日本語記事の重要情報に基づく英文ヘッドライン生成法

畑山満美子　松尾義博　大山芳史　　白井諭
NTTコミュニケーション科学基礎研究所　ATR音声翻訳通信研究所

1．はじめに

我々は日本語新聞記事の速報型翻訳を行うための技術的課題を検討している[畑山 98]. 従来までは本文のみを翻訳対象としており，記事の内容を代表する見出しの翻訳については解析の困難さなどから対応されていなかった. 本稿では，記事対応のついている日英新聞記事の内容的差分の分析を行い，その分析から重要情報の抽出を行う要約ルールを作成し，それを利用して和文記事から英文ヘッドラインを生成する方法について論じる. (以下，ヘッドライン＝HLと略す)

和文記事から英文HLを生成する上でまず考えられるのは，和文見出しをそのまま翻訳することである. しかし，和文見出しは，1文から成ることは少なく，2～3文もしくはそれ以上の数文から成り立つ. また，接続詞や助詞がほとんど使用されず，名詞の羅列になる場合が多いため，機械的に解析することが困難で，構文情報がつかみにくいという特徴がある. このため，直接和文見出しから一文の英文HLを生成することは極めて困難だと考える. そこで本研究では，和文記事本文から英文HLを生成することを考える.

従来，HLを自動生成する研究はあまり知られていない. しかし，HLを生成することは，記事から最も重要な一文を選定する要約，または，重要要素を抽出し一文に構成する要約であるとも考えられる. このような観点から従来の要約研究を考えると，テキスト中の出現頻度によって単語に重み付けを行ない重要文を選定する手法[Edmundson69, Luhn58]，文間関係を利用した手法[Miike94,Marcu97]，心理実験を利用する手法[難波 98]などがあるが，これらはいずれも重要選定文をそのまま抜き出すことが考えられている. また，要約の評価として基準となるものが人間の直感によるところが大きく不安定である. 本手法では，評価基準として人間の直感の他，元となるHLとの整合性を基準にすることができ，これには実験者の主観が入らないという利点がある.

以上のことから，本研究では日英記事の内容的差分から要約ルール及びHL加工ルールを作成することによって，英文HLを自動生成するシステムを構築し，評価を行う.

2．日英新聞記事の比較検討

2.1.対象データ

対応づけ可能な日英記事として，日本経済新聞社の新聞記事に着目し，日経テレコンデータベースから，日本語記事はテレコンBIZ，英語記事はJapan News & Retrieval を対象データとした.

2.2.日英新聞記事の比較

記事対応のついている和文記事と英文記事(図1，2)の比較を行う. この例では，和文記事が5文で記述されているのに対し，英文記事は1文にまとめて記述されている. 例文中，下線部分は英文記事に採用されている内容であるが，複数文に渡り部分的に情報が抽出されているのが分かる. また例文中で，囲み部分は，英文HLに用いられた単語の情報源に相当する語句である.

このように，英文記事は概して和文記事より短く，重要情報のみをまとめて記事にしている傾向が見られた. また，英文HLは,本文よりも更に重要と思われる情報に絞り込まれているのが分かる.

【見出し】 JT株売り出し終了／申込倍率10倍強？
予想下回る500万件台

【和文記事】

1:日本たばこ産業(JT)株の一般売り出しの購入申し込みが、八日で締め切られた。
2:市場関係者によると、申込件数は売出株数の十倍強に当たる五百万―五百五十万件になった模様だ。
3:売出価格が百四十三万八千円(額面は五万円)と高かったうえ、六日上場した日本テレコムの株価が公募価格を割り込んでいることが影響、申込件数は事前の予想を下回った。
4:JT株の売出株数は四十三万六千六百六十六株。
5:購入申し込みの件数は、一千万件を上回るとの観測もあったが、日本電信電話(NTT)株(第一次放出百六十五万株、売出価格百十九万七千円)の千五十九万件、東日本旅客鉄道(JR東日本)株(百四十万株、同三十八万円)の千四十八万件に比べ半分程度の水準となった。

図1:和文記事

【ヘッドライン】
Japan Tobacco draws fewer than expected buyers

【英文記事】

1: Japan Tobacco shares drew 5.0-5.5 million applications, a little more than 10 times the actual number of shares to be offered, stock market sources estimated Thursday, the application deadline.

