形態素情報付きコーパスの再構成手法

田代敏久† 森元逞†

自然言語処理研究者にとって形態素情報付きコーパスは貴重な知識源である．研究のために自由に利用可能な形態素情報付きコーパスもすでに提供されはじめている．しかし，単語認定基準や品詞付与基準などの形態素情報体系には様々なバリエーションが存在するため，求める体系と異なる体系で作成された形態素情報付きコーパスを有効利用することは困難である．そこで，本論文では，形態素調整規則を用いた形態素情報付きコーパスの再構成手法を提案する．独立した2つの研究機関で作成された形態素情報体系付きコーパスの書き換え実験を行い，本手法の有効性を確認した．

A Method to Restructure Tagged Corpora

TOSHIHISA TASHIRO† and TSUYOSHI MORIMOTO†

A part-of-speech tagged corpus is a very important knowledge source for natural language processing researchers. Today we can get several part-of-speech tagged corpora that are available for research use. However, because there are many diversities of a morphological information system (word-segmentation, part-of-speech system, etc.), it is difficult to use tagged corpora that have different morphological information system. In this paper, we propose a method of restructuring tagged corpora using morpheme adjustment rules. We performed rewriting experiments whose targets are tagged corpora that had been developed separately in two research organizations, and verified the effectiveness of this method.

1. はじめに

近年，コーパスに基づく自然言語処理が注目されている．言語コーパスは，単なるテキストファイル（raw corpus），形態素情報付きコーパス（tagged corpus），構文情報付きコーパス（bracketed corpus）等に分類できるが，形態素情報付きコーパスが最も利用価値が高く，多くの研究が形態素情報付きコーパスを利用して行われている[1]．

一般に，コーパスに基づく自然言語処理システムの性能は学習データの量に依存するため，利用できるコーパスはできるだけ大規模であることが望ましい．しかし，大規模な形態素情報付きコーパスの作成は，以下のような困難をともなうため，必ずしも十分なデータを利用できるわけではない．

- データ収集の困難さ：
  コーパスの対象が書き言葉であれ話し言葉であれ，テキストデータを収集すること自体が難しい問題である．書き言葉のコーパスを作成する場合，近年のワープロ等の普及により，電子化されたテキストを手に入れることは比較的容易になってきた．しかし，コーパスとして不特定多数の研究者に利用可能なものとするためには，著作権等の問題を解決する必要がある．一方，話し言葉のデータを収集するためには，話題の設定，会話の収録，書き起こし等，きわめてコストがかかる作業を行う必要がある．
- データ加工の困難さ：
  形態素情報付きコーパスを作成するためには，収集したテキストの形態素解析を行う必要がある．しかし，現状の形態素解析システムで完全に正確な解析を行うことは不可能なので，人手による誤りの発見および修正作業は避けられない．

このデータ量不足の問題を解決する手段として，人間により発見された誤りを形態素解析システムにフィードバックさせることによりシステムの解析能力を向上させ，コーパス作成時の人手の介入を最小限に抑える手法[2]が提案されている．この手法は，“データ加工の困難さ”を軽減する有効な手段であるが，“データ収集の困難さ”を解決することには寄与しない．

“データ収集の困難さ”と“データ加工の困難さ”の

† ATR 音声翻訳通信研究所
ATR Interpreting Telecommunications Research Laboratories

両方を解決するための手段は，言語コーパスを研究者（機関）の間で共有することである．幸いにして，日本においても形態素情報を付与されたコーパスが公開されており，多くの研究者（機関）の間で利用可能になっている[3]．

しかし，単語の認定基準（単語分割の基準）や品詞分類等の形態素情報体系には，標準となる規格が存在しないために，研究者（機関）によって異なっていることが多い．そのため，たとえ共有可能な形態素情報付きコーパスを入手しても，それを有効利用することは難しい，という問題がある．

この問題を解決するためのもっとも単純な方法は，形態素情報付きコーパスの形態素情報を無視して，改めて形態素解析をやり直すという方法である．しかし，形態素解析の精度に限界がある以上，見かけ上どんなに形態素情報体系が異なっていたとしても，多大の労力を注いで作成された情報を無視することは得策でないと思われる．

本論文では，求める体系と異なる体系で作成された形態素情報付きコーパスを有効利用するために，形態素調整規則を用いた形態素情報付きコーパスの再構成手法を提案する．本手法は，異なる形態素情報体系のコーパスの形態素情報を手がかりに求める体系に変換する特殊な形態素解析手法として考えることができる．

本章以降，第2章では形態素情報付きコーパスの再構成手法の概要を述べる．第3章では，実際の形態素情報付きコーパスの再構成実験の結果を報告し，本手法を用いることにより，異なる体系の形態素情報を無視して形態素解析をやり直すより正確に解析が可能であることを示す．

2. 形態素情報付きコーパスの再構成手法の概要

前章で述べたように，形態素情報体系の標準規格が存在しないために，異なる研究機関で作成された形態素情報付きコーパスの形態素情報には様々な相違が存在する．形態素情報の相違には，以下のようなものがある．

