英語ニュース記事ヘッドラインの翻訳

〒212-8582 川崎市幸区小向東芝町１
（株）東芝　研究開発センター　ヒューマンインターフェースラボラトリー
小野 顕司
kenji2.ono@toshiba.co.jp

Abstract

英語ニュース記事は未知語が多い、文体が特殊であるといった理由により機械翻訳が困難であるという問題がある。特にそのヘッドラインは断片的に記述されており、翻訳が困難である。今回英語ニュース記事ヘッドライン翻訳システムを試作して評価を行った。方式のポイントは、類似記事の翻訳文や記事本文を利用して翻訳を行う点である。具体的には英日対訳コーパスから類似記事を見つけ、訳語情報を抽出して利用する。またヘッドライン中の断片的な単語と記事本文中の語句とを対応付け、ヘッドラインを補完して翻訳する。３万件の英日対訳記事を用いて実験を行った結果、１００ヘッドラインについて従来と比べて訳語改善率１７％、文体の改善率２１％という結果を得た。

1　はじめに

近年インターネット上の英語ウェブページ閲覧のために機械翻訳ソフトが利用されることが増えてきた。一般的なユーザがよく見るウェブページとしてワシントンポスト紙、ウォールストリートジャーナル紙といった海外の動向をリアルタイムに伝えてくれるオンラインニュースのページがある。これらの英語ニュース記事は、翻訳辞書に登録されていない新しい固有名詞が多い、あるいは文体が特殊であるといった理由により機械翻訳が困難であるという問題がある。特にそのヘッドライン（記事タイトル）は、英語圏の読者の背景知識を前提として断片的に記述されており、翻訳が困難である。しかし、翻訳ソフトを使ってこれらのページを読むユーザはヘッドラインの訳文を見て記事本文を読むかどうか判断する可能性が高く、その意味ではヘッドライン部分の翻訳は本文部分の翻訳よりも重要性が高いと考えられる。

英日ニュースヘッドライン翻訳の既存研究としては、(吉見 99), (大井 96), (加藤 93) などがある。吉見および加藤のものはヘッドラインの文体に対応した翻訳処理やルールを設けるものである。大井らのものはヘッドラインの翻訳に記事本文の単語を利用するものであり、シソーラスを併用してヘッドライン中の単語の多義性解消を行う。これは訳語選択に有効であるが、もっと直接的に本文をヘッドラインの翻訳に利用することも考えられる。本研究では本文特に先頭文の情報の直接的な利用と、既存対訳コーパスの利用に着目してヘッドライン翻訳方式を提案する。

2　英語ヘッドライン翻訳方式

この翻訳方式の特長は、翻訳対象の英語新聞記事と類似する記事を対訳コーパスから見つけ、訳語情報を抽出して利用する点である。また、記事先頭文から抽出した情報を利用してヘッドラインの翻訳に利用する点である。以下に本方式の処理の流れを示す。また、図１に本方式に基づく翻訳システムの構成図を示す。

(1) 原文解析 英語記事を解析して、ヘッドライン部分と本文部分を区別する。また単語の辞書見出しを取得する。

(2) 英日対訳記事コーパスからの類似記事検索 英語記事に出現する英単語の頻度ベクトルと、コーパス中の英語記事の単語頻度ベクトル間の距離（cos）を計算し、しきい値以上で上位１０位まで抽出する。イベントとして類似していれば有効な訳語情報は取り出せるので、固有名詞を除いて単語インデクスを作成している。コーパス中の記事数（新聞記事）は３万件。訳文は人手で作成されたものである。内訳はクリーン２万件[1]＋ノイジー１万件[2]である。

(3) 類似記事対からの訳語情報抽出 検索された類似英語記事とその翻訳から、訳語情報を抽出する。

(4) フレーズアラインメント ヘッドラインと記事先頭文の間でフレーズアラインメントを行い、ヘッドライン中の名詞句を補完する。

[1] '95～'96 のロイター記事とその人手翻訳文。
[2] インターネットで収集した英語記事とその抄訳記事

(5) 英語記事の翻訳 (3) で抽出された訳語情報を利用して翻訳する。ヘッドラインの翻訳には、(4) のフレーズアラインメント結果を利用する。またヘッドライン翻訳のための特別の翻訳ルール（体言止め, etc.）を適用する。