図2:英文記事

2.3.日英新聞記事の内容的差分を利用した要約

これらのことから，本研究では，和文記事と英文記事の内容的な差分を利用することによって，新聞記事の要約が行なえるのではないかと考え，このアプローチのもと，HLの自動生成の研究を行なっている.実際には，新聞記事にはHLと本文があるため，記事内容を数文にまとめた本文要約と，要約された内容を更に1文に集約し，英文HLの特徴をふまえた英文HLを生成する2つが考えられるが，本稿ではHL生成に言及する.

3．ヘッドライン翻訳システム

日英機械翻訳システム ALT-J/E を用いてHLを生成することを考えた場合，次の3つのパス(図3)が考えられる．

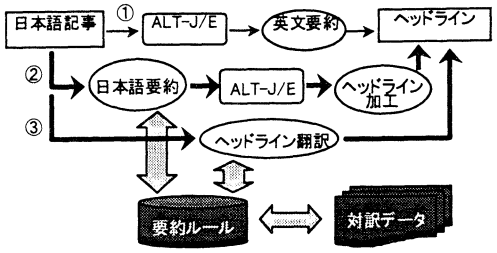

図3：ヘッドライン生成の流れ

パス1は，和文記事全体を英訳してから一文に要約し,HL加工を行なう方法である.この場合，精度の良い全文英訳を行うことが英文要約の精度に大きく影響するため，自動HL生成を考えると難しい問題がある．パス2は，和文記事を一文に和文要約してからALT-J/E を用いて英訳し,その結果にHL加工を行なう方法である.この場合，一度 和文記事から和文要約を生成するため，のちの本文要約にも利用できる枠組みを作成することが可能である．パス3は，ALT-J/E をベースにしたHL生成プログラムを作成する方法である．他にも日本語見出しを加工して翻訳する方法も考えられるが，本研究では，本文要約への展開を念頭に，本文に基づいてHLを生成することを考える．

以上の観点から，本研究ではパス2の手順で英文HLを生成する手段を考える．将来的にはパス3の専用ツールを作成した場合との比較検討を行いたい．

4．ヘッドラインルールの検討

比較検討に用いるデータとして，日英記事対応付け[高橋 97]を行った後，無作為抽出した151記事について分析を行った．

4.1.要約ルール

4.1.1.ヘッドライン生成の方針

HLを生成する際，重要となるのは，重要度付与による重要文の選定と，重要文からの必要な要素の抽出である．そこで，「HLに必要な情報を含んでいる和文」もしくは「第1文など，HL翻訳の元に使用したい和文」を特定し，英文の主語，動詞等，重要要素に絞り込むことを考えた．

ヘッドライン動詞の時制	(%)
現在形（SV／SVO）	40.8
to不定詞（S to V）	36.2
過去分詞	3.3
現在分詞	2.8
助動詞	1.2
過去形	1.0
その他（名詞句など）	14.8

図4：英文ヘッドラインの形態(3000 件中)

HLの形態の分析[白井 97](図4)によるとHLではSVO相当語があれば良いと考えられるため，HLに必要なSVO要素の抽出とそれに付随する不要な修飾語句の除去が必要となる．SVO抽出のためには形態素情報だけで必要な要素を選別するのは困難であり，係り受け解析を行い不要な修飾語句を判定し削除する．係り受け解析により主動詞を決定することによって，その格要素(SVO)を必要情報として抽出できるのではないかと期待される．

HLを構成する単語情報が含まれている和文をターゲット文と呼ぶことにする．記事全体からターゲット文を選出するために次のような方針を立てた．キーワードとキー表現，及び文の位置情報から，1文ごとに重要度を付与し，最も重要度の高い1文をターゲット文として採用する．

