修 士 論 文

意味辞書を利用するための形態素区切り修正規則の自動獲得

北陸先端科学技術大学院大学
情報科学研究科情報処理学専攻

森田 勝

2003 年 3 月

修 士 論 文

意味辞書を利用するための形態素区切り修正規則の自動獲得

指導教官　白井 清昭

審査委員主査　白井清昭 助教授
審査委員　島津明 教授
審査委員　鳥澤健太郎 助教授

北陸先端科学技術大学院大学
情報科学研究科情報処理学専攻

110119 森田 勝

提出年月: 2003 年 2 月

概 要

自然言語処理においては、シソーラスや国語辞典などの意味辞書を用いて解析対象となる文中の形態素の意味クラスや語釈文を調べる機会が多い。また、その前処理として、形態素解析ツールを用いて文を形態素に分割することが一般的である。しかし、形態素解析ツールが出力する形態素と意味辞書中の形態素の表記が一致していなかったり、形態素区切りが一致していないために、意味辞書から意味クラスや語釈文が取り出せないことがある。意味辞書をより効果的に利用するためには、表記や形態素区切りの不一致が生じた際に、それらを修正する必要がある。但し、現在利用可能な形態素解析ツールや意味辞書は複数存在するため、その全ての組み合わせについて人手で修正規則をつくるのは多大な時間と費用がかかる。そこで本研究では、形態素解析ツールの辞書中の形態素と意味辞書中の形態素を照合し、形態素解析ツールの出力を意味辞書での表記や区切りに合わせるように修正する規則を自動的に獲得した。

本研究で獲得する修正規則は次の 2 つである。

1. **表記の不一致を修正する規則**
   異表記などでツールと意味辞書の表記が一致しないときに、これを修正する規則である。「輪なげ」 → 「輪投げ」が例として挙げられる。この規則は、ツールが出力する形態素の表記が「輪なげ」のとき、これを意味辞書での表記「輪投げ」に修正する規則であるし。また、読みだけで意味辞書を検索するナイーブな方法と比べて、取り出されるエントリの数を絞り込む働きをする。本研究ではこれを 1:1 の規則と呼ぶ。
2. **形態素区切りを修正する規則**
   まず、ツールが出力する 1 つの形態素をいくつかに分割して意味辞書での区切りに合わせる規則を獲得する。これを 1:多の規則と呼ぶ。「大量消費」 → 「大量」 + 「消費」が例として挙げられる。この規則は、ツールが「大量消費」という形態素を出力するとき、これを意味辞書にある 2 つの形態素「大量」と「消費」に分割して、それぞれのエントリを取り出すための規則である。また、ツールが出力する複数の形態素を 1 つにまとめて意味辞書での区切りに合わせる規則も獲得する。これを多:1 の規則と呼ぶ。「経済」 + 「成長」 → 「経済成長」が例として挙げられる。この規則は、意味辞書に「経済成長」というエントリがあるとき、ツールが出力する 2 つの形態素「経済」と「成長」を連結して「経済成長」のエントリを取り出すための規則である。

1:1 の規則の獲得は以下のように行う。まず、ツールに登録されている形態素の集合を M=$\{(h_m,y_m,p_m)\}$、意味辞書に登録されている形態素の集合を D=$\{(h_d,y_d,p_d)\}$ とする。ツール及び意味辞書に登録されている形態素は表記 h、読み y、品詞 p の組とする。但し、ツールと意味辞書では一般に品詞体系が異なるので、「名詞」「動詞」のような共通の粗い品詞体系を用意し、それぞれの品詞をこれに合わせることによって両者の差異を吸収する。次に M と D から読みと品詞が一致し、表記がマッチするものを探し、1:1 の規則 $(h_m, y_m, p_m) \rightarrow (h_d, y_d, p_d)$ として推測する。ここで表記がマッチするとは、同じ文字はマッチする、任意のひらがな列は漢字 1 文字とマッチする、という条件の下での DP マッチングに成功することを指す。

一方、1:多の規則の獲得は以下のように行う。まず、固有名詞は分割しても意味がないため、M から固有名詞を除く。次に、M 中の各形態素 (h_m, y_m, p_m) について、次の 4 つの条件を満たす形態素の組を D から探し、1:多の規則 $(h_m, y_m, p_m) \rightarrow (h_{d1}, y_{d1}, p_{d1}) + \cdots + (h_{dn}, y_{dn}, p_{dn})$ として獲得する。1. 表記が一致している $(h_m = h_{d1} \oplus \cdots \oplus h_{dn})$ ここで $\oplus$ は文字列の連結を表わす。2. 品詞が一致している。$(p_m = p_{d1} = \cdots = p_{dn})$3.$h_m$ がひらがな、特殊文字を含まない。4. 規則の右辺の形態素の中に、表記 h_{di} が 2 文字以上のものを必ず含む。また、多:1 の規則の獲得は、M と D を入れ換えて 1:多の規則と同様に行う。ただし、1:多の規則とは異なり、M から固有名詞は除かない。

形態素解析ツールとして JUMAN、茶筌の 2 つ、意味辞書としては岩波国語辞典、分類語彙表、日本語語彙体系、EDR 日本語単語辞書の 4 つ、計 8 通りの組み合わせについて、修正規則を獲得する実験を行った。1:1 の規則は $11,000 \sim 300,000$ 個獲得された。獲得した規則のおよそ $97 \sim 99\%$ は 1:1 の規則として適切であった。一方、獲得された形態素区切りを修正する規則の数は 1:1 の規則に比べて少なく $100 \sim 21,000$ であった。1:多の規則についてはランダムに 50 個選んでその規則が正しいかどうか調べたところ、およそ $80\% \sim 90\%$ の規則が正しかった。また多:1 の規則は、ツールが出力する複数の形態素をまとめて意味辞書での区切りに合わせる規則であるが、このような場合には意味辞書のエントリが常に正しく取り出すことができると考えられる。すなわち獲得した規則は全て正しいとみなした。

次に、毎日新聞の 1997 年の 10,000 記事の形態素解析を行い、獲得した規則を適用し、意味辞書のエントリを取り出すことのできた形態素がどれだけ増加したか調べた。1:1 の規則を用いることによって、意味辞書のエントリを取り出すことができた形態素は $2 \sim 7\%$ 増加した。一方、1:多と多:1 の規則については著しい効果が見られなかった。これは獲得された規則の数が少ないことが原因として考えられる。

目 次

表 目 次

第1章　はじめに

1.1　研究の背景と目的

自然言語処理においては、シソーラスや国語辞典などの意味辞書を用いて解析対象となる文中の形態素の意味クラスや語釈文を調べる機会が多い。また、その前処理として、形態素解析ツールを用いて文を形態素に分割することが一般的である。しかし、形態素解析ツールが出力する形態素と意味辞書中の形態素の表記が一致していなかったり、形態素区切りが一致していないために、意味辞書から意味クラスや語釈文 (以下、これらをまとめて意味辞書のエントリと呼ぶ) が取り出せないことがある。意味辞書をより効果的に利用するためには、表記や形態素区切りの不一致が生じた際に、それらを修正する必要がある。但し、現在利用可能な形態素解析ツールや意味辞書は複数存在するため、その全ての組み合わせについて人手で修正規則をつくるのは多大な時間と費用がかかる。そこで本研究では、形態素解析ツールの辞書中の形態素と意味辞書中の形態素を照合し、形態素解析ツールの出力を意味辞書での表記や区切りに合わせるように修正する規則を自動的に獲得することを目的とする。

本研究では次の 2 つの規則を獲得する。

- **表記の不一致の修正規則**
  異表記などでツールと意味辞書の表記が一致しないときに、これを修正する規則である。例を (1.1) に挙げる。

$$(\text{輪なげ}, \text{わなげ}, \text{名詞}) \quad \rightarrow \quad (\text{輪投げ}, \text{わなげ}, \text{名詞}) \tag{1.1}$$

  この規則は、ツールが出力する形態素の表記が「輪なげ」のとき、これを意味辞書での表記「輪投げ」に修正する規則である。本研究ではこれを 1:1 の規則と呼ぶ。

- **形態素区切りの修正規則**

  1. **1:多の規則**
     まず、ツールが出力する 1 つの形態素をいくつかに分割して意味辞書での区切りに合わせる規則を獲得する。これを 1:多の規則と呼ぶ。例を (1.2) に挙げる。

(大量消費, たいりょうしょうひ, 名詞)

$$\rightarrow \quad (\text{大量}, \text{たいりょう}, \text{名詞}) + (\text{消費}, \text{しょうひ}, \text{名詞}) \quad (1.2)$$

この規則は、ツールが「大量消費」という形態素を出力するとき、これを意味辞書にある 2 つの形態素「大量」と「消費」に分割して、それぞれのエントリを取り出すための規則である。

