慣用表現を利用した形態素情報収集法

6F－2

小倉 健太郎　北 研二　森元 逞

ATR自動翻訳電話研究所

1. はじめに

単語は文の意味や構文を構成する最も基本的な要素であり、言語現象の分析のための言語データの土台となるものである。日本語では文を文字で表記しようとする時、英語のように単語ごとに分かち書きされることはなく、漢字と仮名の混ざったべた書きがされる。そこで、単語を文字列から抽出する必要がある。文字列から単語列を抽出しながら形態素処理することを形態素解析と呼ぶ。

形態素解析においては普通複数個の可能性が生じるが、この中から尤らしい候補を優先的に得る方式として、最長一致法や文節数最小法などを挙げることができる[1][2]。本稿では、慣用表現を利用することにより、尤らしい候補を優先的に得る方法について述べ、言語データベースでの形態素情報収集[3][4][5]に慣用表現を利用した結果を示すことより、その有効性を明かにする。

2. 慣用表現の抽出方法[6]

慣用表現として、いくつかの単語を一まとめとして処理することによる仕事の削減量を基準として、慣用表現を抽出する。

A　…単語列
|A|　…単語列の長さ(単語列中の単語の数)
n(A)　…単語列のテキストでの出現回数

とする。仕事の削減量R(A)は

$R(A)=(|A|-1)*n(A)$

で表わすことができる。すなわち、テキストに|A|個の単語からなる慣用表現Aがあったとする。ひとつの単語を処理するのに1の仕事量がかかるとすれば、もしAのそれぞれの単語を別々に処理したとすれば、|A|の仕事量になる。慣用表現Aが一回出現するごとに、|A|-1の仕事量が削減でき、テキスト全部を処理する場合、(|A|-1)*n(A)の仕事量が削減できる。基本的にはR(A)を計算し、上位のものを慣用表現として選べばよい。

ただし、ある単語列にある単語列が含まれるような場合を考慮する必要がある。AをBの部分文字列であるとする(例えばA＝“ました”、B＝“わかりました”)。

A、B両方を慣用表現としてテキストを処理することを考えると、最長一致で処理を行なう場合、Aの出現回数n(A)のうちn(B)回はBが参照されるので、Aを慣用表現としたことによる実質の仕事の削減量は、

$AR(A)=(|A|-1)\times(n(A)-n(B))$

である。一般的には

$AR(A)=(|A|-1)\times(n(A)-\sum_{i} n(Bi)+\sum_{j} n(Cj))$

(ただしCjはB1……B_iのどれかで、B1……B_iのどれかをふくむような単語列)

で計算する必要がある。

3. 形態素情報収集のための慣用表現の知識表現

ATRでは、単語の形態素情報として表記データ、ひらがな、標準表現、品詞[7]、活用型、活用形、音便を持たせている。

“表記データ”とはテキストに現われたままのものであり、“ひらがな”とはそれをひらがな表記したデータである。“標準表現”は、活用のある単語の場合は、その単語の終止形で、表記にゆれがある場合は典型的に使われる表記である。“品詞”は動詞、名詞、形容詞、助動詞などの分類データである。活用のある単語については、“活用形”に終止形、連体形などのそこでの活用形が入り、“活用型”には、五段活用、下一段活用などの分類が入る。“音便”では、撥音便、促音便、い音便、う音便の区別を行なう。

形態素情報の収集では、慣用表現は処理効率を上げるための一まとまりの処理単位と考えられる。慣用表現から、その慣用表現に含まれる単語の形態素情報を取り出す必要がある。そのため、慣用表現の知識表現は以下のようになっている。

“わかりました”の例

出現表記	わかりました
単語分割	わかり/まし/た
ひらがな表記	わかり/まし/た
標準表現	分かる/ます/た
品詞･活用型･活用形	(本動詞、五段活用、連用形)/(助動詞、、連用形)/(助動詞、、終止形)

4. 慣用表現の利用の実験

形態素解析を行なう時に、慣用表現を優先的に取り扱うことにより、形態素解析の誤り率をどの程度減らすことができるかの実験を行なった。

ここでは、形態素解析はテキストから可能な総ての候補を取り出すではなく、一番尤らしいと思われる一つの単語列を出力する処理とする。

Morphological Analysis Method using Idioms
Kentaro OGURA, Kenji KITA, Tuyoshi MORIMOTO
ATR Interpreting Telephony Research Laboratories

メディア	イディオム	形態素数	誤項目数	誤単語数	誤区切数	項目誤り率	単語誤り率	区切誤り率
KEY	なし	3,395	3,174	990	374	14.93	29.16	11.02
TEL	なし	3,472	2,755	999	280	13.46	28.77	8.06
KEY	あり	3,395	1,698	705	93	8.59	20.77	2.74
TEL	あり	3,472	1,773	773	97	9.01	22.26	2.79