- 表記法（正書法）の相違：
  語の表記法は一通りに定まるとは限らない．日本語においては，送りがなの振り方，難読漢字の扱い，外来語の仮名表記等が問題となる．
- 単語認定（単語分割）の相違：
  日本語のように単語区切りを示すマーカが存在しない言語の場合，単語自体を認定すること（文字列を単語列に分割すること）自体が難しい問題である．日本語においては，活用語の語幹と語尾，複合名詞の分割，“に/関/し/て”のような1つの機能語とみなせる連語の扱い等が問題となる．
- 品詞体系の相違：
  どの程度詳細な品詞体系を用いるかは，コーパスの目的に応じて異なる．日本語においては，助詞を機能によって分類するか，固有名詞を意味的に細分類するか等が問題になる．

上記のような相違がある2つの形態素情報付きコーパスを有効利用するためには，これらの相違を書き換え規則（形態素調整規則）として抽出し，一方の形態素情報体系を他方の形態素情報体系へ変換すればよい．しかし，この書き換え規則を人手により作成する作業は，書き換え元（source）の形態素情報体系と書き換え目標（target）の形態素情報体系の両方を熟知した人間が行う必要があり，決して容易な作業ではない．

そこで，我々は形態素情報付きコーパスの再構成（形態素情報体系の変換）を，

(1) 訓練集合の準備
(2) 形態素調整規則の自動抽出
(3) コーパスの書き換え

という3段階で行う．この3段階の中で，訓練集合の準備には人手の介入が必要であるが，作業者は書き換え元（source）の形態素情報体系の知識を持っていなくても構わない．

なお，前述の3つの相違のうち，表記法（正書法）の相違は本来形態素解析以前の問題であることと，書き換え規則の抽出時に2つの形態素データの文字列長が一致していることを利用していることから，本手法の対象からは除外している．以下，本手法の具体的な内容を記す．

2.1 訓練集合の準備

まず，書き換えの対象となるコーパスから文を選ぶ．選んだ文に対して求めようとする形態素情報体系に基づき単語分割と品詞情報付与を行う．この作業においては，求めようとする形態素情報体系に基づく形態素解析システムと人手による修正作業を必要とする．こうして，同一の生（raw）テキストに対して，2種類の形態素情報を持つ訓練集合を作成することができる．**図1** に訓練集合の例を示す．

2.2 形態素調整規則の自動抽出

日本語のように明示的な語境界マーカを持たない言語では，ある形態素情報体系で1語と扱われる語が，別の形態素情報体系では複数の語に分割されたり（一対多対応），複数の語が1つの語としてまとめて扱われたり（多対一対応），複数の語の分割方法が異なっ

生テキスト	日本の自動車産業に覆されてしまう。
形態素情報 1	((日本 固有名詞)(の 連体助詞)(自動車 普通名詞) (産業 普通名詞) (に 格助詞)(覆 本動詞)(さ 語尾) (れ 助動詞) (てしま 助動詞)(う 語尾)(。記号))
形態素情報 2	((日本 名詞)(の 助詞)(自動車産業 名詞) (に 助詞)(覆 動詞)(さ 語尾)(れ 助動詞) (て 助詞)(しま 動詞)(う 語尾) (。記号))

図 1 訓練集合の例
Fig. 1 An example of the training set.

```
[一対一対応規則]
((今日 名詞) <-> (今日 普通名詞))

[一対多対応規則]
((危険性 名詞) <-> (危険 形容名詞)(性 接尾辞))

[多対一対応規則]
((連合 名詞)(体 接尾語) <-> (連合体 普通名詞))

[多対多対応規則]
((肯 動詞)(じる 語尾) <-> (肯じ 本動詞)(る 語尾))
```

図 2 抽出された規則の例
Fig. 2 Examples of the extracted rules.

たり（多対多対応）することがある．また，単語境界の対応とは独立に品詞情報も多対多に対応する可能性があるため，形態素情報の対応はかなり複雑なものとなる．

これらの対応関係を機械的に抽出するためには，Unix の diff コマンドのような最長部分列（longest common substring）を発見するアルゴリズムを応用することも可能だが，2 つの形態素データの文字列の長さが同じであることを利用すると，単語境界が一致する単位ごとに対応づけるという簡単な手続き[4]で求められる．訓練集合にこの手続きを適用することにより，**図 2** に示すような書き換え規則（形態素調整規則）を抽出することができる．

なお，この対応関係の発見と形態素解析の評価法とは密接な関係がある．3.1 節で改めて検討する．

2.3 コーパスの書き換え

2.3.1 形態素調整規則の適用

形態素情報付きコーパスを書き換えるには，前述の形態素調整規則を通常の形態素解析で用いられる辞書（あるいは正規文法に基づく文法規則[5]）と同様に扱い，別の形態素情報が付与された形態素列（単語列）を入力とする形態素解析を行えばよい．つまり，通常の形態素解析において，入力文字列の各要素（文字）と辞書を照合し，要素を単独で語（1 文字単語）として認定したり，連続した複数の要素をまとめて語（複数文字からなる単語）として認定するように，本手法においては，入力形態素列の各要素（単語）と規則を照合するわけである．

通常の形態素解析において曖昧性が生じるのと同様，

```
日本（固有名詞）
    の（格助詞）
    の（準体助詞）
    の（連体助詞）
        自動車（普通名詞）
            産業（普通名詞）
                に（格助詞）
                    覆（本動詞）
                        さ（語尾）
                            れ（助動詞）
                                て（接続助詞）
                                て#しま（助動詞）
                                    しま（本動詞）
                                        う（助動詞）
                                        う（語尾）
―――――――――――――――――――――――――――――――
日本 の 自動車産業 に 覆 さ れ て しま う
```

