話し言葉機械翻訳のための日本語前編集

坂本　明子　　田中　浩之
株式会社東芝 研究開発センター
{akiko7.sakamoto, hiroyuki34.tanaka}@toshiba.co.jp

1　はじめに

本研究は，講演の話し言葉を理解しやすい簡潔な表現に変換し，これによって機械翻訳の精度を改善することを目的とする．

後段の処理の精度改善を目的とし，統計的な手法を用いて話し言葉を前編集する技術 ([10]，[11]，[12]) が研究されている．これらの前編集の内容は，フィラーや間投詞の除去，口語表現の正規化，言い直し部分の削除，助詞の補完，句点の挿入や節境界検出のように多岐にわたる．統計的手法を用いた開発のためには，音声の収録や，フィラーも含めた，発音に忠実な書き起こし，各種のアノテーションタグを付与したコーパスを大規模に構築する必要がある．しかし，必要な現象を網羅し，さらに正解の前編集結果を付与したコーパスを，大規模に構築するのは難しい．

一方、自然言語処理技術において規則ベースの手法を用いると，比較的少量のコーパスであっても，これを参照して開発を進めることができ，初期段階でも大きな効果を期待することができる．また，適用する処理も制御しやすく，後段の処理に合わせた柔軟な制御が可能である．規則ベースで話し言葉を前編集する先行研究としては，動詞・サ変名詞・形容詞・副詞についての換言辞書を人手で作成して換言に用いる技術 [14]，名詞を国語辞典を用いて換言する技術 [9] が報告されている．また，節境界検出のための規則も既に体系的にまとめられている [7]．

本研究では，規則ベースの方法が有効に働くことが期待できる，機能語相当表現，すなわち，助詞相当表現や助動詞相当表現の前編集規則を重点的に作成する．これらの表現は，一般的に新語が出現しにくく，話者に共通する現象であるため，機械翻訳の翻訳品質改善への貢献度が高い．

2　前編集知識の作成

2.1　前編集規則のタイプ分類

話し言葉の前編集規則を作成する際に着目した，言語現象の例を表 1 に挙げる．

便宜的に，規則を A，B，C と分類する．A はフィラーや間投詞の削除，口語の正規化に関する規則である．これらは，先行研究で既に多く取り上げられている現象である．また，機械翻訳には難易度の高いと思われる細かなニュアンスを担うテ形複合動詞 [8] の一部を削除する規則も含める[1]．

B は，今回重点的に作成した助詞・助動詞相当の表現に関する規則である．規則 B を作成する際には，敬語の誤用を指摘する書籍 [3] や，作文を推敲する際の指南書 [1] を一部参考にし，これらを機械翻訳に適した簡潔な表現に変換することを目的とする．

C は，節境界検出に関する規則である．機械翻訳をする上で，その処理単位を決める必要があるが，話し言葉は，文の区切りを示す句読点を含まない．そこで，本研究では，節を翻訳処理の基本単位とし，これを検出する規則を設けた．

2.2　前編集の動作と，規則の書式

本研究の前編集は，形態素系列における系列パターンに対して，これを別の系列に書き換える動作として定義する．形態素解析には，MeCab[5] と日本語形態素辞書 UniDic2.0[2] を用いた．

前編集規則は，編集対象の形態素系列パターンと，編集後の形態素系列パターンの組からなる．形態素系列のパターンは，形態素の表層に加えて，MeCab+UniDic の解析出力である品詞，活用型，活用形，原型と照合できるようにし，これらの要素のうち必須と定義されたものが一致した系列を検知する．たとえば，「“”, “感動詞-フィラー”, “”, “”, “”」というパターンは，表