以下、各処理について順に述べる。

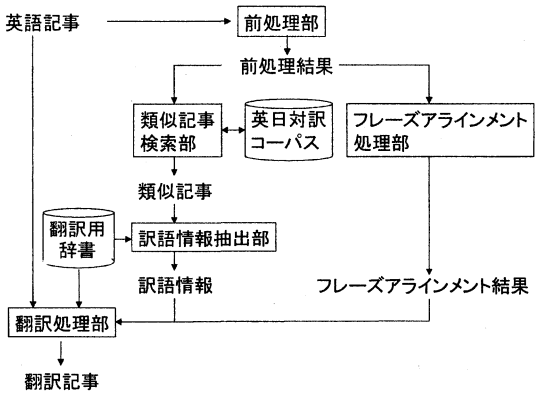

図 1: システム構成図

2.1 対訳コーパスからの訳語情報抽出

対訳記事から、英単語がどの日本語単語に訳されているかを検出する。これは、単語アラインメントとみなすこともできる。具体的には、英日翻訳辞書を参照して、英語記事に出現する単語の訳語候補のうち日本語記事に出現しているものを抽出する。あるいは日英翻訳辞書を参照して、日本語記事に出現する単語の訳語候補のうち英語記事に出現しているものを抽出する。こうして抽出した単語対の中で、頻度が高いものを出力する。

単語アラインメントの方法は各種ある。本方式のメリットは、コンパラブル記事対（訳文が原文にあまり対応しないような翻訳）に対しても適用可能という点にある。インターネット上にはクリーンパラレルデータ（訳文が原文を正確に人手で翻訳したもの）は少ないが、コンパラブルデータは多い。特に新聞記事はそうで、日本語記事は原文の半分以下の要約になっている場合が多い。

2.2 記事先頭文とのフレーズアラインメントによるヘッドライン補完

英語新聞記事のヘッドラインと記事先頭文との間でフレーズアラインメントを行い、ヘッドラインの中の単語（断片的で翻訳困難な場合が多い）を補完する。処理の流れは以下の通りである。

1. 各文を解析し、品詞列を参照して名詞節（フレーズ）を取り出す。
2. 両フレーズを比較し、一致率を計算する。略字なども考慮する。
3. 一致率が最大かつ値が閾値（現在暫定的に５０％に措定）以上の候補を出力する。

2.3 ヘッドライン翻訳用ルール

新聞のヘッドラインの文体は、通常体言や格助詞等で終ることが多い。そこで、体言止めになるように訳出する訳出ルールを作成し、ヘッドライン部分にだけそのルールを適用するようにしている。この他、to 不定詞や分詞構文, “seen” や “may” といったヘッドライン特有の文体に応じた訳出ルールも追加している。

また、英語ニュースヘッドライン特有の文体として、ニュースソースを文末に付加するものがある。これは通常に翻訳すると問題をおこすので、フレーズアラインメント処理で記事先頭文を参照するときに、ヘッドライン末尾の単語がニュースソースであるかどうかを判定し、ニュースソースであると判定した場合には、その部分を分離してそれぞれ翻訳し、あとでマージして訳文を得るようにしている。

3 ヘッドライン翻訳の動作例

英語ヘッドライン翻訳実験システムの動作例を示す。

図 2 は翻訳対象の英語記事の例である。図 3 は図 2 に示した英語記事のヘッドラインと記事先頭文のフレーズアラインメント結果である。図 4 は類似記事検索部によって英日対訳コーパスから検索された類似記事の一覧と類似度である。図 5 は検索された類似記事（英日対訳）の例である。図 6 は訳語情報抽出部で類似記事から抽出された訳語情報の一部である。「●」が付与されているのは、図 2 の英語記事の翻訳に適用されたものである。図 7 に本実験システムによる図 2 の英語記事の翻訳結果を示す。従来の翻訳結果を（従）、本実験システムによる翻訳結果を（本）で示している。また訳文中の相違個所を【...】によって示している。

4 評価

100 ヘッドラインに対して本ヘッドライン実験システムを適用し、翻訳精度の変化を評価した。結果は以下のとおりである。

- 訳語改善は、１７％（改善２６、悪化９）
- 文体の改善（体言止めなど）：２１％（改善２８、悪化７）

タイトル：Disney to buy back up to 104.5 mln shares
本文：BURBANK, Calif., April 23 (Reuter) - Walt Disney Co said its board had approved a stock repurchase program of up to 104.5 million shares. The program replaces a similar program that was in place prior to its acquisition of Capital Cities/ABC, it said on Monday.