図 3 ラティス構造
Fig. 3 Lattice structure.

```
日#本(固有名詞)
日(接尾辞)
日(普通名詞)
    本(接尾辞)
    本(普通名詞)
        の(格助詞)
        の(準体助詞)
        の(連体助詞)
            自#動#車(普通名詞)
            自#動(普通名詞)
                車(普通名詞)
                    産#業(普通名詞)
                    産(普通名詞)
                        業(接尾辞)
                            に(格助詞)
                            に(助動詞)
                                覆(本動詞)
                                    さ(語尾)
                                        れ(助動詞)
                                            て(接続助詞)
                                            て#し#ま(助動詞)
                                                し#ま(本動詞)
                                                    う(助動詞)
                                                    う(語尾)
―――――――――――――――――――――――――――――――
日 本 の 自 動 車 産 業 に 覆 さ れ て し ま う
```

図 4 通常の形態素解析によるラティス
Fig. 4 Lattice made by the ordinary morphological analyser.

形態素調整規則の適用によっても曖昧性が生じる可能性があるため，中間データ構造として**図 3** に示すようなラティス構造を用いる必要がある．しかし，形態素調整規則の適用により生じる曖昧性は，通常の形態素解析により生じる曖昧性よりも小さい場合がほとんどである．**図 4** は，図 3 と同じ文に対して通常の形態素解析を行った場合に作成されるラティスであり，形態素調整によるラティスよりもかなり大きいことがわかる．

2.3.2 未知語処理

図 2 に示すような形態素調整規則では，訓練集合に出現する語しか扱えない☆．実際のコーパスを書き

☆ より正確にいうと，訓練集合に出現した語であってもコンテキストの違いによっては扱えない場合がある．

換えるためには，訓練集合に出現しない語（未知語）を処理する必要がある．そこで，入力中の未知語部分に対しては，以下のように規則の制約を緩和して適用する．

- 単語境界の変化をともなわない規則（一対一対応規則）は，品詞情報が一致すれば適用可能とする．
- 入力の複数の語を1つにまとめあげる規則（多対一対応規則）および複数の語の分割方法を改める規則（多対多対応規則）は，未知語を含む区間の各語の品詞情報と文字列長が一致し，さらに各語のうち最低1つは語の表記が一致する場合に適用可能とする．

  たとえば，入力形態素列が

  ((この 連体詞)(構造 名詞)(体 接尾語)(が 助詞) ‥

  であり，“構造” が未知語である場合には，

  ((連合 名詞)(体 接尾語) → (連合体 普通名詞))

  という規則を

  ((任意の2文字 名詞)(体 接尾語) →
  (“任意の2文字 + 体” 普通名詞))

  のように解釈し適用する．
- 入力の語を複数の語に分割する規則（一対多対応規則）は，未知語の品詞情報と文字列長が一致し，さらに分割後の各部分文字列のうち最低1つは規則の右辺の語の表記と一致する場合に適用可能とする．

  たとえば，入力形態素列が

  ((この 連体詞)(有効性 名詞)(は 助詞) ‥

  であり，“有効性” が未知語である場合には，

  ((危険性 名詞) → (危険 形容名詞)(性 接尾辞))

  という規則を

  ((“任意の2文字 + 性” 名詞) →
  (任意の2文字 形容名詞)(性 接尾辞))

  のように解釈し適用する．

なお，通常の形態素解析システムで用いられるような字種に関するヒューリスティックや，固有名詞の特定機能[6]，文字連鎖の統計情報を用いた単語推定機能[7]等を本手法の枠組で用いることも可能である．

2.3.3 ラティスの探索

書き換え規則を適用後の処理は，通常の形態素解析で行う処理とまったく同じである．つまり，曖昧性をパックした表現であるラティス構造から，何らかの基準で最ももっともらしい解を効率的に抽出するわけである．現在一般に利用されている解のもっともらしさを判断する知識としては，

- 最長一致や文節数最小などのヒューリスティック[6),8),9)]
- 語や規則に人手で付けられたコスト[2),5)]
- コーパスから得られた統計的情報[7),10)]

等があり，また効率的にラティスを探索する方法としては

- 動的計画法を用いたグラフ探索[5),11)]
- 前向き動的計画法と後ろ向き A^* 探索を組み合わせた手法[7)]

等が提案されている．これらの知識や探索方法はすべて本手法に適用可能である．

3. 実験と考察

本手法の有効性を確かめるために，2つの形態素情報付きコーパス（話し言葉を対象とするATRコーパス[12)]と書き言葉を対象とするEDRコーパス[13)]）の形態素情報体系を対象とする書き換え実験を行った．この実験の目的は，異なる体系の形態素情報を利用する本手法が，異なる体系の形態素情報を無視する通常の形態素解析より，精度よく解析が可能であることを示すものである．

3.1 評価方法

形態素解析結果の評価方法として，すでに様々な方法が提案されているが，英語の構文解析評価方法[14)]を日本語形態素解析に応用した永田が提案した方法[7)]がもっとも一般的かつ適切な方法と思われるので，本論文でもこれに従う．

永田の方法の詳細な説明は，文献7)に譲るが，この方法で用いる再現率・適合率および交差数と，本手法で用いている形態素情報の対応関係発見（書き換え規則抽出）手続きとの間には密接な関係がある．

