自然言語処理 90−10
(1992. 7. 17)

固有名詞の特定機能を有する形態素解析処理

木谷 強
NTTデータ通信（株）開発本部
〒210 川崎市幸区堀川町66−2

カーネギーメロン大学 自動翻訳技術開発センター
訪問研究員

形態素解析プログラム MAJESTY(Morphological Analyzer for Japanese Text Analysis) は、処理対象文書を電子化された新聞記事とし、辞書への未登録語を特定するアルゴリズムと、形態素分割候補と品詞候補の中から確からしい候補を選択するアルゴリズムを採用し、高精度な処理を実現した。さらに、企業名の前後に頻繁に出現する接頭語と接尾語を利用して企業名を特定し、特定した企業名を形態素処理結果へ組み込んで、構文・意味解析パーサーへ渡すインタフェースを備える。その出力形式は、パーサーの処理効率と、既存の文法ルールおよびレキシコン体系に容易に適合可能な形式となっている。本論文では処理方式と評価結果を述べ、アルゴリズムの有効性を示す。

A JAPANESE MORPHOLOGICAL ANALYZER WITH A PROPER NOUN DETECTION ALGORITHM

Tsuyoshi Kitani
Development Headquarters
NTT DATA COMMUNICATIONS SYSTEMS CORP.
66-2 Horikawa-cho, Saiwai-ku, Kawasaki-shi, Kanagawa 210, Japan

Visiting Researcher
Center for Machine Translation
Carnegie Mellon University
Pittsuburgh, PA 15213 U.S.A.

A Morphological Analyzer called MAJESTY (Morphological Analyzer for Japanese Text Analysis) segments Japanese texts and tags parts of speech with high accuracy by introducing algorithms detecting unknown words and selecting the most likely output among possible segmentations and parts of speech. It is aimed to process on-line news articles. A company name detection algorithm is also implemented. Detected company names are grouped and incorporated into the analyzer's output. The output format is designed to fit an existing grammar and lexicon notation and to achieve the high efficiency of a parser. This paper describes the algorithms and shows the effectiveness with evaluation data.

1 はじめに

形態素解析処理は、日本語文章を処理する際の第一段階に位置するものであり、文書推敲、OCRの後処理など、様々な分野の日本語処理アプリケーションに組み込まれている[1] [2]。形態素解析の出力を、構文・意味解析処理の前段階として使用する場合は、パーサー内での解析候補の組合せ数の爆発を抑制するために、曖昧さの少ない高精度な出力が必要となる。形態素解析での誤りの多くは、辞書に定義されていない単語（未登録語と呼ぶ）が原因であり、未登録語の多くは、企業名、人名、地名などの固有名詞によって発生する。財務分野の英語の新聞記事では、全単語の4%以上が企業名であったという報告がある[3]。固有名詞を特定することは、形態素解析の精度を向上させるだけでなく、固有名詞そのものが、データベースへの登録情報やデータベースの検索キーとして利用できるため、利用価値が高い。

本研究の目的は、形態素処理結果を構文・意味解析パーサーへの入力とすることを前提とし、高精度な形態素解析処理を実現することである。処理対象文書は、人力に誤りがなく、かつ表記の基準が比較的定まっている電子化された日本語の新聞記事を想定している。提案する形態素解析プログラムMAJESTY(Morphological Analyzer for Japanese Text Analysis)は、未登録語を特定するアルゴリズムと、複数の形態素分割候補と品詞候補の中から確からしい候補を選択するアルゴリズムを有する。さらに、固有名詞の前後に頻繁に出現する接頭語と接尾語を利用して、固有名詞の文字範囲と企業名、人名、地名などの固有名詞の種類を特定する。そして、その結果を形態素処理結果へ反映し、形態素解析処理の精度を向上させる。パーサーへの出力インターフェースは、パーサーの処理効率と、既存の文法ルールおよびレキシコン体系への適合の容易さを考慮して設計した。

本論文では、MAJESTYの処理アルゴリズムの詳細と、固有名詞の特定処理のうちインプリメントが終っている企業名の特定アルゴリズムを説明し、評価結果をもとにアルゴリズムの有効性を示す。

2 従来の形態素解析の処理方式と問題点

従来の一般的な形態素解析アルゴリズムは、以下の処理ステップで実現されていた。

(1) 文字種類の変化点で入力文章を仮文節に区切る。仮文節とは、例えば、平仮名から他の文字種類への変化点、記号文字の前後、非平仮名列から数字列への変化点、数字列から非平仮名列への変化点を区切り位置とみなした時の、区切られてできた文字列を言う[4]。