2. **多:1 の規則**

また、ツールが出力する複数の形態素を 1 つにまとめて意味辞書での区切りに合わせる規則も獲得する。これを多:1 の規則と呼ぶ。例を以下にあげる。

(経済, けいざい, 名詞) + (成長, せいちょう, 名詞)

$$\rightarrow \quad (\text{経済成長}, \text{けいざいせいちょう}, \text{名詞}) \quad (1.3)$$

この規則は、意味辞書に「経済成長」というエントリがあるとき、ツールが出力する 2 つの形態素「経済」と「成長」を連結して「経済成長」のエントリを取り出すための規則である。

1.2 本論文の構成

2 章では、本研究と関連する研究について述べる。3 章では、規則を獲得する手法について述べる。本研究では、表記の不一致を修正する規則、形態素区切りを修正する規則の 2 つを獲得する。また、形態素区切りを修正する規則については 1:多の規則と多:1 の規則を獲得する。4 章では、3 章で述べた手法を用いて 1:1 の規則、1:多の規則、多:1 の規則のそれぞれの規則を獲得する実験、また獲得された規則の評価実験について述べる。5 章では、本研究のまとめと今後の課題について述べる。

第2章　関連研究

本研究は、既存のコーパスの品詞タグ（ソース側の品詞）を別の品詞体系に基づく品詞タグ（ターゲット側の品詞）に変換する研究と関連が深い。ここでは、このような研究をいくつか紹介する。

乾らは、形態素・構文解析器を用いて既存のコーパスを異なる品詞体系に変換するアルゴリズムを提案している [9]。彼らは、品詞タグ付きコーパスの品詞タグをターゲット側の品詞タグに個別に変換する代わりに、ターゲット側の品詞体系に基づく文法、辞書、品詞連接表、文節係り受け表 (文節の係り受けに関する制約) を用意し、これらを用いてコーパスの文を形態素・構文解析することにより、ターゲット側の品詞タグを付与している。通常の形態素・構文解析を行うだけでは曖昧性が多く、正確な解析はできないため、コーパスに付加されたソース側の形態素区切り、品詞情報、文節境界、係り受け情報などを制約とし、これらの制約に違反しないという条件の下で解析を行うことによって曖昧性を抑制している。ソース側の情報を用いる場合と用いない場合とを比較した実験を行い、ソース側の情報を用いることが曖昧性解消に劇的に貢献することを確認している。具体的には京大コーパス [5] の品詞体系を EDR 日本語単語辞書 [6] の品詞体系に変換することを試みている。

下畑らは、異なる体系間で形態素情報を変換する方法について述べている [10]。変換は、語彙化変換と一般変換の 2 つの方法を用いている。語彙化変換は個別の変換対象語ごとの変換規則を利用する方法であり、語境界の変動を伴う変換も可能と書かれている。一般変換は品詞のみを参照する変換規則を利用する方法であり、語彙化変換と比較すると精度はやや低いが、新出語の変換を行うことができるとされている。高頻度の形態素は語彙化変換、それ以外の形態素は一般変換を行う。各変換は決定木を学習することにより行う。全体的には語彙化変換の方が若干よい精度がでている。一般変換では品詞ごとの精度の差が大きいが、語彙化変換では安定していると記されている。また接頭辞、接続詞、助詞などが一般変換において精度が低いが、これらの品詞は、個別の独立性が強い語が多く含まれており、一般変換の単一の決定木ではそれらの語の独自性を十分に表現しきれなかったと書いている。また、この研究では京大コーパスと RWC コーパス [2] を用いて、JUMAN 体系から、THiMCO 体系への変換実験を行っている。

田代らは利用したい品詞体系と異なる体系で作成された形態素情報つ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つきコーパスの再構成手法を提言している [12]。ソース側とターゲット側の両方の品詞体系に基づく品詞タグが付与されたコーパスを訓練データとし、品詞タグや形態素境界を変換する形態素調節規則を獲得し

ている。また、未知語処理では、単語の変化を伴わないものにおいては、品詞情報が一致すれば、適用可能とし、入力の複数の語を１つにまとめあげるもの、および複数の語の分割方法を改めるものにおいて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 語の表記が一致する場合に適用可能としている。そして、入力の語を複数の語に分割するものに対しては未知語の品詞情報と文字列が一致し、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している。この研究では ATR コーパスと EDR コーパス [6] との体系の書き換えを行っている。

これらの研究と本研究は、その目的が少し異なる。相違点をまとめると以下のようになる。

- 本研究は品詞の変換は対象とせず、表記や形態素区切りの変換のみを目的とする。
- 先行研究はコーパスの品詞体系変換を対象とし、コーパスに含まれる全ての単語の品詞を変換することを目的としている。これに対し、本研究は形態素解析ツールに含まれる形態素が変換の対象となる。
- 先行研究は一組の品詞体系についてのみ品詞タグを変換する手法を検討しているのに対し、本研究では任意の形態素解析ツールと意味辞書の組に対して、形態素の変換規則を自動的に獲得できる汎用性のある手法を提案する。具体的には、形態素解析ツールとして JUMAN[4]、茶筌 [1] を、意味辞書として岩波国語辞典 [7]、分類語彙表 [11]、日本語語彙体系 [3]、EDR 日本語単語辞書 [6] を用いた。

本研究では、修正規則を獲得する際に品詞の情報を利用するので、形態素解析ツールと意味辞書の品詞体系の違いが問題となる。したがって形態素解析ツールの品詞体系を意味辞書での品詞体系に変換する、あるいはその逆の変換を行った上で修正規則を獲得することが理想的である。しかし、品詞体系の自動変換は一般に難しく、先行研究でも十分な成果が得られているとは言い難い。そこで本研究では、品詞体系の自動変換を行うのではなく、共通の粗い品詞体系を用意し、2 つの品詞体系を粗い品詞体系に変換することで、形態素解析ツールと意味辞書の品詞体系の違いに対応する。ここでの粗い品詞体系とは、「名詞」、「動詞」など、どの品詞体系からも容易に対応することが可能な品詞体系である。また、修正規則を獲得する際には、このような粒度の粗い品詞体系の情報でも十分利用する価値がある。

第3章　規則の獲得

1 章で述べたように、本研究では以下の 3 つの規則を獲得する。

- 表記の不一致を修正する規則（1:1 の規則）
  この規則の獲得手法は 3.1 節で説明する。
- 形態素区切りの不一致を修正する規則
  - 1:多の規則
    この規則の獲得手法は 3.2.1 項で述べる。
  - 多:1 の規則
    この規則の獲得手法は 3.2.2 項で述べる。

3.1　表記の不一致を修正する規則の獲得

まずツールに登録されている形態素の集合を M、意味辞書に登録されている形態素の集合を D とする。

$$\begin{aligned} M &= \{(h_m, y_m, p_m)\} \\ D &= \{(h_d, y_d, p_d)\} \end{aligned}$$

ここで、ツールに登録されている形態素は表記 h_m、読み y_m、品詞 p_m の組とする。意味辞書の形態素も同様である。但し、ツールと意味辞書では一般に品詞体系が異なるので、「名詞」「動詞」のような共通の粗い品詞体系を用意し、それぞれの品詞をこれに合わせることによって両者の差異を吸収する。具体的な品詞体系変換表は付録 A に記す。

1:1 の規則の一般形を (3.1) に示す

$$(h_m, y_m, p_m) \rightarrow (h_d, y_d, p_d) \tag{3.1}$$

M と D の中から、以下の条件を満たす (h_m,y_m,p_m), (h_d,y_d,p_d) の組を探し、(3.1) の 1:1 の規則として獲得する。

1. 読みが一致している ($y_m = y_d$)

2. 品詞が一致している ($p_m = p_d$)
3. 表記に関する条件
   まず、$h_m \neq h_d$ が規則獲得の条件となる。また、条件 1,2 を満たしても正しい規則が獲得できないことがある。例えば

$$(h_m, y_m, p_m) = (\text{会う}, \text{あう}, \text{動詞}) \tag{3.2}$$
$$(h_d, y_d, p_d) = (\text{合う}, \text{あう}, \text{動詞}) \tag{3.3}$$

   のときには条件 1,2 を満たすが、規則 (3.4) を 1:1 の規則として獲得することは不適切である

$$(\text{会う}, \text{あう}, \text{動詞}) \rightarrow (\text{合う}, \text{あう}, \text{動詞}) \tag{3.4}$$

   一般に h_m と h_d に異なる漢字が含まれるときは規則として不適切な場合が多い。よって表記をチェックする必要がある。表記をチェックする方法として漢字のみのチェックと DP マッチングがある。以下それぞれの処理の概要を述べる。

(a) 漢字のみのチェック
   ここでは 2 つの例を挙げて説明する。１つ目の説明として

$$(h_m, y_m, p_m) = (\text{合い挽き}, \text{あいびき}, \text{名詞}) \tag{3.5}$$
$$(h_d, y_d, p_d) = (\text{あい挽き}, \text{あいびき}, \text{名詞}) \tag{3.6}$$