まず，再現率・適合率について考察する．単語分割情報の対応関係として，一対一対応，一対多対応，多対一対応，多対多対応の4種類があることはすでに述べた．形態素解析の評価において単語分割がすべて正しい（再現率・適合率が100％である）ということは，すべての語が一対一対応しているということを意味する．逆に，単語分割が誤っている場合は，一対一対応しない語が含まれることを意味し，この場合，再現率・適合率から以下のように単語分割の誤りの傾向を推測することが可能である☆．

- 再現率が適合率より高い場合：

  この場合の典型的な例として，AB/Cという形態素列を誤ってA/B/Cと解析した場合を考える．

☆ これはあくまで形態素解析結果の第一候補の単語分割精度に着目した場合の議論である．解析結果として複数候補を出力した場合の再現率・適合率の意味は，一般に使われる場合と同じである．

この場合，単語分割の対応関係はAB⇔A/Bという一対多対応とC⇔Cという一対一対応となり，再現率・適合率はそれぞれ1/2，1/3となる．つまり，再現率が適合率より高い場合には，単語分割の対応が一対多対応することが多い，言い換えれば形態素解析システムが単語を短く分割しすぎる傾向があることを示している．

- 適合率が再現率より高い場合：
  逆に，A/B/Cを誤ってAB/Cと解析した場合には，再現率・適合率はそれぞれ1/3，1/2となる．つまり，適合率が再現率より高い場合には，単語分割の対応が多対一対応することが多く，システムは必要以上に単語を長単位に認定する傾向があることを示している．

一方，交差数とは，英語の構文解析の評価において，解析結果中の致命的な誤りを示す指標として定義されたものである．形態素解析の評価において，交差数は単語分割の対応が多対多対応になっている場合（AB/CDをABC/Dと解析したような場合）の数に相当する．単語分割の誤りを何らかの後処理で回復しようとする場合でも，多対多対応関係を処理するのは困難であると予測されるので，このような誤りを致命的な誤りとするのは妥当な判断であろう．

3.2 比較実験に用いた形態素解析システム

前述のように，本実験は通常の形態素解析システムより本論文の手法が優れていることを示すことが目的である．そこで，まず比較実験に用いた形態素解析システムの概要と解析能力を明らかにしておく必要がある．

比較実験に用いた形態素解析システムは，様々な言語モデルや探索手法の性能を，実際の解析実験を通じて確認するために，言語モデルや探索手法を容易に変更することが可能なツールとして開発されたものである[15]．今回の実験では，できるだけ条件を単純にするために，言語モデルとしては品詞の2つ組確率と品詞別単語出現確率の積を，探索手法としてはビーム探索（ビーム幅200）を用いている．

未知語の処理もきわめて単純な方法で行っている．入力中の辞書引きができなかった部分を未知語候補とし，未知語候補はすべての品詞を取り得ることにした．未知語の品詞別単語出力確率および未知の品詞2つ組確率はきわめて小さい値（対数尤度で−20.0）をとるとした．

このような形態素解析システムで，ATRの音声言語データベースを対象として解析実験を行ってみた．**表1**は品詞の2つ組確率と品詞別単語出現確率を学習した訓練集合の概要，**表2**はオープンおよびクローズドテストの対象，**表3**は，テスト結果をそれぞれ示している．

表1 形態素解析システムの能力評価：訓練集合
Table 1 The ability of the morphological analyzer: Training set.

文数	語数	品詞の数	異なり語数	品詞2つ組の数
12066	195862	32	4235	392

この形態素解析システムの能力を他のシステムと正確に比較することは，評価方法や実験対象が異なるため難しいが，文献7)の実験結果は，本論文と同一の評価方法およびほぼ同等の実験対象に基づくものなので，かなり正確な比較が可能である．文献7)では，最も基本となるオープンテストにおける単語分割および品詞付与の再現率で，第1候補に対して95.1％（本システムは93.4％），第5候補までに対して97.8％（本システムは97.1％）であったと報告している．本システムの精度は1〜2％程度低いわけだが，これは

- 文献7)では品詞の3つ組確率を用いているの対し，このシステムでは品詞の2つ組確率を用いている．
- 文献7)では前向き動的計画法と後ろ向きA^*探索を組み合わせた手法を用いて正確なN-best候補を出力しているのに対し，このシステムではビーム探索による疑似N-best候補を出力している．
- 文献7)では文字列連鎖の統計情報を用いた高度な未知語推定を行っているのに対し，このシステムの未知語処理はきわめて単純である．

こと等に起因すると思われ，きわめて妥当な結果であるといえる．

また，オープンテストとクローズドテストの結果を比較すると，文献7)ではすべての評価項目で2％程度クローズドテストの結果の方が高いが，この形態素システムでは一部逆転が生じている項目もあり，オープン/クローズドの差はあまり激しくない．これは，品詞の2つ組確率という単純な言語モデルを採用しているために，訓練集合に対する過剰学習が避けられていることを示している．解析の評価としては，とかく正解率の絶対的な数値に注目しがちであるが，過剰学習の有無等も重要な評価基準であり，この点ではこの形態素解析システムは良好な特性を有しているといえる．

以上のように，この形態素解析システムは，形態素解析システムとしてほぼ平均的な能力を持つと判断してよいと思われる．次節以降の実験では，この形態素解析システムと本論文の主眼である形態素調整手法とを比較しながら議論を進める．

表2 形態素解析システムの能力評価：実験対象
Table 2 The ability of the morphological analyzer: Experiment targets.