(2) 仮文節の範囲内で辞書検索文字列を生成し、辞書を検索する。

(3) 仮文節の区切りが誤っている場合は修正する。

(4) 文法接続テーブルを用いて、前後の形態素の接続条件を判断し、接続が許される形態素をテーブルに登録する。

(5) 登録された形態素の中から、仮文節単位または句読点の単位に、右方向最長一致法により出力する形態素を選ぶ。

従来の我々のアルゴリズムも、仮文節を利用することを除いては上記のステップに従っていたが、問題点が2つ存在した。第1に、上記のステップ(5)において未登録語の範囲を判定する処理がなく、正しくテーブルに登録されている周囲の形態素へも悪影響を及ぼすことが多かった。たとえば、「アメックスとの提携」において、「アメックス」のみが未登録語である場合でも、未登録語の次の自立語から処理を再開するアルゴリズムであったため、未登録語の範囲を「アメックスとの」としていた。パーサーでは付属語は格の情報を示す重要なマーカーとなるため、付属語を正しく出力することが重要である。第2の問題点は、複数の形態素分割候補と品詞候補が存在する場合、それらを確からしい順に出力していなかったことである。このため、従来の我々の形態素解析プログラムの解析精度は95.5%にとどまっていた[5]。（ただし、今回の評価とは評価対象と評価基準が異なる。）

3 本方式による形態素解析処理

3.1 使用する辞書とテーブル

MAJESTYで使用する辞書は、自立語辞書、付属語辞書、活用語尾変化テーブル、カテゴリ接続テーブル、付属語接続テーブルの5つである。これらは、べた書き仮名漢字変換で使用していたものを流用したもので、一般的な内容と構成になっている。

自立語辞書には、表記文字列に対応して、54個の文法品詞カテゴリと形態素の読みを平仮名で格納している。文法品詞カテゴリには、固有名詞の属性として企業名、地名、人名（姓、名）も含む。自立語の単語数は合計で約9万語である。その内訳を表1に示す。自立語辞書には、名詞と名詞が結合した複合語も含めている。自立語辞書の実行形式は、1ブロック512バイトのブロック形式で単語情報を格納したもので、各ブロックへはインデックステーブルを利用して高速にアクセスする。

表1 自立語辞書の登録語数

種類	一般語	姓	名	地名	企業名
語数	73,742	3,974	3,007	7,199	1,827

付属語辞書には助詞と助動詞を格納しており、格助詞、副助詞などの助詞の細分類、または断定、推量などの助動詞の細分類、付属語の接続チェックに使用する前接続番号と後接続番号、および文節終了性の4つの情報を活用形別に登録している。この他に、付属語辞書には接頭語と接尾語も格納している。

活用語尾変化テーブルは、用言の種類と活用形で決まる活用語尾を登録したものである。カテゴリ接続テーブルは、隣接する2つの形態素の接続の可否を判断するために用いる。前方の単語を前単語、後方の単語を後単語と呼び、カテゴリ間の接続の可否をテーブルに定義している。カテゴリは、文法品詞カテゴリと形態素の長さをもとに、前単語を13種類、後単語を12種類に分けている。

付属語接続テーブルは、助詞、助動詞が後単語となる場合の接続の可否を示すもので、全ての助詞と助動詞について、前番号と後番号から接続の可否が判断できる。

3.2 出力形式作成処理

2節で述べたステップ(4)までの処理により、単語登録テーブルへは接続可能な形態素が全て登録されている。ただし、付属語は接続しなくても単語登録テーブルへ登録している。MAJESTYの出力形式作成処理はステップ(5)に相当し、処理対象文字列に未登録語が存在すれば、まず未登録語の範囲を決定する。そして、未登録語範囲の出力と解析可能範囲の出力とを組み合わせながら、出力形式を作成する。Fig.1にMAJESTYの処理フローを示す。太線で示した部分が、今回新たに組み込んだモジュールである。以下に、未登録語範囲決定処理、および形態素分割候補と品詞候補の並び替え・絞り込み処理について説明する。