[1] 今回は，「てくる」，「ていく」，「てあげる」，「てくれる」，「てみる」とその活用形を削除した．

表 1: 話し言葉の前編集の例

分類	項目	手がかり	前編集規則による言い換え前後の例文
A	フィラー	「あのー」	前: 5) /**あのー**/ 今回初めての方もいらっしゃるので 後: 6) 今回初めての方もいらっしゃるので
	間投詞	「なんか」	前: 7) /**なん/か**/ 小型の端末を持ってる人もいて 後: 8) 小型の端末を持ってる人もいて
	敬語	「いたす/致す」	前: 9) 今回の取り組みと /**いたし**/ ましては 後: 10) 今回の取り組みと /し/ ましては
	助詞相当表現の敬体	「ます」	前: 11) 今回の取り組み /**と/し/まし/て**/ は 後: 12) 今回の取り組み /と/し/て/ は
	口語	「てん」	前: 13) 何し /**てん**/ のかよく分かった 後: 14) 何し /て/いる/ のかよく分かった
	テ形複合動詞	「てくる」	前: 15) 実際にお客さんの声を /かき集め/**て/くる**/ ということをします 後: 16) 実際にお客さんの声を /かき集める/ ということをします
		「てあげる」	前: 17) リアルタイムに処理 /し/**て/あげる** 後: 18) リアルタイムに処理 /する/
B	助詞相当表現	「ふう/風」	前: 19) /と/いう/**ふう**/な/ ことを可能にします 後: 20) /と/いう/ ことを可能にします
		「みたい」	前: 21) 高まってるんじゃないか /**みたい/な**/ ことを 後: 22) 高まってるんじゃないか /と/いう/ ことを
	助動詞相当表現	「こと」	前: 23) 実際にお客さんの声を /かき集める/と/いう/**こと**/を/し/ます/ 後: 24) 実際にお客さんの声を /かき集め/ます/
		「ところ」	前: 25) 食事介助業務 /と/いう/**ところ**/です/ね/ 後: 26) 食事介助業務 /です/
		「かたち/形」	前: 27) 実際には後から入力 /する/**形**/に/なり/ます/ 後: 28) 実際には後から入力 /し/ます/
C	接続助詞相当表現	「けれども」	前: 29) 本日のアジェンダです /**けれど/も**/ 四件ご講演を予定しています 後: 30) 本日のアジェンダです【境界】四件ご講演を予定しています
		「ので」	前: 31) お時間になりました /**の/で**/ 講演会を始めます 後: 32) お時間になりました【境界】講演会を始めます

層や活用型等によらず，品詞が “感動詞-フィラー” であるものとマッチする．

前編集の動作は，「削除」，「変換」，「節境界検出」の３つに分類できる．表 1 の例文 (5) から (32) に，前編集の例を示す．表 1 の「項目」は，例文中の下線が示す表現の文法的な分類を表している．表 1 の「手がかり」に示した語は，下線部の表現をコーパス中から探す際に用いたキーワードを示しており，例文中では太字部分に相当する．なお，下線部については，分かり易さのため形態素を斜線で区切って表示している．分類 C の規則の例文には，同規則で検出する節境界を “【境界】” として表している．それぞれについて，規則の例を示す．

§ 削除規則

「“”, “感動詞-フィラー”, “”, “”, “”」→ null

形態素の消滅は，便宜上 “null” として表現している．表 1 の例文 (5) は，この規則を適用した結果である．

§ 変換規則

「“”, “動詞”, “”, “連用形一般”, “”」「“て”, “助詞-接続助詞”, “”, “”, “て”」「“くる”, “動詞-非自立可能”, “カ行変格”, “終止形-一般”, “くる”」→ 「“”, “動詞”, “”, “終止一般”, “”」

例文 (15) は，この規則を適用した結果であり，「かき集めてくる」を「かき集める」に変換している．変換規則において，活用形を変更するとともに，表層も改めている．

§ 節境界検出規則

「“けれど”, “助詞-接続助詞”, “”, “”, “けれど”」「“も”, “助詞-係助詞”, “”, “”, “も”」→ 「“けれど”, “助詞-接続助詞”, “”, “”, “けれど”」「“も”, “助詞-係助詞”, “”, “”, “も”」「【境界】」

例文 (29) は，この規則を適用した結果であり，「けれど」と「も」の並びを条件と指定し，節【境界】を検出している．

3 実験

3.1 コーパス

規則作成用，および，closed 評価用には，東芝社内で行われた情報分野の講演（５名分）を録音し，人手で書き起こしたコーパスを用いた．書き起こしの仕様

表 2: 実験用コーパス

コーパス用途	文字数	文数	1 文の平均文字数
規則作成	40,278	522	77.1
Closed 評価	4,087	64	63.9
Open 評価	4,037	68	59.4