実験対象	文数	語数	文字列長			未知語数	未知語率
			最長	最短	平均		
クローズドテスト集合	1000	16652	104	3	26.1	-	-
オープンテスト集合	1996	31563	120	3	25.3	1514	4.8%

表3 形態素解析システムの能力評価：実験結果
Table 3 The ability of the morphological analyzer: Experiment results.

実験対象	単語分割のみ				単語分割および品詞付与			
	再現率	適合率	交差数	文正解率	再現率	適合率	交差数	文正解率
クローズドテスト								
(第1候補)	94.5%	98.5%	0.000	81.6%	93.1%	97.0%	0.000	66.5%
(第5候補まで)	97.6%	90.6%	0.013	93.4%	97.3%	78.2%	0.013	87.8%
(第10候補まで)	98.7%	85.2%	0.031	96.3%	98.5%	67.1%	0.034	92.8%
オープンテスト								
(第1候補)	95.2%	97.8%	0.002	79.8%	93.4%	95.9%	0.002	64.3%
(第5候補まで)	97.8%	87.2%	0.021	90.4%	97.1%	76.4%	0.022	82.1%
(第10候補まで)	98.5%	84.0%	0.056	92.7%	98.1%	65.8%	0.064	86.4%

表4 基底実験：訓練集合
Table 4 The base run: Training set.

形態素情報体系	文数	語数	品詞の種類	異なり語数 (辞書サイズ)	品詞2つ組の種類	形態素調整規則の数 (対応関係の数)
EDR 体系	500	10436	15	2963	121	3058
ATR 体系	同上	10282	32	2908	392	同上

表5 基底実験：実験対象
Table 5 The base run: Experiment targets.

実験対象	文数	語数	文字列長			未知語数	未知語率
			最長	最短	平均		
クローズドテスト集合	500	10282	87	6	32.3	-	-
オープンテスト集合	500	10145	102	7	32.7	2073	20.4%

3.3 基底実験

本論文で提案した手法の有効性を確認するための基底実験として，EDR コーパスの形態素情報体系を ATR コーパスの形態素情報体系へ変換する実験を行った．この実験では，本論文の手法と一般的な形態素解析手法を比較するために，前節で述べた形態素解析システムと同じ探索手法（ビームサーチ）と言語モデル（品詞の2つ組確率と品詞別単語出現確率の積）を用いている．

まず，EDR コーパスから 1000 文を抽出し，それらに ATR コーパスの形態素情報を人手により付与した．1000 文を 500 文ずつの訓練集合とテスト集合とに分け，訓練集合から形態素調整規則を抽出し，品詞の2つ組確率と品詞別単語出現確率を学習した．**表4** は訓練集合の概要を，**表5** は，オープンおよびクローズドテスト対象の詳細な情報を示している．

表4 の品詞の種類や異なり語数の項目が示すように，EDR の形態素情報体系と ATR コーパスの形態素情報体系の間には，単語分割基準および品詞付与基準ともかなりの差異がある．これらの差異の主なものとしては，以下のようなものがある．

- EDR 体系では，“に/関/し/て” のような機能的に助詞に相当する表現や，“て/い/る” のような機能的に助動詞に相当する表現を短単位に分割しているのに対し，ATR 体系ではこれらを1語として扱っている．
- 名詞複合語句に関しては，どちらの体系においても，ある場合は1語扱いされたり，ある場合には短単位に分割されたりしているが，その基準は単純ではなく，しかも両体系で異なっている．
- EDR 体系では，名詞類はすべて単に “名詞” と扱っているのに対し，ATR 体系では “普通名詞”，“固有名詞”，“サ変名詞” 等，統語的・意味的な基準により分類している．
- EDR 体系では，助詞類はすべて単に “助詞” と扱っているのに対し，ATR 体系では “格助詞”，“連体助詞”，“接続助詞” 等，統語的・意味的な基準により分類している．

表6 2つの形態素情報体系の違い
Table 6 Differences between the two morphological information system.

比較対象	単語分割のみ			
	再現率	適合率	交差数	文正解率
クローズドテスト集合	90.0%	88.7%	0.010	26.4%
オープンテスト集合	90.1%	88.6%	0.004	27.6%

前述の形態素解析の評価基準は，2つの形態素体系の単語分割基準がどの程度異なっているかを示す指標ともなる．**表6**は，ATR体系の単語分割を正解とみなした場合の，EDR体系の単語分割情報の評価結果である☆．この表から，両体系において全体の約25%の文しか単語区切りさえも一致しないことや，適合率が再現率より低い，つまりEDR体系ではATR体系より単語を細かく分割している傾向があること等を読みとることができる．

これほどの差異を持つEDR体系の形態素列に対して，前章で説明した形態素調整手法を適用した結果を**表7**に示す．表中の丸括弧内の数字は，EDR体系の形態素情報を無視して文字列から通常の形態素解析を行った場合の解析結果である．

まず，クローズドテストの結果を見ると，形態素調整も形態素解析もかなり高い精度で解析しているが，すべての評価項目で形態素調整の方が良い値を示している．ラティス探索のための言語モデルと探索手法は同一であるのに両者の性能に差があるのは，形態素調整は別の形態素体系の情報を利用して探索空間の絞り込みを行っているのに対し，形態素解析は何も行わないことに起因する．

