定型表現を利用した効率的な形態素解析の実現

2H－4

安藤　一秋　藤澤　貴之　獅々堀　正幹　青江　順一
徳島大学工学部知能情報工学科

1．はじめに

定型表現は，慣用表現と並び文章中に頻繁に出現する表現である．また，この表現は，それ自体が独自の意味を持つものと，持たないものがあるが，一つのまとまった単位で処理することで，処理効率や精度を上げることができる[1]．

定型表現や慣用表現の利用は有効であるが，これらは，表現形式の客観的な定義が不明瞭なため，収集作業には時間と労力を要する[2]．しかし，コーパスなどから自動収集する研究も多数報告されている[1][3]．

このような定型表現を，ＯＣＲや音声認識システムなどの自然言語処理システムに導入すると，以下のような長所を生じる．１）短い単語で検索された候補より，長い単語で検索された候補の方が情報量が多いため，各種処理システムの認識率や誤り訂正率を向上できる．２）単語レベルだけではなく，構文・意味解析を交えた文脈レベルの誤り訂正が実現できる．

本稿では，有効とみられる付属語的定型表現の抽出結果，及び，それらの定型表現を利用した形態素解析手法について説明する．また，本手法の有効性を確認するため，約７万文のコーパス[4]を対象にした実験結果を示す．更に，ＯＣＲの誤り訂正への応用も示す．

2．定型表現

本稿では，付属語的定型表現を対象とする．付属語的定型表現には，「に関して」，「に基づく」などの文節と文節の関係を表し，助詞的な働きをする定型表現と，「かもしれない」，「なければならない」などの助述的な働きをする定型表現などがある．これらの表現は，一般に構成語順が不変であり，各構成語間の結合も強く，他の語の挿入も起こらないため，これらの表現を一語として扱うのが自然である．また，処理効率の面からみても一語として取り扱う効果が高く，実用的な場合も多い．

本稿で用いる定型表現は，首藤[2]が抽出，分類した付属語的な表現を基にしている．そして，コーパスから頻度をとり，頻度の高い表現と重要である思われる表現を抽出した．以下にその例を示す．

【助詞的働きをする定型表現】

にもかかわらず，だけでなく，にわたって，について，
によって，にとって，のように，に対する，に対して，
のために，とともに，を中心に，ではなく，とはいえ，
ことから，に関する，としても，を通じて，に向けて，
にしても，を使って，に関して，に基づく，として，
による，のため，などの，ために，に対し，をして，
とする，のでは，もので，・・・・・

【助述的働きをする定型表現】

られている，になっている，のではない，てしまった，
なければならない，かもしれない，わけではない，
ないだろう，ものである，ねばならない，てはならない，
からである，なのである，に違いない，ようになる，
だけではない，なくてはならない，てはいけない，
ように見える，そうである，に過ぎない，できている，
ようになっている，・・・・・

以下，これらの定型表現を利用した形態素解析について述べる．

3．定型表現を利用した形態素解析

3．1．辞書

単語辞書は，自立語の表記と品詞コードを登録した自立語辞書と，助詞，助動詞などの付属語の表記と品詞コードを登録した付属語辞書を用いる．また，本実験で利用する定型表現は，付属語的表現を扱うので付属語辞書に登録する．

定型表現の辞書登録形式は，一般の単語と違い，表記と最前部形態素の品詞コード，最後部形態素の品詞コードを登録する．ここで，最前部形態素とは，定型表現を形態素に区切った結果の最初の形態素を指し，最前部形態素は，最後の形態素を指す．

例えば，“に関して”を辞書登録するとき，表記の他に最前部形態素の“に”の品詞コード（格助詞）と最後部形態素の“て”の品詞コード（接続助詞）を登録する．

3．2．形態素解析

本手法は，最長一致法に接続表を利用した解析手法[5]を用いる．接続表の構成を以下に示す．

A Method for Japanese Morphological Analysis Using Frozen Patterns.
Kazuaki Ando, Takayuki Fujisawa, Masami Shishibori, Jun-ichi Aoe
Faculty of Engineering , The University of Tokushima

隣接した２つの形態素をmorpheme$_1$（前形態素），morpheme$_2$（後形態素）とする．接続表はmorpheme$_1$の品詞コードとmorpheme$_2$の品詞コードの対で構成され，２次元配列 connect で表す．morpheme$_1$の品詞コードをｉ，morpheme$_2$の品詞コードをｊとすると，接続可否はconnect［ｉ，ｊ］の値αにより決定される．