表 3: 前編集規則 (721 規則) の内訳

分類	項目	規則数
A	フィラー・間投詞の除去, 敬語・口語の正規化	310
B	口語的な言い回しの簡素化	315
C	翻訳単位検出	96

は，日本語話し言葉コーパスの転記テキストの仕様 [4] を参考にした．

規則開発や評価には，言い直しや言い淀みといった現象は除き，表 2 に示した分量だけを取り出して利用した．言い直しや言い淀みを除いたのは，これらがパターンマッチングで検出しづらく，バリエーションも多いことから，規則ベースの手法である本研究の対象外としたこと，また，これらを含むことで，後の翻訳結果の評価が難しくなることを避けるためである．尚，表 2 に示す文数は，１秒のポーズで機械的に分割し，さらに，「です」,「ます」,「ました」といった明確な文末表現の末尾で人手分割した数である．

open 評価用には，日本語話し言葉コーパス [6] の Disk6 から無作為に 5 話者[2]の転記テキスト (書き起こし) を選び，話者毎に講演の冒頭を少し過ぎたあたりと中盤以降の文とを合わせて，無作為に約 800 文字ずつ抽出した．

3.2 前編集規則の作成

表 2 に示した規則作成用のコーパスを参照し，721 規則を作成した．表 3 に作成した規則の内訳を示す．

3.3 前編集精度の評価

評価用コーパスに対して，日本語母語話者が前編集すべき箇所を付与した．評価は，付与した正解の前編集がなされたか否かを目視で調査した．表 4 に結果を示す．

表 4: 前編集精度

コーパス	Recall	Precision
closed	94 %	97 %
open	86 %	93 %

[2]A01M0035，A01M0069，A01M0091，A01M0095，M01M0152 の各話者

表 5: 翻訳精度

翻訳方向	日英方向		日中方向	
	整文前	整文後	整文前	整文後
closed	14 %	53 %	55 %	78 %
open	21 %	62 %	16 %	72 %

表 6: 前編集した実験用コーパス

コーパス用途	文字数	翻訳単位数	1 単位の平均文字数
Closed 評価	2,845	108	26.3
Open 評価	3,088	107	28.9

3.4 翻訳品質の評価

翻訳品質は，プロの翻訳者が原文と訳文を目視し，原文の内容が十分に伝わるか否かを 2 値 (「○」か「×」) で判定した．翻訳処理は，前編集前は 3.1 節に説明した文の単位毎に，前編集後は規則で検出した節の単位毎に，東芝の規則ベース翻訳エンジン（Ｔｈｅ翻訳エンタープライズ）[13] を用いて処理した．全単位中，評価が「○」となった単位の数を翻訳精度としてまとめた結果を，表 5 に示す．

4 考察

表 6 に前編集後に後段の処理，すなわち翻訳処理に渡した単位の数と，その平均文字数を示す．表 2 に示した元の平均文字数と比べて，格段に短く簡潔な表現に改められたことが分かる．

表 7，表 8 に，前編集により翻訳の質が改善した例を示す．表 7 の例では，前編集する前はフィラーや間投詞に訳語がついてしまったり，一文が長すぎるために構文構造が正しく英語に反映されなかった．前編集を行った後では，フィラーが削除され，口語的な言い回しは簡素化され，節ごとに分割されたために，機械翻訳の結果も読みやすくなっていることが分かる．

表 8 の例においても，前編集をしない場合はフィラーが訳出されたり「けれども」が逆接の意味を持つ表現に翻訳されてしまっていた．これに対して，前編集を行った場合はそれらの問題が解消されている．

5 おわりに

制御が容易な規則ベースの変換方法を用い，バリエーションが限られている機能語相当の表現を対象にして，話し言葉の前編集規則を作成した．わずか 5 話者分のコーパスを用いて規則を作成しても，規則作成に用いていない別の話者の評価データ用に対して，翻訳性能を上昇させる効果を持つ前編集ができることが分かった．今後は，規則作成用コーパスの話者を増やし，規則のバリエーションを増やす．

表 7: 前編集の前後における日英機械翻訳結果の改善例 (CSJ: A01M0152)