次に，オープンテストの結果を見ると，未知語率20.4%，つまり5語に1語は未知語であるというきわめて苛酷な条件の実験結果であるために，かなり低い正解率となっている．それでもなお，形態素調整は元の形態素情報の差異をかなり解消しているのに対し，通常の形態素解析は，単語分割の情報に限れば元の差異を拡大してしまっている．これも，形態素調整は別の形態素体系の情報を活用しているのに対し，形態素解析はそうではないことが原因である．

今回の実験での未知語処理，特に形態素解析の未知語処理は，あまりに単純な仕組みであるために，オープンテストの結果をそのまま両者の能力の差とするのには少し無理がある．そこで，オープンテスト集合を文中の未知語の数で分類し，両者の解析結果を比較してみたのが，**表8**である．この表から，未知語の数が少ない場合でも，形態素解析より形態素調整の方が高い解析能力を持っていることがわかる．

3.4 辞書情報を利用した性能向上

前節の基底実験結果では，本論文で提案した形態素調整が通常の形態素解析よりも優れていることは示せたが，実際にコーパスを書き換える作業を行うことを考えると，オープンテストの結果には不満が残る．

オープンテストの結果を向上させるための一番単純な手法は，訓練集合を大きくして，より多くの形態素調整規則を抽出し，より正確な品詞の2つ組確率と品詞別単語出現確率を学習すればよい．実際の作業時には，いわゆるbootstrappingの手法を用いて，解析結果を人手で修正したものを新たに訓練集合に加えることを繰り返していけばよいだろう．しかし，その場合においても，人手による修正作業のコストを考えると，もう少し高い精度で解析できることが望ましい．

表8が示すように，オープンテストにおける性能悪化は，主として未知語に起因するものなので，未知語を減少させれば性能が向上することが予測できる．そこで，ATRの形態素体系に基づく54282語からなるATRコーパス作成用辞書を利用し，未知語を減少させることにした．

形態素調整において，訓練集合に含まれない単語の辞書情報を利用するには，ラティス探索に用いる品詞別単語出現確率を変化させるしかない☆☆．そこで，訓練集合には出現しないが形態素辞書に含まれる語の頻度を1として，訓練集合に出現したすべての語の頻度には1を加え，品詞別単語出現確率を再計算した．一方，通常の形態素解析システムの辞書には，訓練集合から抽出した語に加えて，訓練集合には出現しないが形態素辞書には含まれる語も含むこととし，さらにラティス探索に用いる品詞別単語出現確率も形態素調整と同じように再計算したものを使うようにした．これにより，オープンテスト対象中の未知語率は，20.4%から6.1%に減少した．この条件のもとで新たに解析をやり直した結果を**表9**に示す．辞書情報を追加することにより，通常の形態素解析と形態素調整のいずれの精度も向上したが，それでもなお形態素調整の方がすべての項目で良い結果を示している．

3.5 逆方向への書き換え

2.2節で説明した形態素調整規則は，2つの形態素情報の対応関係を表現したものなので，左辺の形態素情報から右辺への書き換えにも，またその逆方向への書き換えにも利用できる．

☆ EDR体系を正解とみなした場合には再現率と適合率が逆になる．

☆☆ 形態素調整規則を増やすためには，各語のEDR体系での品詞や単語区切りがどうなるかを考えなければならないので，訓練集合を大きくする以外の簡単な手法は存在しない．

表 7 基底実験：実験結果
Table 7 The base run: Experiment Results.

実験対象	単語分割のみ				単語分割および品詞付与			
	再現率	適合率	交差数	文正解率	再現率	適合率	交差数	文正解率
クローズドテスト								
第 1 候補	99.5%	99.7%	0.000	95.0%	98.0%	98.2%	0.000	74.0%
	(99.0%)	(99.4%)	(0.000)	(90.6%)	(96.0%)	(96.3%)	(0.000)	(56.6%)
第 5 候補まで	99.9%	96.6%	0.006	99.4%	99.9%	85.3%	0.006	95.2%
	(99.8%)	(93.5%)	(0.012)	(97.4%)	(98.9%)	(82.2%)	(0.012)	(83.8%)
第 10 候補まで	99.9%	95.5%	0.006	99.8%	99.9%	79.7%	0.006	98.0%
	(99.8%)	(90.0%)	(0.020)	(98.4%)	(99.6%)	(74.7%)	(0.020)	(90.4%)
オープンテスト								
第 1 候補	94.3%	93.1%	0.008	49.8%	85.6%	84.6%	0.008	12.8%
	(77.9%)	(70.7%)	(1.012)	(8.2%)	(69.2%)	(62.8%)	(1.012)	(2.8%)
第 5 候補まで	94.4%	92.9%	0.010	50.0%	88.7%	73.9%	0.016	23.4%
	(78.2%)	(69.6%)	(1.070)	(8.2%)	(71.2%)	(54.3%)	(1.528)	(5.2%)
第 10 候補まで	94.5%	92.5%	0.010	50.6%	90.0%	67.0%	0.018	27.2%
	(78.4%)	(68.3%)	(1.168)	(8.2%)	(72.2%)	(49.0%)	(1.984)	(5.4%)

表 8 未知語と正解率の関係
Table 8 Relation between unknown word and accuracy.