   をとりあげる。まず形態素解析ツール側の「合い挽き」に含まれる漢字「合」と「挽」を抜き出し集合 A に入れ、意味辞書側の「あい挽き」に含まれる漢字「挽」を抜き出し集合 B に入れる。つまり、集合 A、B はそれぞれ

$$A = \{\, \text{合、挽} \,\} \tag{3.7}$$
$$B = \{\, \text{挽} \,\} \tag{3.8}$$

   である。ここで集合 A と B の関係は B$\subseteq$A である。このように一方の集合にもう一方の集合が包含されるときの組は 1:1 の規則として獲得する。2 つ目の例として、(3.2)、(3.3) の形態素の組に対して同じ手法を適用した場合を考える。集合 A には「会」を、集合 B には「合」を入れる。そこで関係を調べると、B$\not\in$A となる。よってこれは規則として取り入れない。

(b) DP マッチング
   前の手法で規則として獲得されない形態素の組についても、言い換えれば h_m と h_d に異なる漢字が含まれていても、規則として獲得すべき場合もある。例として

$$(h_m, y_m, p_m) = (\text{あい挽き}, \text{あいびき}, \text{名詞}) \quad (3.9)$$
$$(h_d, y_d, p_d) = (\text{合びき}, \text{あいびき}, \text{名詞}) \quad (3.10)$$

をあげる。この時、形態素解析ツール側の漢字「挽」と意味辞書側の漢字「合う」は異なる漢字である。ところが、規則 3.11 は適切な修正規則とみなせる。

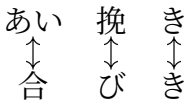
$$(\text{あい挽き}, \text{あいびき}, \text{名詞}) \rightarrow (\text{合びき}, \text{あいびき}, \text{名詞}) \quad (3.11)$$

そこで以下の 2 つの条件の下で DP マッチングを行い、マッチングに成功すれば 1:1 の規則として (3.11) を獲得する。

i. 同じ文字は互いにマッチする

ii. 任意のひらがな列は漢字 1 文字とマッチする

規則 (3.11) の場合、



のように DP マッチングが成功し、条件を満たす。

以上の手法を用いた場合、無駄な規則が獲得されることがある。例えば、以下の 3 つの規則が獲得されたとする。

$$(\text{相うち}, \text{あいうち}, \text{名詞}) \rightarrow (\text{相撃ち}, \text{あいうち}, \text{名詞}) \quad (3.12)$$
$$(\text{相うち}, \text{あいうち}, \text{名詞}) \rightarrow (\text{相打ち}, \text{あいうち}, \text{名詞}) \quad (3.13)$$
$$(\text{相うち}, \text{あいうち}, \text{名詞}) \rightarrow (\text{相討ち}, \text{あいうち}, \text{名詞}) \quad (3.14)$$

ところが、例えば岩波国語辞典では「相撃ち」「相打ち」「相討ち」は全て同じエントリ (「あいうち」の語釈文) を指すものとする。このとき、ツールが出力する形態素「相うち」の表記を規則 (3.12)(3.13)(3.14) を使って 3 通りの表記に修正する必要はなく、どれか 1 つの表記に修正すれば「あいうち」の語釈文を正しく取り出せる。そこで、意味辞書において同じエントリを指す形態素に変換する規則は常に 1 つだけ獲得することにした。

3.2 形態素区切りを修正する規則の獲得

3.2.1 1:多の規則

3.1 項と同様に、ツールに登録されている形態素の集合を M、意味辞書に登録されている形態素の集合を D とする。さらに、D に 1:1 の規則の左辺を追加する。この理由は後述する。また固有名詞を分割しても意味がないため、M から固有名詞を除く。

1:多の規則の一般形を (3.15) に示す。

$$(h_m, y_m, p_m) \rightarrow (h_{d1}, y_{d1}, p_{d1}) + \cdots + (h_{dn}, y_{dn}, p_{dn}) \tag{3.15}$$

M 中の形態素 (h_m, y_m, p_m) について、D の中から以下の条件を満たす形態素の組 (h_{d1}, y_{d1}, p_{d1})、$\cdots$、(h_{dn}, y_{dn}, p_{dn}) を探し、(3.15) の 1:多の規則として獲得する。

1. 表記が一致している $(h_m = h_{d1} \oplus \cdots \oplus h_{dn})$
   但し、$\oplus$ は文字列の連結を表わす。
2. 品詞が一致している $(p_m = p_{d1} = \cdots = p_{dn})$
3. h_m がひらがな、特殊文字を含まない
   h_m がひらがな等を含むときは、以下のような不適切な規則が獲得されることが多かった。

$$\begin{aligned}&(\text{あん黒街}, \text{あんこくがい}, \text{名詞})\\ &\quad\rightarrow\ (\text{あん}, \text{あん}, \text{名詞}) + (\text{黒}, \text{こく}, \text{名詞}) + (\text{街}, \text{がい}, \text{名詞})\end{aligned} \tag{3.16}$$

4. 規則の右辺の形態素の中に、表記 h_{di} が 2 文字以上のものを必ず含む
   ツールの形態素が 1 文字ずつに分割されるとき、以下のように不適切な規則が得られることが多い。

$$\begin{aligned}&(\text{猪武者}, \text{いのししむしゃ}, \text{名詞})\\ &\quad\rightarrow\ (\text{猪}, \text{いのしし}, \text{名詞}) + (\text{武}, \text{む}, \text{名詞}) + (\text{者}, \text{しゃ}, \text{名詞})\end{aligned} \tag{3.17}$$

本研究では 1:多の規則を獲得する際に、ツールの形態素の表記と意味辞書の形態素の表記が一致しているかどうかをチェックしている (条件 1) が、読みについてはチェックしていない。これは、以下のように、読みが一致していなくても正しい規則が獲得できることがあるためである。

連濁の例

$$(\text{寝不足}, \text{ねぶそく}, \text{名詞}) \rightarrow (\text{寝}, \text{ね}, \text{名詞}) + (\text{不足}, \text{ふそく}, \text{名詞}) \tag{3.18}$$

読みが異なる例

$$(\text{二文字}, \text{ふたもじ}, \text{名詞}) \rightarrow (\text{二}, \text{に}, \text{名詞}) + (\text{文字}, \text{もじ}, \text{名詞}) \tag{3.19}$$

ある 1 つの形態素 (h_m, y_m, p_m) について、上記の条件を満たす $h_{d1} \sim h_{dn}$ の組が複数得られることがある。例えば「秋雨前線」を例に挙げる。

(秋雨前線, あきさめぜんせん, 名詞)

$$\rightarrow \quad (\text{秋雨}, \text{あきさめ}, \text{名詞}) + (\text{前線}, \text{ぜんせん}, \text{名詞}) \tag{3.20}$$

$$\rightarrow \quad (\text{秋雨}, \text{あきさめ}, \text{名詞}) + (\text{前}, \text{ぜん}, \text{名詞}) + (\text{線}, \text{せん}, \text{名詞}) \tag{3.21}$$

$$\rightarrow \quad (\text{秋}, \text{あき}, \text{名詞}) + (\text{雨}, \text{あめ}, \text{名詞}) + (\text{前線}, \text{ぜんせん}, \text{名詞}) \tag{3.22}$$

$$\rightarrow \quad (\text{秋}, \text{あき}, \text{名詞}) + (\text{雨}, \text{あめ}, \text{名詞}) + (\text{前}, \text{ぜん}, \text{名詞}) + (\text{線}, \text{せん}, \text{名詞}) \tag{3.23}$$

このとき規則の右辺の形態素列が生成される確率を式 (3.24) で求め、最大の生成確率を持つ形態素の組についてのみ規則を獲得する。

$$\prod_i P(h_{di}) \tag{3.24}$$

式 (3.24) における $\mathrm{P}(h_{di})$ は、形態素 h_{di} の生成確率であり、これは式 (3.25) の文字 bi-gram によって推定する。

$$P(h_{di}) \quad = P(C_1) \prod_{i=2}^{l} P(C_i | C_{i-1}) \tag{3.25}$$

式 (3.25) において $C_1 \cdots C_l$ は h_{di} を構成する文字列である。文字 bi-gram は毎日新聞 1991 年から 1999 年の記事から推定した。

例えば (3.20) $\sim$ (3.23) の生成確率は (3.26) $\sim$ (3.29) のように求める。

$$\begin{aligned}
\text{秋雨} + \text{前線} &= P_w(\text{秋雨}) \times P_w(\text{前線}) \\
&\simeq \{P(\text{秋}) \times P(\text{雨} \,|\, \text{秋})\} \times \{P(\text{前}) \times P(\text{線} \,|\, \text{前})\} && (3.26) \\
\text{秋雨} + \text{前} + \text{線} &\simeq \{P(\text{秋}) \times P(\text{雨} \,|\, \text{秋})\} \times \{P(\text{前})\} \times \{P(\text{線})\} && (3.27) \\
\text{秋} + \text{雨} + \text{前線} &\simeq \{P(\text{秋})\} \times \{P(\text{雨})\} \times \{P(\text{前}) \times P(\text{線} \,|\, \text{前})\} && (3.28) \\
\text{秋} + \text{雨} + \text{前} + \text{線} &\simeq \{P(\text{秋})\} \times \{P(\text{雨})\} \times \{P(\text{前})\} \times \{P(\text{線})\} && (3.29) \\
& && (3.30)
\end{aligned}$$

