の初期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，議論に参加頂いた英日機械翻訳グループの諸氏に感謝します．また，本稿の改善に非常に有益なコメントを頂いた査読者の方に感謝いたします．

参考文献

Hornby, A. S. (1977). 英語の型と語法. オックスフォード大学出版局. 伊藤健三 訳注.

金淵培 江原暉将 (1994). “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, **35** (6), 1018–1028.

Lewis, D. D. (1997). “Reuters-21578 Text Categorization Test Collection, Distribution 1.0.” http://www.research.att.com/~lewis/reuters21578.html.

長尾眞 辻井潤一 (1985). “機械翻訳における訳語選択と構造変換過程.” 情報処理, **26** (11), 1261–1270.

仲尾由雄 (1997). “見出しを利用した新聞・レポートからのダイジェスト情報の抽出.” 研究報告 NL117-17, 情報処理学会.

白井諭, 池原悟, 河岡司, 中村行宏 (1995). “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, **36** (1), 12–21.

白井諭, 大山芳史, 中尾嘉孝, 西垣万亀子, 上田洋美, 小見佳恵 (1997). “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 4B-1, 情報処理学会.

竹田正幸 松尾文碩 (1993). “英文科学技術抄録文における動詞の決定.” 情報処理学会論文誌, **34** (9), 1931–1936.

竹田正幸, 須田淳一郎, 楠本典孝, 松尾文碩 (1995). “英文科学技術抄録文における名詞の決定.” 情報処理学会論文誌, **36** (8), 1828–1837.

上野田守 布施敏夫 (1978). 新聞英語. 朝日実務英語シリーズ. 朝日出版社.

吉見毅彦, 奥西稔幸, 山路孝浩, 福持陽士 (1999). “表題へのつながりに基づく文の重要度評価.” 自然言語処理, **6** (1), 43–57.

略歴

吉見 毅彦: 1987 年電気通信大学大学院計算機科学専攻修士課程修了．1987 年よりシャープ (株) にて機械翻訳システムの研究開発に従事．1999 年神戸大学大学院自然科学研究科博士課程修了．

佐田 いち子: 1984 年北九州大学文学部英文学科卒業．同年シャープ (株) に入社．現在，同社情報システム事業本部ソフト事業推進センター係長．1985 年より機械翻訳システムの研究開発に従事．

(1999 年 6 月 3 日 受付)
(1999 年 10 月 8 日 再受付)
(2000 年 1 月 7 日 採録)