3.2.1 未登録語範囲決定処理

未登録語範囲の決定に際しての方針は、未登録語の周囲に存在する解析できた形態素に影響を与えないために、未登録語の範囲をできるだけ小さくすることである。そのため、一般的な仮文節よりもさらに細かく、平仮名、カタカナ、英字、数字、記号の5種類の文字種類の変化点で文字列を区切る。区切られた文字列を分割文字列と呼ぶ。未登録語の発生位置は、単語登録テーブルに登録された形態素を後方から接続していき、接続が切れる文字位置とする。この文字位置を含む分割文字列の範囲を未登録語範囲とする。ただし、解析可能範囲が未登録語範囲の直前に組み立てられない場合は、未登録語範囲を拡大し、1つ手前の分割文字列も未登録語範囲に含める。

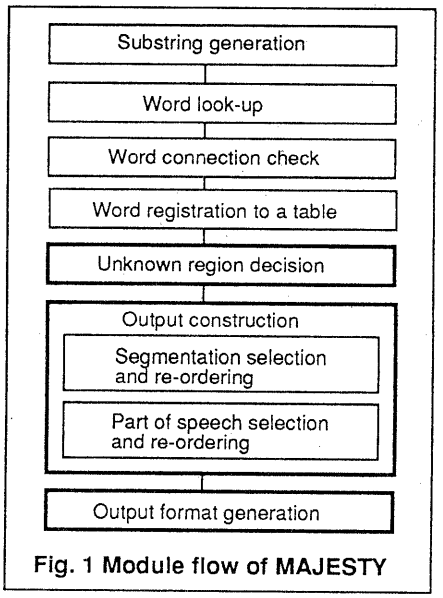


Fig. 1 Module flow of MAJESTY

3.2.2 形態素分割と品詞候補の並び替え・絞り込み

形態素分割が複数生じる場合を分析すると、以下の3通りに分けることができる。

(1) 文脈において1候補のみ正しい定型パターン
例1 {動詞、名詞} + 読点:「狙い、」
候補1:「狙(動詞ワ行五段活用)/い(活用語尾、連用形)/、(記号)」
候補2:「狙い(名詞)/、(記号)」

(2) 複数の正しい候補が存在する定型パターン
例2 助動詞の組合せ:「である」
候補1:「である(指定の助動詞の終止形)」
候補2:「で(断定の助動詞の連用形)/ある(指定の助動詞の終止形)」

(3) 文脈において1候補のみ正しい不定型パターン
例3 名詞の連続:「米国東海岸」
候補1:「米/国東/海岸」
候補2:「米国/東/海岸」
候補3:「米国/東海/岸」

パターン(1)に対しては、統計的にほぼ確からしい候補を第1候補に並び替えて出力するための24個の並び替えルールを定義した。パターン(2)に対しては、文法ルールに適合する1つの候補を出力すれば十分であるため、候補を選択するための4個の絞り込みルールを定義した。

次に、同一の形態素に対して、複数の品詞候補が存在する場合がある。この場合も、形態素分割の候補と同様

に以下の3通りのパターンがある。品詞に対しては並び替えルールを11個、絞り込みルールを4個定義した。なお、ルールは全てソースコード中に定義してある。

(1) 文脈において1候補のみ正しい定型パターン
例4 数字 + {後置助数詞、名詞}:「7月」
候補1:「7（数字）／月（後置助数詞）」
候補2:「7（数字）／月（名詞）」

(2) 複数の正しい候補が存在する定型パターン
例5 助詞または助動詞:「で」
候補1:「で（格助詞）」
候補2:「で（断定の助動詞）」

(3) 文脈において1候補のみ正しい不定型パターン
例6 名詞:「千葉」
候補1:「千葉（名詞－地名）」
候補2:「千葉（名詞－姓）」

3.3 出力インタフェース

文法品詞カテゴリに関しては、既存の文法ルールの文法品詞カテゴリを変更なく利用できるよう、マッピングテーブルを書き換えることで出力する文法品詞カテゴリを変更できる。一般に、パーサー側の文法品詞カテゴリは、解析木の数の爆発を防ぐため、形態素解析内部で使う文法品詞カテゴリよりも少ない数のカテゴリを使用している。したがって、このマッピングテーブルは、カテゴリをまとめる役目も果たしている。