このうち、式 (3.26) の生成確率が最も大きいので、規則 (3.20) のみを 1:多の規則として獲得する。

M 中の形態素 (h_m, y_m, p_m) に対して、1 $\sim$ 4 の条件を満たす形態素の組 (h_{d1}, y_{d1}, p_{d1}) 、$\cdots$ 、(h_{dn}, y_{dn}, p_{dn}) を D の中から見つけること、及びそれらの組に対して式 (3.24) の生成確率を求める作業は MSLR パーザ [8] を用いて行った。

以上で述べた手法は形態素解析ツールと意味辞書の区切りの違いについては考慮しているか、表記の違いについては考慮していない。例えば、M に (国ない法, こくないほう, 名

詞)、D に (国内, こくない, 名詞)、(法, ほう, 名詞) という形態素があるとき、規則 (3.31) を獲得することが望ましい。

$$(\text{国ない法}, \text{こくないほう}, \text{名詞}) \rightarrow (\text{国内}, \text{こくない}, \text{名詞}) + (\text{法}, \text{ほう}, \text{名詞}) \quad (3.31)$$

ところが、この規則は条件 1、すなわち意味辞書中の複数の形態素を連結した文字列 $h_{d1} \oplus \cdots \oplus h_{dn}$ が形態素解析ツールの形態素の表記 h_m と異なるため、規則として獲得されない。

このような表記のずれは 1:1 の規則として獲得されていることがある。例えば、規則 (3.32) が 1:1 の規則として獲得されていたとする。

$$(\text{国ない}, \text{こくない}, \text{名詞}) \rightarrow (\text{国内}, \text{こくない}, \text{名詞}) \quad (3.32)$$

この規則によって、ツールでは「国ない」という表記が意味辞書では「国内」という表記で表されていることが学習されている。そこで、本項の冒頭でも述べたように、意味辞書の形態素の集合 D に 1:1 の規則の左辺を追加する。これにより、(3.33) の規則が獲得される。

$$(\text{国ない法}, \text{こくないほう}, \text{名詞}) \rightarrow (\text{国ない}, \text{こくない}, \text{名詞}) * + (\text{法}, \text{ほう}, \text{名詞}) \quad (3.33)$$

ここで*のついている形態素は、意味辞書の形態素ではなく、1:1 の規則の左辺に現れた形態素である。このような形態素は、元の 1:1 の規則の右辺に置き換えて、最終的な 1:多の規則とする。この例では、規則 (3.33) の (国ない, こくない, 名詞) を規則 (3.32) の右辺 (国内, こくない, 名詞) に置き換えて、規則 (3.31) を得る。

3.2.2 多:1 の規則

多:1 の規則は、1:多の規則とは逆に、ツールにある複数の形態素を連結して意味辞書にある 1 つの形態素をつくる規則である。多:1 の規則の一般形を (3.34) に示す。

$$(h_{m1}, y_{m1}, p_{m1}) + \cdots + (h_{mn}, y_{mn}, p_{mn}) \rightarrow (h_d, y_d, p_d) \quad (3.34)$$

多:1 の規則の獲得は、M と D を入れ換えて、1:多の規則と同様に行う。まず、今まで同様に M と D を作成する。さらに M に 1:1 の規則の右辺を追加する。理由は後述する。ただし、1:多の規則の場合は M から固有名詞を除いていたが、多:1 の規則の獲得においては固有名詞は除かない。次に、D 中の形態素 (h_d, y_d, p_d) に対して、以下の条件を満たす形態素の組 (h_{m1}, y_{m1}, p_{m1})、$\cdots$、(h_{mn}, y_{mn}, p_{mn}) を M から見つけ (3.34) の多:1 の規則として獲得する。

1. 表記が一致している $(h_{d1} \oplus \cdots \oplus h_{dn} = h_m)$
   但し、$\oplus$ は文字列の連結を表わす。
2. 品詞が一致している $(p_{m1} = \cdots p_{mn} = p_d)$

3. h_d がひらがな、特殊文字を含まない
   h_d がひらがな等を含むときは、以下のような不適切な規則が獲得されることが多かった。

$$\begin{aligned}(\text{決}, \text{き}, \text{名詞}) + (\text{まり}, \text{まり}, \text{名詞}) \\ \rightarrow \quad (\text{決まり}, \text{きまり}, \text{名詞})\end{aligned} \tag{3.35}$$

4. 規則の右辺の形態素の中に、表記 h_{di} が 2 文字以上のものを必ず含む
   ツールの形態素が 1 文字ずつに分割されるとき、以下のように不適切な規則が得られることが多い。

$$\begin{aligned}(\text{蒼}, \text{あお}, \text{名詞}) + (\text{白}, \text{しろ}, \text{名詞}) + (\text{い}, \text{い}, \text{名詞}) \\ \rightarrow \quad (\text{蒼白い}, \text{あおじろい}, \text{名詞})\end{aligned} \tag{3.36}$$

ある 1 つの形態素 (h_m,y_m,p_m) について、上記の条件を満たす $h_{d1} \sim h_{dn}$ の組が複数得られることがある。このとき、規則の右辺の形態素列が生成される確率を式 (3.37) で求め、最大の生成確率を持つ形態素の組についてのみ規則を獲得する。

$$\prod_i P(h_{mi}) \tag{3.37}$$

式 (3.37) における P(h_{mi}) は、形態素 h_{mi} の生成確率であり、これは式 (3.38) の文字 bi-gram によって推定する。

$$P(h_{mi}) \quad = P(C_1) \prod_{i=2}^{l} P(C_i|C_{i-1}) \tag{3.38}$$

式 (3.38) において $C_1 \cdots C_l$ は h_{mi} を構成する文字列である。文字 bi-gram は、1:多の規則と同じように毎日新聞 1991 年から 1999 年の記事から推定した。

以上で述べた手法は形態素解析ツールと意味辞書の区切りの違いについては考慮しているか、表記の違いについては考慮していない。例えば、M に (空き, くうき, 名詞)、(銃, じゅう, 名詞)、D に (空気銃, くうきじゅう, 名詞) という形態素があるとき、規則 (3.39) を獲得することが望ましい。

$$(\text{空き}, \text{くうき}, \text{名詞}) + (\text{銃}, \text{じゅう}, \text{名詞}) \rightarrow (\text{空気銃}, \text{くうきじゅう}, \text{名詞}) \tag{3.39}$$

ところが、この規則は条件 1、すなわちツール中の複数の形態素を連結した文字列 $h_{m1} \oplus \cdots \oplus h_{mn}$ が意味辞書の形態素の表記 h_m と異なるため、規則として獲得されない。

このような表記のずれは 1:1 の規則として獲得されていることがある。例えば、規則 (3.40) が 1:1 の規則として獲得されていたとする。

$$(\text{空き}, \text{くうき}, \text{名詞}) \rightarrow (\text{空気}, \text{くうき}, \text{名詞}) \tag{3.40}$$

この規則によって、ツールでは「空き」という表記が意味辞書では「空気」という表記で表されていることが学習されている。そこで、本項の冒頭でも述べたように、意味辞書の形態素の集合 M に 1:1 の規則の右辺を追加する。これにより、(3.41) の規則が獲得される。

$$(\text{空気}, \text{くうき}, \text{名詞}) * +(\text{銃}, \text{じゅう}, \text{名詞}) \rightarrow (\text{空気銃}, \text{くうきじゅう}, \text{名詞}) \tag{3.41}$$

ここで*のついている形態素は、意味辞書の形態素ではなく、1:1 の規則の右辺に現れた形態素である。このような形態素は、元の 1:1 の規則の左辺に置き換えて、最終的な 1:多の規則とする。この例では、規則 (3.41) の (空気, くうき, 名詞) を規則 (3.40) の左辺 (空き, くうき, 名詞) に置き換えて、規則 (3.39) を得る。

第4章　評価実験

4.1　修正規則の獲得

2 つの形態素解析ツール (JUMAN、茶筌) と 4 つの意味辞書 (岩波国語辞典、分類語彙表、日本語語彙体系、EDR 日本語単語辞書)、計 8 通りの組み合わせについて、3 章で述べた手法により修正規則を獲得する実験を行った。以下、岩波国語辞典を “岩波”、分類語彙表を “分類”、日本語語彙体系を “語彙”、EDR 日本語単語辞書を “EDR” と略記する。