図1. 慣用表現を利用する場合としない場合の違い

最長一致法と単語の連接条件を使った形態素解析を基準とし、慣用表現を利用し優先的にそれを取り扱った場合と、慣用表現をまったく利用しない場合を比較した(図1)。

項目、単語単語の区切りの三つの誤り率で評価した結果どの場合も慣用表現の有効性を示している。特に、単語の区切り誤りは、10%から3%弱になり、慣用表現の利用が単語の区切りに有効であることが分かった。また、メディアによる慣用表現の有効性の違いは見られなかった。

慣用表現を抽出するために使ったテキストデータはおよそ1万語、テストデータは、慣用表現を抽出するために使ったテキストデータとは異なるおよそ3千語のテキストデータである。形態素解析に使用した辞書はおよそ14,000語である。

5. メディアによる慣用表現の違い

通信メディアの違いによる慣用表現の違いを調べるため、国際会議の問い合わせを対象領域として、キーボードを使った会話と電話会話から慣用表現を抽出し比較した。図2にキーボード会話からの慣用表現の例を、図3に電話会話の慣用表現の例を示す。

でしょ,う,か
です,か
し,たい,の,です,が
の,です,が
わかり,まし,た
の,でしょ,う,か
はい,わかり,まし,た
の,です,か
ます,か
どうも,ありがとう,
ございまし,た
し,て
の,です
お,願い,し,ます
て,おり,ます
ませ,ん
そう,です,か
たい,の,です,が
し,ます
と,思い,ます
まし,た

図2. キーボード会話からの慣用表現

でしょ,う,か
ん,でしょ,う,か
ん,です,けれども
の,方
あ,そう,です,か
そう,です,か
はい,わかり,まし,た
と,いう,こと
ん,です
です,ね
ます,でしょ,う,か
ああ,そう,です,か
し,て
と,思い,ます
な,ん,です,けれども
の,方,に
て,おり,ます
の,方,は
わかり,まし,た
に,なっ,て,おり,ます

図3. 電話会話からの慣用表現

ともに、仕事の削減量ARが上位20個の慣用表現である。対象領域が同じであったため、語彙の違いによる慣用表現の大きな違いは見受けられなっかった。同じ慣用表現が多数見られる。

電話会話では、“んでしょうか”、“んですけれども”、“んです”、“なんですけれども”のような準体助詞の“ん”になっているのに対して、キーボード会話では“のですが”、“のでしょうか”、“のです”のように準体助詞の“の”が含まれるややかたい慣用表現になっている。ただし、“の”と“ん”の違いを除けば同じである慣用表現が多い。

また、電話会話では、接続助詞の“けれども”や終助詞の“ね”がついたものが多い。婉曲的な表現を表わす接続助詞の“が”が付いたものはキーボード会話と電話会話に共通に見られる。これは、新聞記事のデータから慣用表現を抽出した時には見られないものである。これらは、キーボード会話が話言葉と書き言葉の中間的なところに位置していることを示唆している。

また、電話会話の慣用表現の“の方”、“の方に”、“の方は”などは、“の”の前のものを直接指すことをさけた婉曲的な表現であり、話言葉一般に共通する慣用表現といえる。

6. おわりに

本稿で示した方法により、慣用表現を組織的に抽出することができる。また、このような方法で抽出した慣用表現が有効に活用できることを、慣用表現を利用した形態素情報の収集の結果により明かにした。

また、慣用表現を実際に形態素情報収集に利用するために、慣用表現を登録、修正、削除できる慣用表現管理システムを用意している。形態素解析システムでは、慣用表現管理システムで管理する慣用表現辞書を利用して、形態素解析を行なう。慣用表現管理システムは形態素解析システムとは独立したシステムであり、他のシステムで利用することもできる。

さらに、分野を限定すれば、メディアによる慣用表現の大きな違いはなく、特徴的な違いを考慮すれば、慣用表現をキーボード会話と電話会話で共用できる可能性を示した。

<参考文献>

[1] 吉村·日高·吉田、日本語文の形態素解析における最長一致法と文節数最小法について、情報処理学会NL研30-7
[2] 吉村·武内·津田·首藤、コスト最小法を用いた日本語文の形態素解析、情報処理学会NL研87-NL-60-1
[3] 小倉·篠崎·森元、形態素情報収集支援システム、情報処理学会第38回全国大会4E-1、1989
[4] 篠崎·水野·小倉·吉本、形態素情報利用解説書、ATRテクニカルレポートTR-I-0077、1989
[5] 小倉·橋本·森元、言語データベース統合管理システム、情報処理学会NL研88-NL-69-4
[6] 北·森元、テキストデータベースからの慣用表現の自動抽出、情報処理学会第37回全国大会
[7] 吉本、日本語品詞の分類、ATRテクニカルレポートTR-I-0008、1987