図 2: 入力英語記事例

Disney → Walt Disney Co

図 3: フレーズアラインメント結果

類似度	ヘッドライン
0.58	Northwest to buy back up to 5 mln shares
0.57	Cisco increases buyback program
0.53	Dell Computer increases share buyback
0.51	Microtest Inc bought back 164,500 shares
0.46	PaineWebber increases share buyback plan
0.46	Gillette sets 10-15 mln share buyback
0.44	Campbell heir continues share sale
0.43	Texaco has bought 1.5 mln shrs
0.43	ADM to buy back up to 20 mln of its shares

図 4: 類似記事検索結果

タイトル：Northwest to buy back up to 5 mln shares
本文：MINNEAPOLIS, Dec 6 (Reuter) - Northwest Airlines Corp said Friday its board had approved a program to buy back up to five million shares of Class A common stock. (以下略)

タイトル：ノースウェスト航空（米）、役員会が普通株500万株の買い戻しを承認
本文：[ミネアポリス　6日　ロイター]　米ノースウェスト航空は、同航空の役員会が、クラスA普通株を最大５００万株買い戻す計画を承認した。(以下略)

図 5: 対訳コーパス中の類似記事

適合率（訳が変化したもののうち、訳が改善したものの割合）は 75 % 程度だが、改善でも悪化でもない訳語変化が多かったので、それを考慮して非悪化分全体を対象と考えれば適合率は９０%となる。記事本文も訳語精度が向上しているが、計測していない。

改善例を３つ以下に示す。英語原文、人手翻訳例、従来翻訳システムによる訳文、本ヘッドライン実験システムにおける訳文の順で表示してある。訳文の相違点は【】でマークアップしてある。

approve(v) →	承認＜両性名詞＞
board(n)	取締役会＜名詞＞●
buy(v) →	買い戻す＜５段動詞＞●
common stock(n) →	普通株＜名＞
dilute(v) →	希薄＜形容詞＞
employee(n) →	従業員＜名＞
exercise(n) →	行使＜両性名詞＞
offset(v) →	相殺＜両性名詞＞
program(n)	計画＜両性名詞＞●
repurchase(n) →	買い戻し＜両性名詞＞, 買い戻す＜５段動詞＞
repurchase(v) →	買い戻し＜両性名詞＞, 買い戻す＜５段動詞＞
say(v) →	述べる＜下１段動詞＞●
stock(n) →	株＜名詞＞●
stock option(n) →	株式オプション＜名＞

図 6: 訳語情報抽出結果

(従)ディズニーは、１億 450 万株までバックアップを買【う。】
(本)【ウォルト・】ディズニー【社】は、１億 450 万株までバックアップを買【い戻す】

(従)バーバンク (カリフォルニア)、４月 23 日 (ロイター)—ウォルト・ディズニー社は、その【ボード】が１億 450 万株以内の【ストック】買い戻し【プログラム】を承認したと【言っ】た。
(本)バーバンク (カリフォルニア【州】)、４月 23 日 (ロイター)—ウォルト・ディズニー社は、その【取締役会】が１億 450 万株以内の【株】買い戻し【計画】を承認したと【述べ】た。

(従)【プログラム】がキャピタル・シティーズ/ABC のその獲得に先立って適所にあった、類似した【プログラム】を交換する、とそれは月曜日に【言っ】た。
(本)【計画】がキャピタル・シティーズ/ABC のその獲得に先立って適所にあった、類似した【計画】を交換する、とそれは月曜日に【述べ】た。

図 7: 翻訳結果

(原)Burmese in heavy attack on Karen border base
(人)ミャンマー政府軍、タイ国境のカレン民族同盟の拠点に激しい攻撃
(従)カレン【境界】【基礎】に対する激しい攻撃【においてビルマである。】
(本)カレン【国境】【拠点】に対する激しい攻撃【でのビルマの政府軍およびカレン反逆者】

この例で、“border”:「境界」→「国境」,“base”:「基礎」→「拠点」は対訳コーパスからの訳語情報抽出処理に、またまた、“Burmese in” の訳語の変化は、記事先頭文とのフレーズアラインメントによるヘッドライン補完によっている。

(原)S.Korea to raise ceiling on banks' CD issuance
(人)韓国、銀行のＣＤ発行枠拡大へ＝韓国銀行
(従)韓国は銀行の【CD】配給で【天井】を調達【する。】
(本)韓国、銀行の【3 か月の譲渡可能定期預金証書の】配給で【上限】を調達