ツールと意味辞書に登録されている形態素数を表 4.1 に示す。1:多の規則を獲得する際にはツールの形態素の集合 M から固有名詞を除いたので、表 4.1 のツールの (　) 内に固有名詞を除いた形態素数も示した。

表 4.2、表 4.3 は、獲得された 1:1 の規則の詳細である。1:1 の規則を獲得する際、表記をチェックする方法として 2 つの方法を用いた。ここで、“漢字のみをチェックする方法” で獲得された規則は “DP マッチングでチェックする方法” でも必ず獲得できる。言い換えれば、1:1 の規則 $(h_m, y_m, p_m) \rightarrow (h_d, y_d, p_d)$ において、規則 (4.1) のように h_m と h_d に異なる漢字が含まれていないときには “漢字のみをチェックする方法” と “DP マッチングでチェックする方法” の両方で獲得できるのに対し、規則 (4.2) のように h_m と h_d に異なる漢字が含まれているときには “DP マッチングでチェックする方法” でしか獲得されない。

$$(\text{合い挽き}, \text{あいびき}, \text{名詞}) \rightarrow (\text{あい挽き}, \text{あいびき}, \text{名詞}) \tag{4.1}$$

$$(\text{あい挽き}, \text{あいびき}, \text{名詞}) \rightarrow (\text{合いびき}, \text{あいびき}, \text{名詞}) \tag{4.2}$$

表 4.2 の上段は “漢字のみをチェックする方法” で獲得した規則数であり、下段は “DP マッチングする方法” で獲得した規則数から “漢字のみをチェックする方法” で獲得した

表 4.1: ツールと辞書の形態素数

	ツール	
	JUMAN	茶筌
形態素数	562,817(531,637)	224,828(91,175)

	意味辞書			
	岩波	分類	語彙	EDR
形態素数	68,385	88,004	457,902	280,465

表 4.2: 1:1 の規則の獲得

	ツール	規則数	正解率
(JUMAN, 岩波)	506,944	138,934 (1,003)	99.45% (76.24%)
(JUMAN, 分類)	516,533	134,714 (3,506)	97.58% (93.14%)
(JUMAN, 語彙)	463,458	176,361 (3,386)	98.08% (85.29%)
(JUMAN,EDR)	381,631	292,104 (14,327)	95.29% (96.04%)
(茶筌, 岩波)	171,158	14,635 (30)	99.87% (63.33%)
(茶筌, 分類)	185,241	18,112 (128)	99.36% (90.38%)
(茶筌, 語彙)	102,408	11,888 (70)	99.53% (80.00%)
(茶筌,EDR)	153,120	21,439 (274)	98.86% (89.09%)

表 4.3: 1:1 の規則の集約率

	規則の総数	集約率	TypeA	集約率	TypeB	集約率
(JUMAN, 岩波)	138,934	69.41%	48,229	97.37%	90,705	60.22%
(JUMAN, 分類)	183,900	73.25%	40,307	96.11%	94,407	66.50%
(JUMAN, 語彙)	176,361	45.76%	63,399	67.10%	112,962	38.82%
(JUMAN,EDR)	292,104	42.89%	82,208	48.80%	209,896	40.95%
(茶筌, 岩波)	14,635	69.32%	7,660	89.03%	6,975	55.76%
(茶筌, 語彙)	18,112	73.81%	5,295	92.59%	12,817	68.11%
(茶筌, 語彙)	11,888	38.00%	4,701	51.02%	7,187	32.56%
(茶筌,EDR)	21,439	62.20%	15,035	91.55%	6,404	35.49%

規則数を引いた数である。一方、“正解率”は獲得された 1:1 の規則がどれだけ正しいかを評価したものである。まず、規則 (4.1) のように h_m と h_d に異なる漢字が含まれていないときには、獲得された規則は全て正しいとみなした。一方、規則 (4.2) のように h_m と h_d に異なる漢字が含まれているときにはランダムに選択した 100 個の規則を人手で調べて正解率を求めた。正しくない規則の例を (4.3) に挙げる。

$$(\text{祝い}, \text{いわい}, \text{名詞}) \rightarrow (\text{いわ井}, \text{いわい}, \text{名詞}) \tag{4.3}$$

表 4.2 の “正解率”の下段は、上記のようにランダムに選択した規則の正解率である。一方、表 4.2 の規則の “正解率”の上段は、全ての 1:1 の規則に対する正解率を式 (4.4) で見積もった値である。

$$\frac{(K_1 - K_2) + K_2 \times P_1}{K_1} \tag{4.4}$$

(4.4) において K_1 は 1:1 の規則の総数、K_2 は h_m と h_d に異なる漢字が含まれている規則の数である。また、P_2 は表 4.3 の下段の正解率である。式 (4.4) の分子の第一項 $(K_1 - K_2)$ は、h_m と h_d に異なる漢字が含まれていない規則の数であり、ここでは全て正しい規則とみなしている。一方第二項 $K_2 \times P_1$ は、h_m と h_d に異なる漢字が含まれている規則のうち、正しい規則の推定規則数である。したがって式 (4.4) は全ての 1:1 の規則に対する正解率となる。

1:1 の規則は表記を修正する規則であるが、ツールと意味辞書の表記の違いに対処する方法としては、意味辞書の中から読みだけが一致するエントリを取り出すことが考えられる。この場合、読みが同じで表記や品詞が全く異なるエントリが取り出されることがある。1:1 の規則の獲得は、3.1 節で述べたように、ツールと意味辞書の表記や品詞のチェックを事前に行うことにより、読みだけで意味辞書を検索する方法と比べて意味辞書から取り出されるエントリの数を絞り込む効果がある。表 4.3 の “集約率”はこの効果を評価したものである。集約率は、1:1 の規則の左辺に含まれる全ての形態素に対する式 (4.5) の値の平均である。

$$\frac{1:1\text{の規則によって得られる意味辞書のエントリ}}{\text{読みが一致する意味辞書のエントリ}} \tag{4.5}$$

ツールと意味辞書の組み合わせにもよるが、およそ 40%~70%程度、検索される形態素の数を絞り込む効果があることがわかる。ここで本研究で獲得した 1:1 の規則を次のように分ける。

TypeA 1:1 の規則の左辺の表記 h_m がひらがなだけの規則

TypeB 1:1 の規則の左辺の表記 h_m がひらがな以外の文字も含む規則

表 4.4: 1:多の規則, 多:1 の規則の獲得

	1:多		多:1
	規則数	正解率	規則数
(JUMAN, 岩波)	17,715	84.31%	4,047
(JUMAN, 分類)	9,777	90.20%	5,308
(JUMAN, 語彙)	20,927	88.24%	13,201
(JUMAN,EDR)	108	88.24%	29,004
(茶筌, 岩波)	1,345	84.62%	1,099
(茶筌, 分類)	1,986	92.31%	3,932
(茶筌, 語彙)	2,079	82.35%	14,255
(茶筌,EDR)	804	82.05%	15,067

TypeA の規則のとき、ツールと意味辞書の形態素の読みと品詞が同じなら表記は常にマッチする。つまり、品詞をチェックすることしか絞り込みの効果がない。これに対し、TypeB の規則は、品詞と表記の両方の条件で検索される単語数が絞り込まれる可能性がある。このことから、TypeA の規則は TypeB の規則と比べて集約率は大きいと予想される。このことを裏付けるために、TypeA、TypeB の規則の集約率を調べて表 4.3 に示した。TypeA の集約率は 51% ～ 97%であるのに対し、TypeB の集約率は 32% ～ 68%となり、TypeA の規則の集約率の方が大きくなった。

獲得された 1:多および多:1 の規則の数を表 4.4 に示す。獲得された規則数は 1:1 の規則に比べて少なかった。また、1:多の規則については、ランダムに 50 個選んでその規則が正しいかどうかを調べた。表 4.4 に示したように、80%~90%の正解率で正しい規則が獲得できたことがわかった。一方、多:1 の規則は、ツールが出力する複数の形態素をまとめて意味辞書での区切りに合わせる規則だが、このような場合には意味辞書のエントリが常に正しく取り出すことができると考えられる。すなわち獲得した規則は全て正しいとみなした。

4.2 修正規則の評価

4.2.1 新聞を用いた実験

毎日新聞の 1997 年の 10,000 記事の形態素解析を行い、獲得した規則を適用し、意味辞書のエントリを取り出すことのできた形態素数がどれだけ増加したかを調べた。結果を表 4.5~ 表 4.11 に示す。表 4.5 は形態素解析後の形態素の数を示している。ここでは形態素解析された形態素の数と、そのうちの「未知語」を除いた評価対象となる形態素の数、さらに自立語の数、そのうちの「未知語」を除いた評価対象となる形態素の数を記した。

表 4.5: 形態素解析した後の形態素数

	形態素（全）	評価対象	形態素（自）	評価対象
JUMAN	3,261,794	3,145,400	1,595,108	1,478,714
茶筌	3,604,029	3,584,907	1,958,500	1,939,378