Fig.2は、形態素解析結果の出力例である。この例では、SGMLのタグにより、形態素文字列、品詞、ローマ字読みの各要素を分離している。形態素分割の可能性が複数候補存在する部分は、パーサーでの冗長なレキシコン検索処理を防ぐため、タグ"OR"によって局所的に出力へ埋め込まれている。また、品詞が複数候補ある場合は、タグ"FPOS"内に複数個並べている。出力形式は、タグを定義しているテーブルを変更することにより変更できる。たとえばSGMLタグを、LISPで用いるリスト形式で出力することは容易に可能である。

4 企業名の特定処理

4.1 処理方式

4.1.1 企業名パターン

企業名の特定処理は、MAJESTYの出力を入力とする。新聞記事における企業名の出現パターンは、次のように表すことができる。

[企業名接頭語 [同格語]] 企業名 [企業名接尾語]"(" 説明 ")"

ここで、"[ ]"は省略可能であることを示す。「説明」部分の文字列は任意の文字列であり、" "で囲まれた文字は入力文字列中に存在することを示す。企業名接頭語、同格語、企業名接尾語の例は次の通りである。

(1) 企業名接頭語の例
系、グループ、大手、企業、銀行

(2) 同格語の例
の、である

(3) 企業名接尾語の例
社、系列、協会、銀行、航空、自動車、グループ

たとえば、文字列「大手のABC社（本社、ニューヨーク）」は企業名のパターンに適合し、「大手」が企業名接頭語、「の」が同格語、「ABC」が企業名、「社」が企業名接尾語、「本社、ニューヨーク」が説明となる。

4.1.2 企業名の特定アルゴリズム

以下の3つのステップで企業名を特定する。

(1) 企業名接尾語と企業名接頭語によるパターンマッチング
正規表現で表した企業名接頭語および企業名接尾語と、形態素解析の結果分割された文字列とのパターンマッチングを行なう。正規表現にマッチした形態素の前後の形態素を調べ、企業名パターンにマッチするものを捜す。

(2) 形態素解析で企業名の属性が付与された形態素の処理
形態素解析で企業名とした形態素は、そのまま出力する。

(3) 省略された企業名に対するパターンマッチング
新聞記事では、一度出現した企業名を2回目以降は一部省略して表記することがある。例えば、「野村証券」や「ジェーシービー(JCB)」が、2回目以降は「野村」、「JCB」のように表現されることがある。このような省略に対応するため、ステップ(1)、(2)で特定した企業名から、部分文字列からなる企業名の正規表現をダイナミックに作成し、省略されて表記される企業名を特定する。

4.2 形態素解析結果への組み込み

企業名として出力する部分は、前述の企業名パターンのうち、「企業名」と「企業名接尾語」のみである。企業名を構成する形態素をグルーピングし、企業名の属性を付与する。Fig.3はFig.2の出力に対し、企業名を構成する形態素をグルーピングした例である。これがパーサーとのインタフェース形式となる。

5 評価対象と評価基準

5.1 評価対象文書

評価対象文書は、企業の業務提携に関連する 31 の新聞記事とした。これに含まれる全文字数は 16,428 文字であり、全形態素数は 9,451 形態素であった。使用した自立語辞書に登録されていない未登録語は、評価対象中に 320 形態素（全形態素数の 3.4%）存在した。未登録語の種類の内訳を Fig.4 に示す。評価対象が企業の業務提携に関する記事であったため、一般語の未登録語に次いで企業名が未登録語の 38.4% を占め、商品名、地名、人名（姓）、人名（名）と続いた。なお、
MAJESTY の開発と企業名のパターンマッチングルールの作成には、評価対象とは異なる約 80 の新聞記事を使用した。使用した辞書は仮名漢字変換で使用していた一般的な辞書であり、今回の開発および評価に際して、新たに単語の追加はしていない。また、経済分野の専門用語辞書も使用していない。