前編集の前後における日英機械翻訳の変化例			評価
前	日:	もまずえとワイヤレス化ま前回あのーライブホンに関しては有線を使ったということでえーとワイヤレスにしておりますまこれにはえと一昨年にえー制定されましたあの難聴者補助用の専用の電波というものが割り当てられましたのでそれを採用しています	×
	英:	not rubbing – sexagenary-cycle wireless-ized ま前回 – the cable was able to be used about that - live phone, and I am making it - and wireless – since that electric wave for exclusive use for hearing loss person assistance that is obtained to まこれ in the sexagenary-cycle year before last and by which - establishment was carried out was assigned, it has been adopted.	
後	日 1:	ワイヤレス化前回ライブホンに関しては有線を使ったということで	○
	英 1:	About the live phone, the cable was used last time [wireless-ized].	
	日 2:	ワイヤレスにしています	○
	英 2:	Wireless is used.	
	日 3:	これには一昨年に制定されました難聴者補助用の専用の電波が割り当てられました	○
	英 3:	The electric wave for exclusive use for the hearing loss person assistance enacted in the year before last was assigned to this.	
	日 4:	それを採用しています	○
	英 4:	It is adopted.	

表 8: 前編集の前後における日中機械翻訳結果の改善例 (CSJ: A01M0152)

前編集の前後における日中機械翻訳の変化例			評価
前	日:	で特徴としてはこういう特徴があってま今回主に御報告するのはこの二つなんですけれど	×
	中:	出作为特征有这种特征ま今回虽然是这二个可是主要报	
後	日:	特徴としてはこの特徴があって今回主に御報告するのはこの二つです	○
	中：	作为特征有这个特征这次主要报告的是这二个	

参考文献

[1] 阿部鉱久. 文章力の基本. 日本実業出版社, 2009.

[2] 伝康晴, 小木曽智信, 小椋秀樹, 山田篤, 峯松信明, 内元清貴, 小磯花絵. コーパス日本語学のための言語資源：形態素解析用電子化辞書の開発とその応用. pp. 101–122, 2007.

[3] 小林作都子. そのバイト語はやめなさい. 日経ビジネス人文庫. 日本経済新聞出版社, 2008.

[4] 小磯花絵, 西川賢哉, 間淵洋子. 転記テキスト. 国立国語研究所報告 124：日本語話し言葉コーパスの構築法, pp. 23–132. 2006.

[5] Taku Kudo, Kaoru Yamamoto, and Yuji Matsumoto. Applying conditional random fields to japanese morphological analysis. In *Proc. of the EMNLP-2004*, pp. 230–237, 2004.

[6] 前川喜久雄. 『日本語話し言葉コーパス』の概要. 日本語科学, Vol. 15, pp. 111–133, 2004.

[7] 丸山岳彦, 柏岡秀紀, 熊野正, 田中英輝. 日本語節境界プログラム CBAP の開発と評価. 自然言語処理, Vol. 11, No. 3, pp. 39–68, 2004.

[8] 益岡隆志, 田窪行則. 基礎日本語文法 改訂版. くろしお出版, 1992.

[9] 美野秀弥, 田中英輝. 国語辞典を使った放送ニュースの名詞の平易化. 言語処理学会第 16 回年次大会論文集, pp. 760–763, 2010.

[10] Graham Neubig, 秋田祐哉, 森信介, 河原達也. 文脈を考慮した確率的モデルによる話し言葉の整形. 情報処理学会研究報告, SLP-79-17, pp. 93–98, 2009.

[11] 尾嶋憲治, 河原達也, 秋田祐哉, 内元清貴. 話し言葉の成形作業における削除箇所の自動同定. 情報処理学会研究報告, SLP-71-13, NL-185-13, pp. 85–91, 2008.

[12] 下岡和也, 南條浩輝, 河原達也. 講演の書き起こしに対する統計的手法を用いた文体の整形. 自然言語処理, Vol. 11, No. 2, pp. 67–83, 2004.

[13] 東芝ソリューション株式会社. The 翻訳シリーズ, 2014. `http://pf.toshiba-sol.co.jp/prod/hon_yaku/index_j.htm`.

[14] 吉倉孝太郎, 山本和英. 用言等換言辞書を用いた換言結果の考察. 信学技報, 第 113 巻 of *NLC2013-1*, pp. 57–62, 2013.