上記の方針で，情報抽出，HLを生成するための要約ルールを作成する．情報取得手段の検討のため，つぎの4点に重点を置きデータ分析を行った．

1. ターゲットとなる和文の選択
2. 和文主語と英文主語の対応
3. 和文動詞と英文動詞の対応
4. 目的語，補語の検討

次節以降にそれぞれに対する分析結果を示す．

4.1.2.ターゲット文の選択

ターゲット文をどのように選択するかを調べるために,英文HLと和文本文の対応[畑山 98]をとったところ，約7割のHLが和文1段落1文目から生成できることが期待できた．

- 和文1段落1文目だけで良いもの(72%)
- 1段落2文目以降も必要なもの(24%)
- 2段落目も必要なもの(2.5%)
- 和文見出しが必要なもの(1.3%)

2文目以降が必要な場合は，次のように分類される．

(a) 1文目の動詞が2文目に呼応するもの

例えば，「～を発表した。それによると～」「～計画を発表した。計画では～」のような文では，1文目に具体内容がなく，2文目に具体内容が記述されている．このパタンの場合，動詞とその目的語を手がかりにすることができる．

(b)1文目が導入文の場合

例えば，「選挙戦が始まった。A党のB氏は～」のように，和文記事の特徴的な書き方として，1文目に短い導入文，2文目に具体的・現実的な内容の文を書く,という形式が見られる．このように，1文目が短く具体内容がない場合，2文目以降も参照する必要がある．

(c) 本文中に強調のカギカッコが使用されているもの

第 1 段落の第2文目以降で「」内に名詞句がある場合，カッコ内の名詞句はHLの情報に使用されやすい傾向がある.本文中カッコをつけている名詞句は強調したい語句であり，伝えたい内容が凝縮されている場合が多いため，HLに使用されやすいと考えられる.しかし，これだけでは条件が少ないため，主動詞を条件にするなど他の判断材料が必要だと思われる．

これらの結果から，英文HL生成に必要な情報の多くは，和文記事の本文第1文目または第2文目から得られる見込

みであるため，本研究の第一ステップとして，和文第1段落第1文目からHLをつくることを目標とした．

以下の分析は，第1文目をターゲット文である記事（100記事）を対象とした分析結果である．

4.1.3.和文主語，英文主語の対応

ターゲット文の構文上の主語が，どの程度英文HLの主語と一致しているか，また，一致しないのはどのような場合かを分析する．一致の割合は以下のようである．

- 主語が一致する場合(74%)
- 部分的に一致する(7%)
- 一致しない場合(19%)

このうち，主語が一致するのは，(a)企業・官庁名(81%)，(b)利益・価格など(11.6%)，(c)人物名(7%)であった．部分的に一致するのは，主語の意味的補足，並列のカット，意味的にまとめる，などの場合であった．一致しないのは，主語が利益・収益などの場合である．

このように，(a)(c)が主語の場合はほぼ一致すると見て良いが，(b)が主語の場合，一致しない場合も見られるため，他の判断材料が必要となる．しかし，下の例のように，定式化できるパタンが大半である．

和文: *[企業]社の[期間]期経常利益は，*
[数]くらい[増加/減少]した．
英文: *Earnings: [企業] [増加/減少] [数]*
in pretax profits in [期間]

4.1.4.和文動詞，英文動詞の対応

同様に，ターゲット文の構文上の動詞が，どの程度英文HLの動詞と一致しているか，また，一致しないのはどのような場合かを分析する．

- 動詞が一致する場合(47%)
- 部分的に一致する(14%)
- 一致しない場合(40%)

一致しない場合は次のようなパタンがある．

(a)様相表現的な動詞(上記 40%のうちの 65%)

例えば，「Vすることに決めた。」「Vする方針だ。」「Vする(した)ことが明らかになった。」のような場合で，この時，英文HLに採用される動詞はVの部分である．

(b)様相表現的な動詞の二重使用(同 17%)

