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

5R－2

畑山満美子　白井諭　大山芳史　瀬下貴加子　清末三恵子
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

新聞記事は見出し，リード，本文から構成され，機械翻訳ではリードや本文が対象とされてきた．新聞記事全体の翻訳を行うには見出しの翻訳も必要となるが，日本文記事の見出しは文の体裁を持っていないため，そのままでは機械翻訳の対象にできないという問題がある．また，日英記事を比較すると，本文も文対文には単純に対応していないため，そのまま翻訳しただけでは英字新聞に対応できない．

以上から日本文新聞記事の速報型翻訳を行うための技術的課題の検討を開始した．新聞記事翻訳の第1ステップとして，記事対応のついている日本文新聞記事と英文記事の間の内容的差分の分析を進めている．本研究ではこの対訳データの分析から重要情報の抽出を行う要約ルールを作成し，それを利用した本文要約，ヘッドライン翻訳を行うことを考える．これらのうち，本稿では，ヘッドライン翻訳について論じる．予備実験によれば，比較的単純な方法で7割程度のヘッドラインが生成される見込みである．

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．このうち，比較検討に用いるデータとして，日英記事対応付け[1]を行った後，無作為抽出した80記事について分析を行った．

2.2. 日英新聞記事の比較

記事対応のついている日本文記事と英文記事(図1，2)の比較を行う．この例では，日本文記事が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．このように，英文記事は概して日本文記事より短く，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．また例文中で，囲み部分は，英文ヘッドラインに用いられた単語の情報源に相当する語句である．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
【日本文記事】
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：日本文記事

【英文ヘッドライン】
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(図1の対訳)

ヘッドラインの生成としては日本語の見出しを加工して翻訳する方法も考えられるが，本文要約への展開を念頭に，本研究では本文に基づいてヘッドラインを生成することを考える．

3. ヘッドライン翻訳について

3.1. ヘッドライン翻訳の実現

ヘッドラインを生成するには，3つのパス(図3)が考えられる．本稿では方式検討を優先するため，当面②を採用するが，将来的には③の専用ツールを作成する方針である．

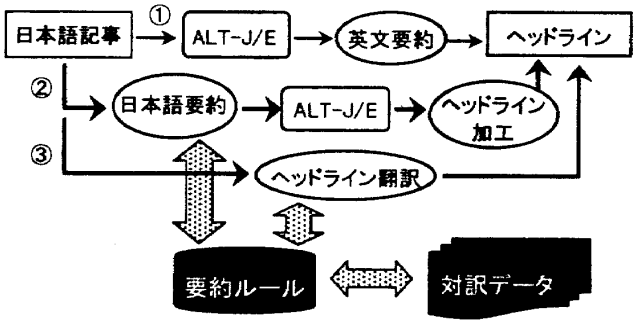

図3：ヘッドライン生成の流れ

Generating English Headlines from Japanese Articles
* Mamiko HATAYAMA, Satoshi SHIRAI, Yoshifumi OOYAMA, NTT Communication Science Laboratories.
** Takako SESHIMO, Mieko KIYOSUE, NTT Advanced Technology Corporation.

3.2. ヘッドライン生成の方針

ヘッドラインを生成する際，重要となるのは，重要度付与による重要文の選定，重要文からの必要な要素の抽出，不必要な要素（修飾語句）の除去である．そこで，「ヘッドラインに必要な情報を含んでいる和文」もしくは「第1文など，ヘッドライン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る．ヘッドラインではSVO相当語があれば良いと考えられるので，ターゲットとなる文からのSVO要素の抽出と不要な修飾語句の除去が課題となる．

また，英文ヘッドラインには次のような特徴が見られる[2]．1）現在形を使う，2）be動詞の省略，3）短い単語に置換される．これらに対応することも必要である．

3.3. 分析結果

情報取得手段として，以下の5点を重点に分析を行った．

(1) ターゲット文の選択

ヘッドライン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する．ターゲット文は，和文1段落1文目だけで良いもの(73.4%)，2文目以降も必要なもの(26.6%)に大別される．2文目以降が必要な場合は，本文中に強調の「」が使用されているもの，1文目が導入文で具体的内容が2文目以降に書かれているもの（例:「～公表した．それによると～」），1文目の動詞が2文目に呼応するもの（例:「計画を発表した．計画では～」）がある．この結果から，英文ヘッドライン生成に必要な情報の多くは，日本文記事の本文第1文目または第2文目から得られる見込みである．以下は，第1文をターゲットにできるもの(59 記事)の分析である．

(2) 和文主語，英文主語の対応

ターゲット文の構文上の主語・動詞が，どの程度英文の主語・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．

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%)がある．一致するのは主語が人物・企業・官庁名の場合．部分的に一致するのは，主語の意味的補足・並列のカット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．

(3) 和文動詞，英文動詞の対応

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，（40%のうちの）65%が様相表現的な動詞（例:「～することに決めた」「～する方針だ」「～ことが明らかになった」），17%が様相表現的な動詞の二重使用であった（例:「～する見通しになったと発表した」）．

(4) 英文動詞の時制の決定基準

英文動詞の時制は，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．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．様相表現では，文面で決定できるものと，直前の動詞の時制により決定できるものに分けられる．

(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，英文ヘッドライン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．

3.4. ヘッドライン生成結果

以上の分析結果に基づいてヘッドライン翻訳システムを試作した．これによる生成結果を以下に示す（図4, 5）．

(ニューヨーク2日＝近藤勝義)特定の国や地域の株式などに集中投資する会社型投資信託(カントリーファンド)を、日本の大手証券会社が相次いでニューヨーク証券取引所に上場する。野村証券は中東諸国の株式を組み込む「エマージング・ミドルイースト・ファンド」の主幹事となるのをはじめ、年内に四本のファンド上場を予定。大和証券、山一証券も日本やアジア株に投資するファンドの主幹事案件を抱えており、この分野で先行する米大手証券会社との引き受け競争が激しさを増している。

図4:入力文

システムによるヘッドライン生成結果
A brokerage in Japan to list country funds in NYSE
日経のヘッドライン
Major brokerages to list country funds on NYSE

図5:ヘッドライン生成結果

この例では，第1文がターゲット文として選定され，「国や会社型投資信託を，日本の大手証券会社がニューヨーク証券取引所に上場する。」と要約されたのち，図5を得た．主語の単複などの問題は今後の課題である．

4. おわりに

本稿では，対訳データの比較検討から得た要約ルールによりヘッドライン翻訳を行った．実験結果では，7割弱がヘッドライン風に翻訳されたが，評価についてはこれからの検討となる．システムは試作の段階であり，今後は検討データを増やし3．3節の分析を更に深めてゆくことが課題となる．また，この要約ルールを用いて，ヘッドライン生成だけではなく，本文要約にも取り組んでゆく予定である．

参考文献

[1] 高橋大和，白井諭，大山芳史，渡辺いづみ，上田洋美．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)．
[2] 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)．