未知語を除いた理由は、本研究では形態素解析ツールの辞書中にある形態素について規則の獲得を行っているために、ツールの辞書にない未知語は研究の対象外としているためである。

表 4.6 における “辞書” の列は、形態素解析ツールの出力と意味辞書での表記や区切りが一致しているため、修正規則を適用しなくても意味辞書のエントリを取り出せる形態素の割合である。一方、“1:1” の列は、1:1 の規則によって表記を修正することにより、新たに意味辞書のエントリを取り出すことのできた形態素の割合である。また下段は修正規則を適用せず、意味辞書のエントリを取り出すことのできなかった形態素に対する、1:1 の規則を用いることにより意味辞書のエントリを取り出すことのできた形態素の割合である。意味辞書のエントリと同じ表記区切りの形態素の割合はおよそ 40%∼70%であった。本研究で獲得した 1:1 の規則を用いると、その割合はおよそ 1.0%∼5.6% 増加した。また、規則なしで意味辞書からエントリを取り出すことのできない形態素のうち、41% ∼ 71% の形態素が 1:1 の規則により意味辞書からエントリを取り出すことができた。

表 4.3 に示したように、1:1 の規則の集約率は、TypeA と TypeB の規則で大きく異なる。そこで、TypeA と TypeB の規則が実際に使われた回数を調べた。表 4.7 は、その回数と 1:1 の規則全体に対する割合である。1:1 の規則における TypeA の規則の使用回数の割合は 50% ∼ 90%を占める。集約率の大きい TypeA の方がよく使われていることから、意味辞書のエントリを絞り込む効果はあまりないことがわかった。

また、意味辞書を用いる対象は、助詞や助動詞などの付属語よりも、名詞や動詞などの自立語が多いと考えられる。そこで、自立語のみの形態素においても同様に調べる。それを表 4.8 に示す。評価対象となる形態素について、修正規則を使用しなくてもおよそ 60%∼85%の形態素を意味辞書から取り出せる。本研究で獲得した 1:1 の規則を用いると、この割合がおよそ 2.0%∼9.7%増加した。また、規則なしで意味辞書からエントリを取り出すことのできた形態素に対する割合は 61%∼85%であった。さらに、表 4.9 に示したように、1:1 の規則において TypeA の規則が占める割合は 54% ∼ 90% であり、TypeB の規則は 9% ∼ 45%であった。TypeA のほうが TypeB よりも占める割合が多いことがわかる。自立語については全ての形態素を評価対象としたときと比べて、予想通り若干意味辞書から形態素を引き出せる割合が大きかった。

表 4.6: 1:1 の規則の効果（全形態素）

	辞書	割合	1:1	割合
(JUMAN, 岩波)	1,824,558	58.01%	148,309	4.72% (11.23%)
(JUMAN, 分類)	1,792,423	56.99%	97,939	3.11% (7.24%)
(JUMAN, 語彙)	1,366,940	43.46%	84,647	2.69% (4.76%)
(JUMAN,EDR)	2,217,923	70.51%	29,897	0.95% (3.22%)
(茶筌, 岩波)	2,084,393	58.14%	202,231	5.64% (13.48%)
(茶筌, 分類)	1,880,792	52.46%	97,874	2.73% (5.74%)
(茶筌, 語彙)	1,500,362	41.85%	107,765	3.01% (5.17%)
(茶筌,EDR)	2,430,833	67.81%	40,495	1.13% (3.51%)

表 4.7: 使用された 1:1 の規則の内訳（全形態素）

	TypeA	割合	TypeB	割合
(JUMAN, 岩波)	115,101	77.61%	33,208	22.39%
(JUMAN, 分類)	49,366	50.40%	48,573	49.60%
(JUMAN, 語彙)	70,293	83.04%	14,354	16.96%
(JUMAN,EDR)	24,120	80.68%	5,777	19.32%
(茶筌, 岩波)	173,082	85.59%	29,149	14.41%
(茶筌, 分類)	52,676	53.82%	45,198	46.18%
(茶筌, 語彙)	97,283	90.27%	10,482	9.73%
(茶筌,EDR)	33,373	82.41%	7,122	17.59%

表 4.8: 1:1 の規則の効果（自立語）

	辞書	割合	1:1	割合
(JUMAN, 岩波)	916,164	61.96%	143,293	9.69% (25.42%)
(JUMAN, 分類)	1,030,703	69.70%	84,913	5.74% (18.95%)
(JUMAN, 語彙)	1,139,354	77.05%	84,647	5.72% (24.94%)
(JUMAN,EDR)	1,258,967	85.14%	29,849	2.02% (13.58%)
(茶筌, 岩波)	1,199,944	61.87%	179,665	9.26% (24.30%)
(茶筌, 分類)	1,286,126	66.32%	90,720	4.68% (13.89%)
(茶筌, 語彙)	1,428,722	73.67%	107,765	5.56% (21.10%)
(茶筌,EDR)	1,501,716	77.43%	40,312	2.08% (9.21%)

表 4.9: 使用された 1:1 の規則の内訳（自立語）

	typeA	割合	typeB	割合
(JUMAN, 岩波)	110,100	76.84%	33,193	23.16%
(JUMAN, 分類)	84,913	54.08%	38,988	45.92%
(JUMAN, 語彙)	84,647	83.04%	14,354	16.96%
(JUMAN,EDR)	29,849	80.66%	5,773	19.34%
(茶筌, 岩波)	179,665	83.78%	29,145	16.22%
(茶筌, 分類)	90,720	54.95%	40,871	45.05%
(茶筌, 語彙)	107,765	90.27%	10,482	9.73%
(茶筌,EDR)	40,312	82.64%	6,999	17.36%

一方 1:多と多:1 のそれぞれの規則で形態素の区切りを修正することにより、新たに意味辞書のエントリを取り出すことのできた形態素の割合を表 4.10 に示し、またそのうち自立語に対して取り出せた形態素の割合を表 4.11 に示す。表 4.10 から、本研究で獲得した 1:多の規則を使うことにより、およそ 0.01%∼0.81%の形態素を新たに意味辞書から取り出せる。また多:1 の規則を使うことにより、およそ 0.02%∼1.81%増加した。1:1 の規則を用いて行った評価の時と同様に、自立語のみの形態素においても調べた。表 4.11 から、本研究で獲得した 1:多の規則を使うことにより、およそ 0.03%∼1.73%の形態素に対して意味辞書からエントリを取り出せる。また多:1 の規則を使うことにより、およそ 0.04%∼3.35%の形態素について意味辞書からエントリを取り出せる。ここでも自立語については、全形態素で評価したときと比べて、予想通り若干意味辞書から形態素を引き出せる割合が大きかった。また 1:多、多:1 の規則については 1:1 の規則と比べて著しい効果が見られなかった。これは獲得された規則の数が少ないことが一因として考えられる。また、ツールが出力する形態素の区切りと意味辞書での区切りが一致しないことがあまり起こらなかったためかもしれない。この原因については今後調査していきたい。

獲得した変換規則をつかっても意味辞書中のエントリを引き出せなかったものは意味辞書には登録されていない形態素がほとんどであった。具体的には以下のような形態素があった。

- 固有名詞:高浜、リマ、日本など
- 数字:１５、１２、１７など
- 形容詞:順調だ、大急ぎだ、まともだ
- ツールの区切りが間違っているもの
  - 一日千秋の思い／で … の “一日千秋の思い” の形態素

表 4.10: 形態素区切りを修正する規則の効果（全形態素）

	1:多	割合	多:1	割合
(JUMAN, 岩波)	25,545	0.81%	4,503	0.14%
(JUMAN, 分類)	8,889	0.28%	17,256	0.55%
(JUMAN, 語彙)	13,017	0.41%	21,270	0.68%
(JUMAN,EDR)	424	0.01%	27,454	0.87%
(茶筌, 岩波)	9,088	0.25%	720	0.02%
(茶筌, 分類)	6,960	0.19%	15,935	0.44%
(茶筌, 語彙)	5,190	0.14%	65,001	1.81%
(茶筌,EDR)	1,999	0.06%	8,908	0.25%

表 4.11: 形態素区切りを修正する規則の効果（自立語）

	1:多	割合	多:1	割合
(JUMAN, 岩波)	25,545	1.73%	4,503	0.30%
(JUMAN, 分類)	8,889	0.60%	17,256	1.17%
(JUMAN, 語彙)	13,017	0.88%	21,270	1.44%
(JUMAN,EDR)	424	0.03%	27,454	1.86%
(茶筌, 岩波)	9,088	0.47%	720	0.04%
(茶筌, 分類)	6,960	0.36%	15,935	0.82%
(茶筌, 語彙)	5,190	0.27%	65,001	3.35%
(茶筌,EDR)	1,999	0.10%	8,908	0.46%