この例で、“CD” :「CD」→「3か月の譲渡可能定期預金証書」は記事先頭文とのフレーズアラインメントによるヘッドライン補完に、“ceiling” :「天井」→「上限」は対訳コーパスからの訳語情報抽出処理に、また体言止めになっているのは、節2.3で説明したヘッドライン翻訳用ルールによる。この例では訳文は改善しているが、“raise” の訳語が改善していないため意味がまだとりにくい。

(原)HKMA nearing full control of HK banking-analysts
(人)香港金融管理局、銀行制度の全面的管理へ近づく=アナリスト
(従)【HKMA】は、HK【銀行業務-アナリスト-の十分なコントロール】に近づいている【。】
(本)【香港金融局】は、HK【金融の十分な規制】に近づいている【--アナリスト】

この例では、フレーズアラインメント処理で“HKMA”が本文中の “Hong Kong Monetary Authority” の略字であることを判定している。また、文末の “analyst” がニュースソースであることを判定し、分離して翻訳している。

つぎに改悪例を以下に示す。

(原)Chechen war shifts from Grozny, no end in sight
(人)チェチェンでの戦闘は首都を離れたが、終戦の見通しはない
(従)チェチェン【戦争】はグロズヌイ(光景中の【終了】なし)から変わる【。】
(本)チェチェン【軍事】はグロズヌイ【の廃墟】(光景中の【最後】なし)から変わる

この例では “war” の訳語が「戦争」から「軍事」となり、悪化している。“end” :「終了」→「最後」の訳語の変化は、改善でもないが悪化でもない。「廃虚」は記事先頭文とのフレーズアラインメントによるヘッドライン補完によって挿入されたものだが、これも訳文の改善には寄与していない。

(原)China fishing boats cut undersea cable to Japan
(人)中国の漁船が日本への海底ケーブルを切断
(従)中国漁船は、日本への海中のケーブルを切断する。
(本)日本への中国の漁船カット非常に現代の海中のテレコミュニケーション・ケーブル

この例では、“China fishing boats” が先頭文中の “Chinese fishing boats” に[3]、フレーズアラインメントされているが、このため “cut” の訳語が「切断」から「カット」に劣化している[4]。

また、“undersea cable” が “super-modern undersea telecommunications cable” (訳文は「非常に/現代の/海中の/テレコミュニケーション/・/ケーブル」) にフレーズアラインメントされている。“super-modern” の訳語が正確でないため、わかりずらい文字列をヘッドライン訳文にもちこむ結果となっている。

これらの悪化例を分析して問題点を整理した。

- 類似記事が見つかる割合は5割程度と低い。また類似記事が見つかっても実質的に効果のある訳語情報がとれない場合もある。
- 除去すべき訳語情報がある。現在の訳語情報抽出はプログラムで自動的に行っているが、ノイズがある。人手で修正する必要がある。
- 改悪事例はフレーズアラインメント処理に基づくものが目立つ。これらはユーザ辞書登録の際に語句の単数・複数情報を保存するようにフレーズアラインメント処理のインプリメントを改善することと、フレーズアラインメントを行う対象を企業名や人名などの Named Entity に限定することにより、かなり対処できると思われる。

5 おわりに

英語ニュースヘッドライン翻訳実験システムを試作した。方式のポイントは、類似記事の翻訳文や記事本文を利用してヘッドラインの翻訳を行う点である。具体的には、3万件の英日対訳コーパスから類似記事を見つけ、訳語情報を抽出して利用する。またヘッドライン中の断片的な単語と記事本文(先頭文)中の語句とを対応付け、ヘッドライン中の語句を補完して翻訳する。

１００ヘッドラインに対する翻訳結果を従来の翻訳結果と比較して評価を行い、訳語改善：１７％(改善２６、悪化９), 文体の改善(体言止めなど)：２１％(改善２８、悪化７)という結果を得た。適合率(訳文が変化したもののうち、翻訳が悪化しなかった割合)は９０％である。

References

Takehiko Yoshimi, Ihciko Sata; Improvement of Translation Quality of English Newspaper Headlines by Automatic Preediting, MT Summit VII, pp. 496-500, 1999.

大井一郎、角田達彦：英字新聞の本文の語彙的結束性による見出し中の名詞の多義性解消, 信学技報 NLC96-2, 1996.

加藤直人、相沢輝昭：外電ニュースの定型文抽出とその英日機械翻訳 NL 研, 1993.

[3]原文ないし訳文がおなじになるようなフレーズアラインメントは棄却するようにしているが、この例では訳文が「中国漁船」と「中国の漁船」となり、棄却されなかった。

[4]現在の実験システムのインプリメントではフレーズアラインメント結果をユーザ辞書に単数単語として登録しているため