例えば，「V する見通しになったと発表した。」「V に乗り出すことで合意した。」などである．この場合，Vが英文動詞に相当するが，どのように主動詞を特定するかは問題である．

「部分的に一致する」場合は，(上記 14%のうちの)63%が「ヲ格+動詞」であった．例えば，

力を入れている → boost
契約を結んだ → tie up
検討を始めた → consider

などで，「ヲ格+和文動詞」の場合，ひとまとまりで英文動詞に訳出できれば良いと考えられる．

4.1.5.目的語，補語の選定基準

必須格を伴う英文動詞は 80%であった．分析の結果，英文HLに必要な情報は和文主動詞の必須格，不必要な情報は和文主動詞の必須格の修飾語・任意格であった．この仮定に基づいてルールを作成した．

4.2.ヘッドライン加工ルール

4.2.1.英文ヘッドラインの特徴

英文HLには次のような特徴がある[藤井 96]．

- 現在形を使う
  S will see → S to see
- be 動詞の省略
  S be to V → S to V
- 短い単語に置換される(略語化)
  The Ministry of International Trade and Industry
  → MITI
  S will approve → S to OK

4.2.2.英文動詞の時制の決定基準

HLの動詞の時制は，現在形(49%)，to 不定詞 (47.5%)に大別され，事実や出来事は現在形，予定や計画は to 不定詞で表現されることが分かった．稀に過去形，過去分詞形があるが，受け身の場合などであり，過去の事柄を過去形で表している例は見受けられなかった．

和文動詞との対応を見ると，和文動詞が現在形の場合，英文動詞の時制は「to 不定詞」．和文動詞が過去形の場合，英文動詞は「現在形」となる．様相表現の場合，主動詞によって変化する．

5．ヘッドライン翻訳の実験と結果

5.1.実験方法と入力データ

前節の分析結果を元に要約ルール，HL翻訳システムを作成した．このシステムに対し，前出のデータ100件を入力し，HLの元になる日本語要約文と英文HLを出力した．

5.2.実験結果

図6の和文を入力とした場合の生成結果を以下に示す(図7)．この例では，第1文がターゲット文として選定され，「社会保障制度審議会は公的介護保険制度の導入を提言した。」と要約されたのち，図7を得た．冠詞等の問題は今後の課題である．

社会保障制度審議会（首相の諮問機関、会長・隅谷三喜男東大名誉教授）は八日、社会保障の将来像についての報告書を発表、老齢者の介護サービスを保障する公的介護保険制度の導入を提言した。厚生省はこの報告を踏まえただちに本格的な検討に入るが、早ければ九七年度の導入をめざし（1）六十五歳以上を保険給付の対象とし（2）二十歳以上のすべての国民から、月収の一％弱相当の保険料を徴収する―などを考えている。年内にも具体案を提示するが、大幅な負担増に強い反発も予想され、実現までには曲折が予想される。

図6：入力和文

システムによるヘッドライン生成結果
Panel proposes the creation of a public nursing insurance
日経のヘッドライン
Panel proposes creation of public nursing insurance

図7：生成結果と正解

6．評価と考察

実験で得られた自動生成英文HLについて評価を行った．評価の対象は，要約文の生成と英文HLスタイル加工の2つの観点から行う．前者は，HL生成の途中過程で得られる和文要約について，どれだけ要約・情報抽出が出来ているかの観点で評価を行った．後者は，理想的な和文要約ができたと仮定したとき，どれだけ英文HLスタイルに適した翻訳がされたか，という観点で評価を行った．なお，どちらも日経HLを正解基準とした．