表 4.12: 形態素解析した後の形態素数

	形態素（全）	評価対象	形態素（自）	評価対象
JUMAN	243,747	240,950	107,642	104,845
茶筌	276,966	274,035	125,551	122,620

4.2.2 小説を用いた実験

芥川龍之介の短編小説 50 編の形態素解析を行い、獲得した規則を適用し、意味辞書のエントリを取り出すことのできた形態素数がどれだけ増加したかを調べた。結果を表 4.12 ～ 表 4.18 に示す。JUMAN を用いて形態素解析を行った場合、また、茶筌を用いた場合の形態素数を表 4.12 に示す。ここでは形態素解析された形態素の数と、そのうちの「未知語」を除いた評価対象となる形態素の数、さらに自立語の数、そのうちの未知語を除いた評価対象となる形態素の数を書いた。

表 4.13 ～ 表 4.18 の内容は、新聞記事での実験結果の表 4.6 ～ 表 4.11 と同じである。表 4.13 から、1:1 の規則については、修正規則を使用しないで意味辞書のエントリを取り出すことのできた形態素の割合はおよそ 35%∼80%であった。本研究で獲得した 1:1 の規則を用いると、この割合がおよそ 2.0%∼11.0% 増加した。さらに、表 4.13 から、1:1 の規則のうち TypeA の規則が用いられた割合は 36% ～ 89%であった。またほとんどのツールの辞書の組において、TypeA のほうが TypeB よりも使用される割合が大きいことがわかる。自立語に対する結果を表 4.15、4.16 に示す。評価対象となる形態素について、修正規則を適用しなくてもおよそ 70%∼90%の形態素を意味辞書から取り出せる。本研究で獲得した 1:1 の規則を用いると、この割合がおよそ 3.4%∼12.5%増加した。さらに、表 4.13 から、1:1 の規則のうち TypeA の規則が用いられた割合は 34% ～ 86%であった。ほとんどのツールと辞書の組において、TypeA のほうが TypeB よりも使用される割合が大きい。自立語については、全形態素での評価に比べて、予想通り若干意味辞書から形態素を引き出せる数が多かった。

一方、1:多と多:1 のそれぞれの規則で形態素の区切りを修正することにより、新たに意味辞書のエントリを取り出すことのできた形態素の割合を表 4.17 に示し、またそのうち自立語に対して取り出せた形態素の割合を表 4.18 に示す。評価対象となる形態素において、表 4.17 から、本研究で獲得した 1:多の規則を使うことにより 0.00%∼0.19%の形態素を意味辞書から取り出せる。また多:1 の規則を使うことにより、およそ 0.01%∼0.43%の形態素に対して意味辞書のエントリを取り出せた。自立語のみを評価対象としたとき、表 4.18 から本研究で獲得した 1:多の規則を使うことにより 0.00%∼0.45%の形態素について意味辞書からエントリを取り出せる。また多:1 の規則を使うことにより、0.02%∼0.96%の形態素に対して意味辞書のエントリを取り出せる。ここでも自立語については、全形態素を評価対象としたときと比べて、予想通り若干意味辞書から形態素を引き出せる数が多かった。また 1:多、多:1 の規則については、毎日新聞で行ったとき同様、著しい効果が見

表 4.13: 1:1 の規則の効果（全形態素）

	辞書	割合	1:1	割合
(JUMAN, 岩波)	159,480	66.19%	17,964	7.46% (22.05%)
(JUMAN, 分類)	144,596	60.01%	18,811	7.81% (19.52%)
(JUMAN, 語彙)	83,384	34.61%	11,867	4.93% (7.53%)
(JUMAN,EDR)	193,045	80.12%	4,748	1.97% (9.91%)
(茶筌, 岩波)	194,464	70.96%	28,768	10.50% (36.15%)
(茶筌, 分類)	162,467	59.29%	18,685	6.82% (16.75%)
(茶筌, 語彙)	94,934	34.64%	15,806	5.77% (8.83%)
(茶筌,EDR)	223,925	81.71%	6,063	2.21% (12.10%)

表 4.14: 使用された 1:1 の規則の内訳（全形態素）

	TypeA	割合	TypeB	割合
(JUMAN, 岩波)	13,519	75.26%	4,445	24.74%
(JUMAN, 分類)	6,808	36.19%	12,003	63.81%
(JUMAN, 語彙)	8,231	69.36%	3,636	30.64%
(JUMAN,EDR)	3,701	77.95%	1,047	22.05%
(茶筌, 岩波)	25,296	87.93%	3,472	12.07%
(茶筌, 分類)	7,298	39.06%	11,387	60.94%
(茶筌, 語彙)	12,998	82.23%	2,808	17.77%
(茶筌,EDR)	5,144	84.84%	919	15.16%

表 4.15: 1:1 の規則の効果（自立語）

	辞書	割合	1:1	割合
(JUMAN, 岩波)	71,939	68.61%	17,043	16.26% (31.79%)
(JUMAN, 分類)	71,685	68.37%	17,824	17.00% (33.01%)
(JUMAN, 語彙)	73,013	69.64%	11,867	11.32% (22.59%)
(JUMAN,EDR)	98,353	93.81%	4,683	4.47% (17.22%)
(茶筌, 岩波)	89,666	73.13%	23,721	19.35% (71.98%)
(茶筌, 分類)	92,992	75.84%	18,286	14.91% (61.72%)
(茶筌, 語彙)	90,902	74.13%	15,806	12.89% (49.83%)
(茶筌,EDR)	112,410	91.67%	5,979	4.88% (58.56%)

表 4.16: 使用された 1:1 の規則の内訳（自立語）

	TypeA	割合	TypeB	割合
(JUMAN, 岩波)	12,633	74.12%	4,410	25.88%
(JUMAN, 分類)	6,138	34,44%	11,686	65,56%
(JUMAN, 語彙)	8,231	69.36%	3,636	30,64%
(JUMAN,EDR)	3,666	69.36%	1,017	30.64%
(茶筌, 岩波)	20,265	85.43%	3,456	14.57%
(茶筌, 分類)	7,071	38.67%	11,215	61.33%
(茶筌, 語彙)	12,998	82.23%	2,808	17.77%
(茶筌,EDR)	5,118	85.60%	861	14.40%

表 4.17: 形態素区切りを修正する規則の効果（全形態素）

	1:多	割合	多:1	割合
(JUMAN, 岩波)	472	0.20%	52	0.02%
(JUMAN, 分類)	185	0.18%	178	0.17%
(JUMAN, 語彙)	455	0.19%	205	0.09%
(JUMAN,EDR)	4	0.00%	154	0.06%
(茶筌, 岩波)	127	0.05%	20	0.01%
(茶筌, 分類)	93	0.03%	186	0.07%
(茶筌, 語彙)	116	0.04%	1,178	0.43%
(茶筌,EDR)	2	0.00%	145	0.05%

られなかった。

最後に獲得した変換規則で意味辞書のエントリを引き出すことの出来なかった形態素を示す。4.2.1 項での実験と同様に、意味辞書に登録されていない形態素がほとんどであった。

- 動詞:遇わす
- 形容詞:あらたかだ
- 固有名詞:長崎
- 副詞:時々
- 名詞:精進
- ツールの区切りの間違い
  - 寛／永／か ••• の “寛” の形態素
  - その頃

新聞記事と小説での評価を比較すると、芥川龍之介の小説のほうが毎日新聞に比べ、1:1 の規則は多く使われ、1:多や多:1 の規則の効果は小さかった。このように、修正規則の効果は解析対象となるテキストのジャンルによって異なる。

表 4.18: 形態素区切りを修正するための規則の効果（自立語）

	1:多	割合	多:1	割合
(JUMAN, 岩波)	472	0.45%	52	0.05%
(JUMAN, 分類)	185	0.08%	178	0.17%
(JUMAN, 語彙)	455	0.43%	205	0.20%
(JUMAN,EDR)	4	0.00%	154	0.15%
(茶筌, 岩波)	127	0.10%	20	0.02%
(茶筌, 分類)	93	0.08%	186	0.15%
(茶筌, 語彙)	116	0.09%	1178	0.96%
(茶筌,EDR)	2	0.00%	145	0.12%

第5章　おわりに

本研究では意味辞書を効率よく利用するために、形態素解析ツールの表記や区切りを意味辞書での表記や区切りに修正する規則を獲得する手法を提案した。表記や区切りの修正を行う手法はいろいろあるが、本研究では解析前にできる処理を事前に行い、その結果を修正規則として表現することにより、解析時の処理の負担を軽減できる点に特徴がある。

今後の課題としては、1:多や多:1 の規則を獲得する手法の改良が挙げられる。本研究では、規則を獲得する際にツールと意味辞書の形態素の品詞が同じであることを前提としていたが、規則 (5.1) のように品詞の異なる形態素を連結しても正しい修正規則が得られる場合もある。

$$\begin{aligned}&(\text{阿呆}, \text{あほ}, \text{名詞}) + (\text{くさい}, \text{くさい}, \text{形容詞})\\ \rightarrow\quad &(\text{阿呆くさい}, \text{あほくさい}, \text{形容詞})\end{aligned} \tag{5.1}$$