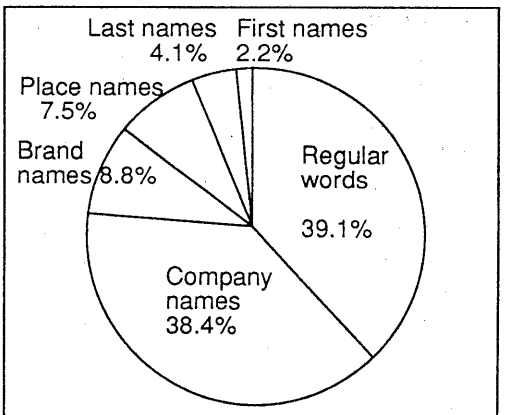


Fig. 4 Distribution of unknown words

```
<SEG>
<TOK><FTOK> J </FTOK><FPOS>ALPH</FPOS><FROM>J</FROM></TOK>
<TOK><FTOK> ・ </FTOK><FPOS>SMBL</FPOS><FROM>.</FROM></TOK>
<TOK><FTOK> A </FTOK><FPOS>ALPH</FPOS><FROM>A</FROM></TOK>
<TOK><FTOK> ・ </FTOK><FPOS>SMBL</FPOS><FROM>.</FROM></TOK>
<TOK><FTOK> ジョーンズ </FTOK><FPOS>?? N</FPOS><FROM>jo-Nzu</FROM></TOK>
<TOK><FTOK> 社 </FTOK><FPOS>N N-PLACE</FPOS><FROM>sha yashiro</FROM></TOK>
<TOK><FTOK> と </FTOK><FPOS>P</FPOS><FROM>to</FROM></TOK>
<OR><DIF>
<TOK><FTOK> 近く </FTOK><FPOS>N ADV</FPOS><FROM>chikaku</FROM></TOK>
</DIF><DIF>
<TOK><FTOK> 近 </FTOK><FPOS>ADJ</FPOS><FROM>chika</FROM></TOK>
<TOK><FTOK> く </FTOK><FPOS>INFL-MIZO</FPOS><FROM>ku</FROM></TOK>
</DIF></OR>
<TOK><FTOK> 提携 </FTOK><FPOS>NSA</FPOS><FROM>teikei</FROM></TOK>
<TOK><FTOK> する </FTOK><FPOS>INFL-SYU7</FPOS><FROM>suru</FROM></TOK>
<TOK><FTOK> 。 </FTOK><FPOS>SMBL</FPOS><FROM>.</FROM></TOK>
</SEG>
```

Fig. 2 An Example of the Morphological Analyzer's output

```
<GRP>
<GTOK> J ・ A ・ ジョーンズ社 </GTOK>
<GPAR>COMPANY</GPAR>
<TOK><FTOK> J </FTOK><FPOS>ALPH</FPOS><FROM>J</FROM></TOK>
<TOK><FTOK> ・ </FTOK><FPOS>SMBL</FPOS><FROM>.</FROM></TOK>
<TOK><FTOK> A </FTOK><FPOS>ALPH</FPOS><FROM>A</FROM></TOK>
<TOK><FTOK> ・ </FTOK><FPOS>SMBL</FPOS><FROM>.</FROM></TOK>
<TOK><FTOK> ジョーンズ </FTOK><FPOS>?? N</FPOS><FROM>jo-Nzu</FROM></TOK>
<TOK><FTOK> 社 </FTOK><FPOS>N N-PLACE</FPOS><FROM>sha yashiro</FROM></TOK>
</GRP>
```

Fig. 3 An Example of Incorporating a Proper Noun Group into the Morphological Analyzer's Output

5.2 形態素解析処理の評価基準

形態素解析の評価項目は、形態素への分割の精度と品詞の付与精度とした。正解データ作成のため、広辞苑[6]と三省堂の国語辞典[7]を基準として用いた。前者は採録語数が多く主に形態素分割の判定に用い、後者は主に品詞の判定に用いた。MAJESTYの自立語辞書に登録されている複合語は、2つの辞典[6][7]のエントリとは異なっていても正解とした。形態素分割、品詞の付与に関して、以下の評価基準を設定した。

(1) カタカナ文字列（外来語）
外来語の連続に対しては、一般外来語が連続する場合は英単語単位に分割し、固有名詞の場合は分割する意味がないため、分割しないこととした。

(2) 英文字列と数字文字列
英文字列と数字文字列は、形態素分割が正しいものを正解とした。

(3) 固有名詞の属性
形態素解析の段階では固有名詞の属性は考慮せず、名詞であれば正解とした。

6 評価結果と考察

6.1 形態素解析単体の評価

6.1.1 未登録語の特定精度

未登録語の特定精度は、評価文書中の全未登録語数に対する、MAJESTYが正しく分割した未登録語数の割合として求めた。評価の結果、MAJESTYによる未登録語の特定精度は76.9%であった。未登録語特定アルゴリズムを組み込まない場合の特定精度は23.8%であり、未登録語の特定精度は大幅に向上した。特定できなかった単語の多くは、カタカナまたは平仮名の未登録語が部分的に登録語として解析された場合（例「エア／ロス／ベース」、「エアロスベース」が未登録語）と、未登録語の前後にある1文字の漢字が未登録語に併合された場合（例「新業態」、「業態」が未登録語）であった。未登録語の特定に失敗すると、周囲の解析結果へ影響を及ぼし、形態素解析の精度の低下につながる。しかし、文字種類を拠り所にしたアルゴリズムの他には有効なものがなく、これ以上の大幅な精度向上は困難であろう。