文中の未知語の数	文数	平均語数	第 1 候補の再現率	
			単語分割のみ	単語分割と品詞付与
0	14	11.1	100%	98.7%
			(100%)	(98.1%)
1 以下	76	12.7	98.0%	93.2%
			(94.3%)	(89.4%)
2 以下	156	13.3	95.9%	90.3%
			(90.2%)	(84.5%)
5 未満	317	16.0	95.3%	87.4%
			(84.3%)	(77.1%)
全体	500	20.3	94.3%	85.6%
			(77.9%)	(69.2%)

表 9 辞書情報の利用
Table 9 Use of lexical information.

	単語分割のみ				単語分割および品詞付与			
	再現率	適合率	交差数	文正解率	再現率	適合率	交差数	文正解率
クローズドテスト								
第 1 候補	99.5%	99.8%	0.000	95.6%	98.2%	98.5%	0.000	77.4%
	(97.3%)	(98.5%)	(0.004)	(77.0%)	(94.1%)	(95.3%)	(0.004)	(48.8%)
第 5 候補まで	99.9%	97.2%	0.006	98.8%	99.7%	85.3%	0.006	95.4%
	(98.6%)	(93.2%)	(0.046)	(88.4%)	(97.7%)	(81.6%)	(0.046)	(74.2%)
第 10 候補まで	99.9%	96.2%	0.006	99.8%	99.9%	80.2%	0.006	98.2%
	(99.1%)	(88.7%)	(0.076)	(91.4%)	(98.6%)	(72.5%)	(0.080)	(81.8%)
オープンテスト								
第 1 候補	94.9%	94.8%	0.006	56.0%	90.6%	90.5%	0.006	28.8%
	(91.5%)	(86.3%)	(0.096)	(32.2%)	(86.6%)	(81.6%)	(0.096)	(19.0%)
第 5 候補まで	95.1%	93.9%	0.012	57.2%	92.5%	78.3%	0.018	37.8%
	(92.7%)	(83.0%)	(0.158)	(37.0%)	(89.9%)	(70.7%)	(0.206)	(28.0%)
第 10 候補まで	95.2%	93.0%	0.012	57.4%	93.0%	70.7%	0.022	39.8%
	(93.1%)	(80.3%)	(0.196)	(38.4%)	(90.8%)	(63.6%)	(0.284)	(30.8%)

表 10 は，前節までの実験の逆方向の実験，つまり ATR の形態素情報体系から EDR の形態素情報体系への書き換え実験の結果を示している．なお，この実験は 3.3 節で述べた基底実験の逆方向実験であり，使用した形態素調整規則や言語モデルは訓練集合から求めたもののみを用いている．この実験結果においても，全体的な傾向は EDR から ATR への書き換え実験とほぼ同じとなった．

文献 4) では，細かい品詞体系から粗い体系への結果の方が高い精度で解析できるとしているが，今回の結果ではそれほどの差は確認できなかった．これは，品詞体系の粗さは確かに探索空間を狭めるが，同時に探索時の知識である品詞 2 つ組モデルの制約も緩めてしまうことによると思われる．一方，文献 4) では，探索時の知識として単語そのものの 2 つ組確率を用いているため，品詞 2 つ組モデルの制約の緩みがそれほど問題にならないわけである．

3.6 誤り傾向と対策

3.3 節において，ATR の形態素情報体系と EDR の形態素情報体系の差異について述べた．本節では，形

表10 逆方向への書き換え
Table 10 Rewriting in the opposite direction.

	単語分割のみ				単語分割および品詞付与			
	再現率	適合率	交差数	文正解率	再現率	適合率	交差数	文正解率
クローズドテスト								
第1候補	99.3%	99.6%	0.000	93.4%	97.4%	97.7%	0.000	63.4%
	(98.4%)	(99.0%)	(0.002)	(85.4%)	(95.5%)	(96.1%)	(0.002)	(49.2%)
第5候補まで	99.9%	97.4%	0.002	98.6%	99.7%	88.0%	0.002	93.6%
	(99.4%)	(93.8%)	(0.026)	(94.6%)	(98.8%)	(83.1%)	(0.026)	(82.4%)
第10候補まで	99.9%	96.7%	0.004	99.4%	99.9%	85.3%	0.004	98.2%
	(99.6%)	(90.5%)	(0.044)	(96.4%)	(99.3%)	(76.1%)	(0.044)	(89.8%)
オープンテスト								
第1候補	94.7%	92.6%	0.020	44.2%	89.4%	87.5%	0.020	19.4%
	(78.5%)	(72.5%)	(0.748)	(7.0%)	(65.5%)	(65.5%)	(0.748)	(3.6%)
第5候補まで	94.9%	91.9%	0.022	45.4%	92.0%	77.1%	0.022	30.8%
	(79.2%)	(71.1%)	(0.796)	(7.0%)	(73.8%)	(57.2%)	(1.186)	(5.2%)
第10候補まで	94.9%	91.6%	0.022	46.4%	92.6%	71.1%	0.022	34.0%
	(79.4%)	(70.2%)	(0.812)	(7.0%)	(74.7%)	(52.0%)	(1.494)	(5.8%)

表11 誤り種別の変化
Table 11 Changes of error types.