6.1.和文要約評価

(a)評価項目と基準

1. 正解HLの和訳と比較して，必要な情報(文節単位)がどれだけ抽出されているかを再現率と適合率で判定する．再現率は，(正解和文に含まれる要約後和文の文節数／正解和文の文節数)で表される．適合率は，(要約後和文に含まれる正解和文の文節数／要約後和文の文節数)で表される．
2. アナリストの直感による文の意味判定．和文要約のみを見て，記事の内容が分かるかどうかを評価する．評価基準は以下の通りである．
   ◎:意味的に正しい要約になっており，正解HLと語句的にも一致している．
   ○:意味的に正しい要約になっているが，語句が正解HLと一致しない．
   ×:要約になっていない．
3. 機械生成された英文HLについて2と同様の評価基準で文判定を行う．

(b)評価結果

以下のような評価結果を得た．

1.再現率の平均	69．0%
1.適合率の平均	84．5%

	◎	○	×
2.和文意味判定	33	24	43
3.英文意味判定	32	22	46

(%)

日経HLを正解要約とした場合，再現率・適合率の評価から，必要情報の大部分が正しく抽出できていると考えられる．日経HLを正解と設定しない場合，5～6割が意味的に要約として評価することができることが分かった．なお，英文HLに加工した場合にも，内容の劣化はあまりみられなかった．

6.2.英文ヘッドラインスタイル評価

理想的な和文要約が出来た場合，どれだけ英文HLスタイルに変換できるかを判定する．理想和文要約を人手で作成し，HLを生成した結果を評価する．

(a)評価項目と基準

1. 略語化されているかどうか
2. 正しい動詞かどうか．
   ◎:正解HLと一致している，
   ○:一致していないが意味は同じ，
   ×:一致していない．
3. 時制が整っているかどうか
4. 主語の一致(トピックがつかめているかの判定)．判定基準は2と同様．
5. 文の意味判定．判定基準は6.1節の2と同様．

(b)評価結果

以下のような評価結果を得た．ただし，略語化判定の場合，対象となる語が存在しない場合があるため，合計が100%にならない．

	◎	○	×
1.略語化	85	−	8
2.動詞	38	37	25
3.時制	58	−	42
4.主語	60	18	22
5.文の意味	73	0	27

(%)

HL加工ルールとしてみると，略語変換，動詞の選定は，7割以上が正しく訳されていることが分かる．同様に，主語が8割近く取れていること，意味判定が7割以上取れていることから，このルールを用いることによって，トピックを押さえたHL加工が可能であることが分かる．ただし，時制の変換については機械翻訳機ALT-J/Eの時制処理との不整合により，正解率が低下していることが分かったため，今後改良を行う．

7．おわりに

本稿では，対訳データから要約ルール，HL加工ルールを作成し，和文記事から英文HLを自動生成した．この結果，情報抽出として7割程度，要約として5～6割程度の結果を得ることができた．これからは，2文目以降をターゲットとした要約を考えていく予定である．

参考文献

[Edmundson69] H.P.Edmundson. New methods in automatic abtracting. Journal of ACM, Vol..16, No.2 (1996).

[Luhn58] H.P.Luhn. The automatic creation of literature abstracts.IBM Journal of Research and Development, Vol..2, No.2 (1958).

[Miike94] S.Miike, et al.. A full-text retrieval system with a dynamic abstract generation function. In Proc. of the 18th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval (1997).

[Marcu97] D.Marcu. From discourse structures to text summaries. In Proc. of the ACL Workshop on Intelligent Scalable Text Summarization (1997).

[難波 98] 難波英嗣, 奥村学. 観点に基づいた新聞記事の重要文抽出に関する心理実験と考察. 言語処理学会第4回年次大会(1998).

[藤井 96] 藤井章雄. ニュース英語の翻訳プロセス, 早稲田大学出版部(1996).

[白井 97] 白井諭, 他. 英文記事ヘッドラインの特徴について. 情報処理学会第54回全国大会(1997).

[高橋 97] 高橋大和, 他. 日英新聞記事の記事対応コーパス自動作成, 言語処理学会第3回年次大会(1997).

[畑山 98] 畑山満美子, 他. 日本文新聞記事からの英文ヘッドライン生成法について, 情報処理学会第 57 回全国大会 (1998).