未登録語に対しては、品詞を推定して付与した。ほとんどの未登録語は名詞であるが、活用語尾として推定したものが246単語中、1単語のみ存在した。推定した品詞は全て正しかった。その結果、今回の評価対象記事に対して、形態素分割の第1候補と品詞の第1候補に着目すると、従来の形態素解析処理に未登録語特定アルゴリズムのみ組み込んだ場合の処理精度は、93.9%となった。

6.1.2 形態素分割候補と品詞候補並び替え・絞り込みルールの有効性

評価記事に対して、形態素分割の候補が複数個存在した個数は251個であり、このうち、並び替え・絞り込みルールが働いたのは196件(78.1%)であった。ルールの作用別にその効果を示したものがFig.5である。動作したルールのうち、並び替えルールにより、正解の分割候補を第1位に位置付けたものは59.2%を占め、その内訳は、句読点が直後に続く場合の名詞系と動詞系の品詞の選択に関わるルール(3.2.2節の例1)と、動詞の活用種類の選択に関わるルール(例えば、動詞下一段活用と五段活用の選択)とが上位を占めた。次いで効果が大きかったものは、複数の分割候補を1つに絞るルール群で、動作したルールの29.6%を占めた。動作したルールは、助詞と助動詞の選択に関わるルールがほとんどであった（同、例2）。なお、誤って分割候補を並び替えた割合いは4.6%であり、誤って分割候補を絞り込んだ例はなかった。絞り込みルールにより、複数個の形態素分割がある箇所を251件から193件にまで絞り込むことができた。形態素分割の並び替え・絞り込みルールが動作しなかった部分は、名詞が連続する場合の分割（同、例3）が多かった。この並び替えには、単語の使用頻度を利用することが有効であろう。

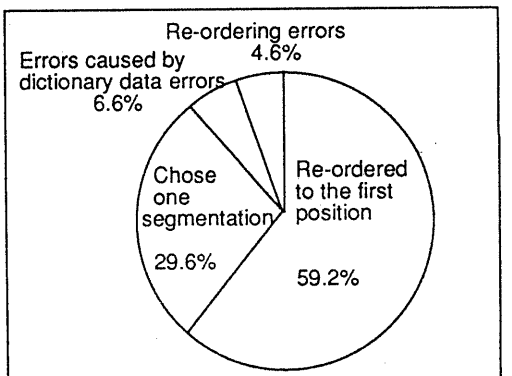


Fig. 5 Results of applied segmentation rules

次に、同一の形態素に対して複数の品詞候補が存在した形態素数は1,029個であり、このうち、品詞の並び替え・絞り込みルールが働いたのは、568件(55.2%)であった。Fig.6にルールの作用別の効果を示す。品詞の並び替えが動作しなかった部分は、企業名、人名、地名などの固有名詞の属性を複数持つ名詞（同、例6）が50.3%を占め、次いで、名詞と、形容動詞または副詞であるものが30.4%であった。固有名詞の属性の選択

に関しては、固有名詞特定プログラムによって大部分は解決できる。名詞と、形容動詞または副詞の選択については、広辞苑 [6] でそれらの区別を明確に付けていないように、それらの間には用法上の差が少なく、構文・意味解析パーサーに判断を任せるのが妥当である。

並び替え・絞り込みルールを組み込むことによって、今回の評価対象記事に対して形態素分割の第1候補と品詞の第1候補に着目した場合、処理精度を 98.2 % にまで引き上げることができた。形態素解析処理が高精度であるなら、パーサーは形態素解析出力の上位候補のみを処理対象とすることができ、高速で高精度な処理が可能となる。