種別	書き換え前	基底実験	辞書追加
誤り全体	668	387	314
複合名詞	301	276	228
助詞・助動詞類	327	71	54
その他	41	40	32

態素調整規則の適用により2つの体系の差異がどのように変化したかを述べる．

表11は，EDRの形態素体系のテスト集合の書き換え前，書き換え後，さらに辞書情報を追加した書き換え後の形態素列集合と，ATRの形態素体系のテスト集合との間の単語分割の誤り（差異）を内容別に分類したものである．この表から，助詞・助動詞等の単語分割の誤りは，訓練集合の形態素調整規則だけでもかなり修正できるのに対し，複合名詞等の誤りを規則によって修正することは難しく，辞書情報のような個別的な知識の追加によって，ひとつひとつ対応するしかないということがわかる．

複合名詞，特に接尾辞をともなう名詞派生語の解析を的確に行うために，確率文法とシソーラスを用いる手法がすでに提案されている[16]．しかし，これを形態素情報付きコーパスの再構成に応用するには，シソーラスの単語認定基準とコーパスの単語認定基準の差異も考慮に入れる必要があり，問題はかなり複雑である．1つの解決策として，EDRコーパスに付与されている概念情報を活用する方法を現在検討中である．

4. おわりに

本論文では，求める形態素情報体系と異なる体系に基づいて作成された形態素情報付きコーパスを有効利用するために，形態素調整規則を用いた形態素情報付きコーパスの再構成手法を提案した．独立に作成された2つの形態素情報付きコーパスを対象とする実験を行い，本手法の有効性を確認した．本手法は，入力を形態素列とする特殊な形態素解析手法であり，通常の形態素解析で用いられる言語知識や探索手法を容易に応用することができる．

今後の課題として，本手法の性能をさらに高めるために，

- 通常の形態素解析で用いられる未知語に関するヒューリスティックを応用する．
- 形態素調整規則自体に何らかのコストを付与し，ラティス探索時に利用する．

こと等を予定している．

謝辞 本研究の機会を与えて下さったATR音声翻訳通信研究所の山崎泰弘社長に感謝します．御指導下さった第4研究室の皆様に感謝します．

参考文献

1) 野美山他：「コーパスを利用した自然言語処理」サーベイ，情報処理学会自然言語処理研究会，104-12 (1994).
2) Maruyama, H., Ogino, S., Hidano, M.: The Mega-Word Tagged-Corpus Project, *TMI-93*, pp.15–23 (1993).
3) 竹沢，末松：音声・テキストコーパスとその構築技術，標準化動向，人工知能学会誌，Vol.10, No.2, pp.168–180 (1995).
4) Tashiro, T., Uratani, N., Morimoto, T.: Restructuring Tagged Corpora with Morpheme Adjustment Rules, *COLING-94*, pp.569–573 (1994).
5) 丸山，荻野：正規文法に基づく日本語形態素解析，情報処理学会論文誌，Vol.35, No.7, pp.1293–1299 (1994).
6) 木谷：固有名詞の特定機能を有する形態素解析処理，情報処理学会自然言語処理研究会，90-10 (1992).
7) Nagata, M.: A Stochastic Japanese Morpho-

logical Analyzer Using a Forward-DP Backward-A^* N-Best Search Algorithm, *COLING-94*, pp.201–207 (1994).
8) 吉村，日高，吉田：文節数最小法を用いたべた書き日本語文の形態素解析，情報処理学会論文誌，Vol.24, No.1, pp.40–46 (1983).
9) 中村，今永，吉田：接続コスト最小法による日本語形態素解析の評価実験，電子情報通信学会言語理解とコミュニケーション研究会，91-24 (1991).
10) 村上，嵯峨山：HMM を用いた形態素解析，情報処理学会第 45 回全国大会，Vol.3, pp.161–162 (1992).
11) 久満，新田：接続コスト最小法による形態素解析の提案と計算量の評価について，電子情報通信学会言語理解とコミュニケーション研究会，90-116 (1990).
12) Morimoto, T., et al.: "A Speech and Language Database for Speech Translation Research", ICSLP-94, pp.1791–1794 (1994).
13) 日本電子化辞書研究所：EDR 電子化辞書仕様説明書，第 9 章，日本電子化辞書研究所，東京 (1995).
14) Black, E., et al.: "A Procedure for Quantitatively Comparing the Syntactic Coverage of English Grammars", *DARPA Speech and Natural Language Workshop*, pp.306–311 (1991).
15) 田代，森元：音声言語処理のための構文解析ツールキット，情報処理学会自然言語処理研究会，106-12 (1995).
16) 市丸，中村，宮本，日高：シソーラスと確率文法による派生語解析，情報処理学会論文誌，Vol.36, No.4 pp.849–858 (1994).

(平成 7 年 7 月 20 日受付)
(平成 7 年 10 月 5 日採録)

田代 敏久（正会員）

昭和 39 年生．平成元年東京大学文学部卒業．同年，(株) CSK に入社．機械翻訳システムの研究開発に従事．平成 3 年より ATR 自動翻訳電話研究所に出向．以来，音声言語翻訳システム，特に，音声言語統合方式，音声言語理解方式などの研究に従事．現在，ATR 音声翻訳通信研究所研究員．

森元 逞（正会員）

昭和 21 年生．昭和 43 年九州大学工学部電子工学科卒業．昭和 45 年同大大学院修士課程終了．同年，日本電信電話公社に入社．以来，同社電気通信研究所にて，オペレーティングシステム等の研究開発に従事．昭和 62 年より，ATR 自動翻訳電話研究所にて，音声言語翻訳システム，特に，音声言語統合方式，音声言語理解方式などの研究に従事．現在，ATR 音声翻訳通信研究所，第 4 研究室室長．電子情報通信学会，人工知能学会，言語処理学会，各会員．