002023 形態素レベルで見た日本語の会話文と記述文の違い

計算機工学講座 池原研究室　小林 和晃

1　はじめに

記述文を対象にした形態素解析や構文解析の技術が進歩してきた．機械翻訳はそれらの解析結果をベースに処理されている．会話文の機械翻訳を目指すとき，会話文と記述文とでは話し言葉特有の表現や敬語表現など構造の差異 [1,2] が問題になるのだが，形態素解析の段階で差異が吸収できるならば，機械翻訳の処理がそのまま利用できることになる．

そこで本研究では，形態素解析の段階で検出される構造の差異を定量的に分析し，その違いを吸収する方法について明らかにする．

2　特徴的な会話文の収集

2.1　収集の目的と調査対象

会話特有の表現を持った文を収集する．ATR 音声翻訳通信研究所では音声翻訳研究のためにバイリンガル旅行会話コーパス [3,4] を構築した．本研究ではコーパス中から 3,206 事例を対象に分析を行う．事例はいくつかの文で構成されているので文単位で調査する．対象となる日本語文は 3,712 文である．

事例　かしこまりました。では、ご昼食で、三千円以下で、禁煙席をご希望ですね。

2.2　記述文と差異のある会話文の抽出方法

記述文との差異を持つ会話文は，記述文のための解析ツールでは正しく解析されないことが予想される．記述文用の形態素解析システム ALTJAWS は単語単位で 99.5%の精度があるので [5]，このシステムを用いて解析に誤る会話文を特徴的な会話文として収集する．

解析誤りの判定は，形態素境界・文節境界の誤り，語義の誤りに注目して行う．ただし，漢字表記に直せば正しく解析される箇所や，固有名詞のために誤る箇所は，会話文特有の誤りではないため収集はしない．

2.3　抽出結果

2.1 節で述べた文を ALTJAWS で解析したところ，解析に成功した文は 2,632 文 (70.9%)，失敗した文は 1,080 文 (29.1%) であった．1 文中に複数の誤り箇所が含まれており，失敗した箇所は 1,195 箇所であった．以後の分析はこの 1,195 箇所を対象に行う．

3　会話文の特徴分析

3.1　誤り箇所の分類

形態素解析に失敗した品詞は表 1 のようになった.「複数の品詞」とは，本来複数の品詞として解析すべき所が一つの形態素として解析してしまった部分のことである．例えば「〜なんです」といった箇所では，助動詞「な」と形式名詞「ん」と解析すべき所を代名詞「何」と解析した．

表 1: 失敗箇所の品詞

品詞	数 (%)	品詞	数 (%)
複数の品詞	257(21.5)	名詞	103(8.6)
助詞	163(13.6)	補助動詞	91(7.6)
助動詞	155(13.0)	動詞	69(5.8)
接続詞	118(9.9)	その他	132(11.0)
感動詞	108(9.0)	**計**	1195

3.2　誤り分析

形態素解析を誤る原因は会話文でしか使われない単語があることが挙げられる．そこで単語が辞書に無いがために失敗した (未知語誤り) 箇所と，辞書に単語があるにもかかわらず失敗した (既知語誤り) 箇所とに失敗箇所を分けて，分析を進める．1,195 箇所の誤りの中で，未知語誤り箇所は 597 箇所 (50.0%)，既知語誤り箇所は 535 箇所 (44.8%)，そしてどちらとも取れず保留としたものは 63 箇所 (5.2%) であった．表 2，表 3 にそれぞれの誤り箇所の多いものを示す．

表 2: 未知語誤り箇所

箇所	出現数
では (接続詞)	116
けど (助詞)	99
てる (補助動詞)	84
ええ (感動詞)	58
って (助詞)	35
いくら (名詞)	30
実は (副詞)	23
らっしゃる (動詞)	11
いえ (感動詞)	9
どういった (連体詞)	8
その他	124
計	597

表 3: 既知語誤り箇所

箇所	出現数
ていただけます	79
なん	73
はい。	34
お〜いただけます	33
お一人	32
、もし	23
お〜くださいませ	22
それでしたら	20
お待たせ	20
ありがとうございます	17
その他	182
計	535

4　記述文との差異の吸収方法

4.1　対策を行う対象

未知語誤り箇所は，今後辞書に登録するとことで対処できると考えられるので，既知語誤り箇所について考える．さらに ALTJAWS が [*] というコードを出力すると明確な誤りである．この場合，後の処理が完全に止まるので対策が優先される．このように，既知語誤り箇所のうち，明らかに誤りと検出される箇所に関して対策を考える．

4.2　対策

if-then ルールで書き換え規則を作成する．次にルールの例を挙げる．

if V/ていただけ (2719, て頂く)\n ます ([*]1100)
then V/ていただけ (2719, て頂く)/ます (7236)

これにより「ます」を正しく助詞に直しかつ，文節境界を直すことができる．

4.3　ルールの概要

既知語誤り箇所 535 箇所の中で 4.1 節で述べた対象となる箇所は 204 箇所 (38.1%) であり，誤り方の種類数は 25 種類であった．表 4 に出現回数の多い誤り方を示す．この 25 種類について机上シミュレーションで修正可能であることを確認した．

表 4: ルールの適用できる誤り箇所

箇所	出現回数 (%)
ていただけます	79(38.7)
お〜いただけます	33(16.2)
お〜くださいませ	22(10.8)
ありがとうございます	17(8.3)
願えますか	13(6.4)
その他	40(19.6)
計	204

5　おわりに

今回の調査で，誤った箇所においては，50.0%は単語辞書には無い会話文特有の単語が原因であり，44.8%は単語辞書に単語があるが，会話文特有の使われ方をしているために失敗をしていることが分かった．

さらに単語が辞書にあるのに誤った箇所は，接続する単語によって決まった誤り方があり，それを if-then ルールによって正しく書き換える方法を考察した．これにより，単語に辞書があるのに誤った箇所の 38.1%を正しく直すことが出来ると期待できる．今後の課題は書き換え規則の実装と実験，および残りの誤り箇所への対策と検討が挙げられる．

参考文献

[1] 竹沢, 白井, 大山:バイリンガル旅行会話コーパスに見られる話し言葉の特徴分析, 情処研報,2001-NL-141,pp.137-144,2001.[2] 松本, 伝:話し言葉の形態素解析, 情処研報,2001-SLP-36,pp.9-14,2001.[3]T.Morimoto et.al.:A speech and language database for speech translation research,Proc.International Conference on Spoken Language Processiong,pp.1791-1794,1994.[4] 竹沢, 中村, 隅田:ATR の会話音声翻訳研究用データベース, 音声研究,Vol.4,No.2,pp.16-23,2000.[5] 白井, 横尾, 池原, 奥山, 宮崎:多段解析法による日本語形態素解析の精度, 情処大,Vol.3,pp.37-38,1995.