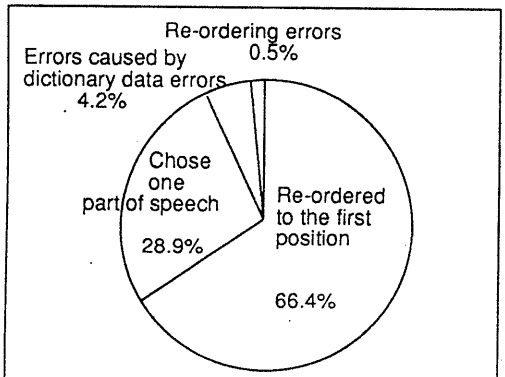


Fig. 6 Results of applied part of speech rules

6.1.3 MAJESTY の処理精度

MAJESTY の処理精度は、全正解形態素数に対する出力の正解形態素数の割合として算出した。表 2 に示すように、形態素分割と品詞付与の両方に着目した場合の処理精度は、形態素分割の第1候補と品詞の第1候補に対して 98.2%、形態素分割の第1候補に対し品詞の全候補を対象とすると 98.4% となった。なお、品詞には着目せず形態素分割の第1候補のみ着目した時の処理精度は 98.5%、第2候補以下も含めると 99.1% となった。

表 2 MAJESTY の処理精度

形態素分割＼品詞	第1候補のみ	全候補	着目せず
第1候補のみ	98.2%	98.4%	98.5%
全候補	−	−	99.1%

形態素分割の第1候補における分割誤りの内訳を Fig.7 に示す。外来語の未登録語によるものが 30.7% を占め、これに、人名などの固有名詞と経済分野の専門語、および、平仮名の単語が登録されていなかったものによる分割誤り 29.3% を含めると、未登録語による分割誤りは全体の分割誤りの 60.0% であった。次いで、並び替えルールによる複数候補の選択の誤りが 22.0% であった。品詞の誤りは、23 例のうち 13 例がルールによる複数候補の選択誤りであり、10 例は辞書データの誤りとプログラムのバグとによるものであった。

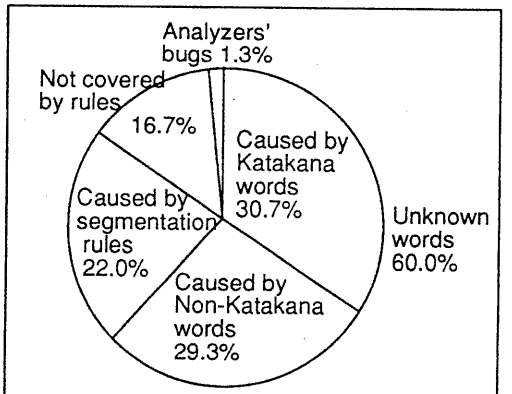


Fig. 7 Causes of segmentation errors

6.2 企業名の特定精度

31 記事に存在した企業名は 312 件であり、このうち 263 件が企業名として特定できた (再現率 84.3%)。一方、企業名でないものを企業名とした件数は 60 件であり、適合率は 81.4% であった (表 3)。

表 3 企業名の特定精度

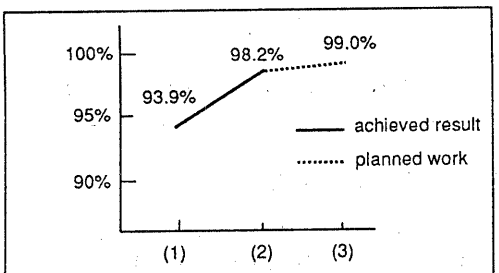
	接頭・接尾語	形態素辞書	省略対応	精度
正解	42.6%	40.7%	16.7%	84.3%
誤り	22.9%	−	77.1%	81.4%

企業名特定アルゴリズムおよび省略語対応アルゴリズムによって正しく特定できた企業名には、未登録語だけでなく、形態素解析で一般語として処理された形態素も含まれていた。企業名が特定できなかった原因は、企業名の前後に企業名接頭語、接尾語がなかったものが 64.7%、企業名の範囲の決定アルゴリズムに起因するものが 32.7% であった。

誤りの原因のうち、省略語対応アルゴリズムによって誤って取り出したもの 77.1% の中の 32.8% は他の種類の固有名詞であり、企業名以外の固有名詞の特定アルゴリズムを組み込むことにより、適合率の向上が期待できる。企業名接頭語、接尾語のマターンマッチングによって誤って取り出したものは 22.9% であった。良く知られている企業名は、接頭語または接尾語を付けずに表記されることが多い。適合率・再現率をさらに向上させるためには、形態素解析辞書に企業名を追加登録したうえ、誤ったマッチングを少なくするよう、パターンマッチングの判定条件を厳しくすることが有効である。