$\alpha=1$ ：morpheme$_1$とmorpheme$_2$は接続可能

$\alpha=0$ ：morpheme$_1$とmorpheme$_2$は接続不可能

$\alpha=1$で接続可能な場合，morpheme$_2$を前形態素として解析を続ける．

但し，morpheme$_1$またはmorpheme$_2$が定型表現である場合は．以下の操作が必要となる．

・morpheme$_1$が定型表現である場合，morpheme$_1$の最後部形態素の品詞コードをｉとする．

・morpheme$_2$が定型表現である場合，morpheme$_2$の最前部形態素の品詞コードをｊとする．

“形態素に関する資料”の接続に関する解析例（図１）を以下に示す．

morpheme$_1$を“形態素”とし，morpheme$_2$を“に関する”とする．ここでは，morpheme$_2$が定型表現であるので“形態素”の品詞コードと，morpheme$_2$の最前部形態素である“に”の品詞コードで接続検定を行う．この場合，名詞と助詞であるから接続可能となる．次に，morpheme$_1$を“に関する”とし，morpheme$_2$を“資料”とする．ここで，morpheme$_1$が定型表現であるので，morpheme$_1$の最後部形態素である“関する”の品詞コードと“資料”の品詞コードで接続検定を行う．この場合は，動詞（連用形）と名詞であるから接続可能となる．

４．実験結果

４．１．コーパスに対する実験

本手法の有効性を確認するため，約７万文コーパスに対して実験を行った．単語辞書には，約９７０００語の自立語辞書と約１８００語の付属語辞書を用いた．

定型表現数，最長定型表現，最長定型表現文字数，平均定型表現文字数，最大定型表現形態素数，平均定型表現形態素数，定型表現含有率，計算コスト減少率をまとめたのが表１である．ここで，定型表現含有率は，全付属語形態素中に，定型表現が含まれる割合を示したものである．計算コスト減少率は，定型表現を一つにまとめて処理することにより，接続に関する計算コストが減少する割合である．

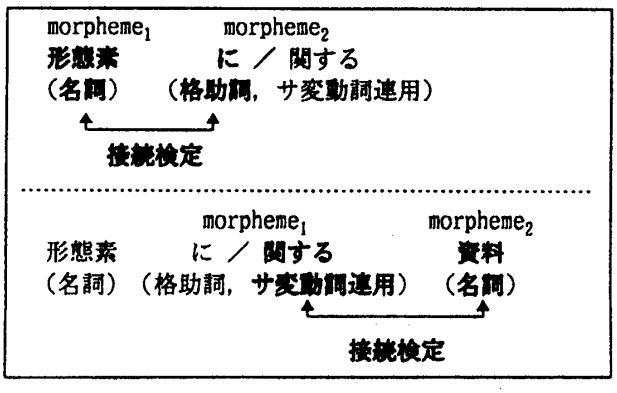

図１：“形態素に関する資料”の接続検定

表１：実験結果

定型表現数	１５５３
最長定型表現文字数	１１
平均定型表現文字数	５．０
最大定型表現形態素数	６
平均定型表現形態素数	２．８
定型表現含有率	１１．５％
計算コスト減少率	６．６％

実験結果より，定型表現を形態素解析に利用することで，計算コストの減少を約６．６％確認することができた．

４．２．ＯＣＲの誤り訂正への応用

認識率が約８８．６％のサンプル文（２９５１文字）に対して，定型表現の有効性の確認を行った．このサンプル文中には，付属語的なものの認識誤りが約１９．３％存在した．この誤りの中から，定型表現に関する誤りの例を示すと，「と・うより」（というより），「らオ１ている」（られている），「かあるのか」（があるのか），「とみ扮１る」（とみられる）などがあった．

これらの定型表現を辞書に登録すれば，誤り回復は容易に行え，その結果として，付属語の認識誤りが約１２．０％に減少し，認識率は，約９０．０％に向上する．

５．まとめ

本稿では，定型表現を利用した形態素解析の実験結果について述べた．定型表現を利用することで接続検定の計算コストの減少が確認できた．

今後の課題として，ＯＣＲや音声認識システムに定型表現を利用し，構文・意味処理を交えた文脈レベルの誤り訂正を行う予定である．

【参考文献】

[1]北研二，小倉健太郎，森元逞，矢野米雄：“仕事量基準を用いたコーパスからの定型表現の自動抽出”，情報処理学会論文，Vol. 34, No. 9(1993).

[2] 首藤公昭：“文節構造モデルによる日本語の機械処理”，福岡大学研究所報，No. 44(1979).

[3]新納浩幸，井佐原均：“コーパスからの関係表現の自動抽出”，情報処理学会論文誌，Vol. 35, No. 11(1994).

[4]EDR電子化辞書，（株）日本電子化辞書研究所.

[5]田中穂積，“自然言語解析の基礎”，産業図書.