6.3 形態素解析結果と企業名の特定結果を組み合わせた精度

企業名の特定結果を形態素解析結果に組み込むことにより、形態素分割と品詞の誤りを修正することができる。たとえば、「日航系」は形態素解析では1つの未登録語として判断されたが、企業名特定アルゴリズムによって「日航（企業名）／系（企業名接尾語）」のように修正することができた。これにより、形態素分割と品詞の第1候補に着目した場合の処理精度は0.1%向上し、98.3%となった。Fig.8は採用した各アルゴリズムの有効性を示す図である。外来語を中心とした辞書の拡充と、地名および人名（姓、名）の特定アルゴリズムを組み込むことにより、最終的には処理精度を99%程度にまで改善できる見通しを得ている。



(1) Original morphological analyzer + unknown word identification
(2) (1) + Rules for re-ordering and selecting output
(3) (2) + Proper noun detection + Dictionary update

Fig. 8 Accuracy improvement of the morphological analyzer for the most likely segmentation and part of speech

6.4 処理速度

MAJESTYはC言語で記述されており、処理速度はSPARC STATION IPX上で、732文字／秒である。固有名詞特定処理はGNU AWKの日本語版、JGAWKで記述されており、企業名の特定と形態素解析結果への組み込み速度は、今回の評価対象に対して164文字／秒、または2.6企業名／秒であった（処理速度は、全てtime コマンドによる測定）。MAJESTYはパーサーと比較しても、十分に速い処理速度を実現できた。

7 おわりに

本論文では、形態素解析処理の高精度化のため、企業の業務提携分野の新聞記事を対象とし、未登録語の特定アルゴリズムと形態素分割候補および品詞候補の並び替え・絞り込みアルゴリズムを提案し、その有効性を示した。さらに、形態素解析処理の出力に対して、企業名の特定アルゴリズムが適合率・再現率ともに80%以上の精度で動作することを示した。この固有名詞の特定結果を形態素解析に組み込んだ結果、形態素分割の第1候補と品詞の第1候補に対して、98.3%の処理精度を達成することができた。

新聞記事を処理対象とした場合、音声認識の出力結果のように誤りを多く含むものと比較すれば、形態素解析処理で発生する曖昧さは少なく、形態素解析処理とパーサーを独立したプロセスとして扱う構成でも高い精度が得られる。このような背景から、MAJESTYはパーサーとは独立したプロセスとして扱った。パーサー側から見ると、MAJESTYの出力を取り込むことにより、膠着語である日本語を英語のように単語分割された言語と同様に扱え、日本語以外を対象とした既存のパーサーをほとんど変更なく利用できるという利点がある。また、形態素分割と品詞の候補を確からしい順に、かつ冗長性がない形式で出力するので、パーサーは処理時間と処理精度を考慮して、必要な候補数だけ取捨選択して解析することができる。さらに、固有名詞特定機能により、固有名詞を1つのグループとしてパーサーに渡すため、パーサーは効率的な処理が可能となる。

今後は、企業名以外の固有名詞の特定機能を実現し、パーサーの処理効率と処理精度の向上度を評価することが課題である。

謝辞

本研究の場を与えて下さったカーネギーメロン大学Center for Machine Translationの関係各位に感謝致します。なお、評価に使用した新聞記事は、情報抽出プロジェクトTIPSTERの研究のため、DARPAから同センターに提供されたものである。

参考文献

[1] 木谷：”文書推敲処理と目次・索引作成処理を利用した文書作成支援システム”、画像電子学会誌、Vol. 17, No. 5, pp. 337-345, 1988

[2] T. Kitani: “An OCR Post-processing Method for Hand-written Japanese Documents”, Natural Language Processing Pacific Rim Symposium (NLPRS), pp. 38-45, 1991

[3] L. Rau: “Extracting Company Names from Text”, Seventh IEEE Conference on Artificial Intelligence for Applications, Vol. 1, pp. 29-32, 1991

[4] 田中: “自然言語解析の基礎”, pp. 138-142, 1989, 産業図書

[5] 木谷、今福：”文書推敲処理における形態素解析の一検討”、信学全大、626, pp. 2-293, 1987

[6] 新村：”広辞苑 第四版”、岩波書店 (1991)

[7] 金田一：”三省堂国語辞典 第2版”、三省堂 (1974)