しかし、全ての品詞の組を連結可能とすると不適切な規則が誤って獲得されることがある。したがって連結可能な品詞の組を良く検討する必要がある。また形態素解析ツールが未知語処理を行い、未知語として形態素を出力する時がある。ところが本研究ではツールに登録されている形態素のみを処理の対象としているため、ツールに登録されていない形態素については対処できない。未知語への対応も重要な課題の一つである。

そして、1:多と多:1 の規則において、左辺の形態素にひらがなが 1 字でも含まれる場合は誤った規則を獲得する傾向が多々みられたので、規則の獲得を行わなかった。しかし、ひらがなを含む形態素はツール内のエントリや意味辞書内のエントリに多く含まれている。したがって、これらを対象に規則を獲得する手法を考案することでさらなる規則の獲得を望める。

1:多、多:1 の規則は名詞に対する規則が大半を占めており、動詞に対する規則にはほとんど獲得されなかった。例えば動詞については、(5.2) のような規則を獲得するべきである。

$$(\text{走り}, \text{はしり}, \text{動詞}) + (\text{回る}, \text{まわる}, \text{動詞}) \rightarrow (\text{走り回る}, \text{はしりまわる}, \text{動詞}) \tag{5.2}$$

ところが、本研究は形態素は全て基本形で取り扱っているので、(5.2) のように連用形と終止形を連結する規則は獲得される。そこで、動詞や形容詞については連用形の表記も考慮する必要がある。

謝辞

本研究を進めるにあたり，終始熱心な御指導を賜わりました白井清昭助教授に心から感謝致します．さらに数多くの御教授を頂きました島津明教授に厚く御礼申し上げます．山田寛康助手ならびに東京外国語大学 外国語学部 望月 源講師、自然言語処理学講座の皆様には，貴重な御意見、討論をして頂きました事を感謝致します．

最後に，本研究を行うにあたり、励まし、また経済的援助をして頂いた両親をはじめ親族にこころから深く感謝致します。

参考文献

[1] http://chasen.aist-nara.ac.jp/index.html.ja.

[2] Koiti Hasida, Hitoshi Isahara, Takenobu Tokunaga, Minako Hashimoto, Shiho Ogino, Wakako Kashino, Jun Toyoura, and Hironobu Takahashi. The RWC text databases. In *Proceedings of the first International Conference on Language Resources and Evaluation*, pp. 457–462, 1998.

[3] 池原悟, 宮崎正弘, 白井諭, 横尾昭男, 中岩浩己, 小倉健太郎, 大山芳史, 林良彦. 日本語語彙体系 — 全 5 巻 —. 岩波書店, 1997.

[4] http://www-nagao.kuee.kyoto-u.ac.jp/nl-resource/juman.html.

[5] 黒橋禎夫, 長尾眞. 京都大学テキストコーパス・プロジェクト. 人工知能学会全国大会論文集, pp. 58–61, 1997.

[6] 日本電子化辞書研究所. EDR 電子化辞書仕様説明書第 2 版. Technical Report TR–045, 1995.

[7] 西尾実, 岩淵悦太郎, 水谷静夫. 岩波国語辞典 第五版. 岩波書店, 1994.

[8] 白井清昭, 植木正裕, 橋本泰一, 徳永健伸, 田中穂積. 自然言語解析のための MSLR パーザ・ツールキット. 自然言語処理, Vol. 7, No. 5, pp. 93–112, 11 2000.

[9] 乾健太郎, 脇川浩和. 品詞タグつきコーパスにおける品詞体系の変換. 情報処理学会自然言語処理研究会, Vol. 132, No. 12, pp. 87–94, 1999.

[10] 下畑光夫, 隅田英一郎. 形態素体系間の情報変換手法. 情報処理学会自然言語処理研究会, Vol. 141, No. 25, pp. 157–162, 2001.

[11] 国立国語研究所. 分類語彙表, 増補版, 1996.

[12] 田代敏久, 森元逞. 形態素情報付きコーパスの再構成手法. 情報処理学論文誌, Vol. 37, No. 1, pp. 18–22, 1996.

付 録A　品詞体系変換表

本研究で用いた品詞体系の一覧を以下に挙げる。左辺が元の品詞で右辺が変換後の粗い品詞である。

岩波国語辞典

以下のように品詞に変換規則を行う。例えば「愛育」の品詞は “名・ス他”であるが、以下にある規則 “名・ス他 → 名詞”から、品詞を “名詞”に変換する。

カ変自 → 動詞
サ変自 → 動詞
サ変自他 → 動詞
サ変他 → 動詞
サ変他自 → 動詞
ス自 → 動詞
ス他 → 動詞
ダ<047>→ 形容動詞
ダ<047>・ス自 → 形容動詞
ダ<047>・副 → 形容動詞
ダナ → 形容動詞
ダナ・ス自 → 形容動詞
ダナ・ト　タル → 形容動詞
ト　ス自 → 形容動詞
ト　タル → 形容動詞
ト　タル・ス自 → 形容動詞
ニ　ナル → 副詞
ラ変自 → 動詞
下一自 → 動詞
下一自他 → 動詞
下一他 → 動詞
下一他自 → 動詞
下二他 → 動詞
格助 → 助詞
感 → 感動詞
間助 → 助詞
係助 → 助詞
形 → 形容詞
五自 → 動詞
五自他 → 動詞
五他 → 動詞
五他自 → 動詞
四自 → 動詞
終助 → 助詞
助 → 助詞
助動 → 助動詞
上一自 → 動詞
上一自他 → 動詞
上一他 → 動詞
上一他自 → 動詞
接 → 接続詞
接助 → 助詞
接助・接 → 助詞
接頭 → 接頭語
接尾 → 接尾語
造・ダ<047>→ 形容動詞

造・ダナ → 形容動詞
造・代 → 名詞
造・副 → 副詞
造・名 → 名詞
代 → 名詞
副 → 副詞
副<038>→ 副詞
副・ス自 → 副詞
副・ダ<047>→ 副詞
副・ダナ → 副詞
副・ト　タル → 副詞
副・感 → 副詞
副・接 → 副詞
副・名 → 副詞
副助 → 助詞
名 → 名詞
名<038>→ 名詞
名<038>・ス自 → 名詞
名<038>・ス自他 → 名詞
名<038>・ス他 → 名詞
名<038>・ト　タル → 名詞
名<038>・副 → 名詞
名・ス自 → 名詞
名・ス自・ダ<047>→ 名詞
名・ス自・ダナ → 名詞
名・ス自他 → 名詞
名・ス他 → 名詞
名・ス他・ダナ → 名詞
名・ス他自 → 名詞
名・ダ<047>→ 名詞
名・ダナ → 名詞
名・ダナ・ス自 → 名詞
名・ダナ・ス他 → 名詞
名・ト　タル → 名詞
名・代 → 名詞
名・副 → 名詞
名・副・ダナ → 名詞
名・副助 → 名詞
連語 → その他
連体 → 連体詞
連体・ダナ → 連体詞

分類語彙表

分類コード表の最初の数字で品詞分けを行う。

- 1 の場合「体の類」
- 2 の場合「用の類」
- 3 の場合「相の類」
- 4 の場合「その他」

例えば「こちら」の分類コードは “11000 1 1 4”である。そこで、分類コードの最初の数字は “1”であることより、「こちら」の品詞は「体の類」とする。

日本語語彙体系

意味クラスを表わす ID の桁数から以下の 3 つに分ける。

- 4 桁の場合「名詞」
- 3 桁の場合「名詞」
- 2 桁の場合「用言」

例えば「あーさき」というエントリは ID が 049 で、桁数は 3 桁であるので、品詞は「名詞」とする。

EDR

以下のような品詞の変換規則を行う。例えば「あ」の品詞は “JIT”であるので、変換規則 JIT → 名詞から、品詞を「名詞」に書き換える。

JN1 → 名詞
JN2 → 名詞
JN3 → 名詞
JN4 → 名詞
JN7 → 名詞
JVE → 動詞
JAJ → 形容詞
JAM → 形容動詞
JD1 → 副詞
JD2 → 副詞
JNM → 連体詞
JC1 → 接続詞
JC3 → 接続詞
JT1 → 接頭語
JT2 → 接頭語
JT3 → 接頭語
JT4 → 接頭語
JN5 → 名詞
JB1 → 接尾語
JUN → 名詞
JN6 → 名詞
JEV → 動詞
JEA → 形容詞
JEM → 形容動詞
JJO → 助詞
JJ1 → 助詞
JJD → 助動詞
JJP → 助動詞
JAX → 動詞
JIT → 名詞
JSY → 名詞

JUMAN

JUMAN の品詞体系は以下のように階層構造をもつ。

名詞　　普通名詞
動詞　　五段活用　　連用形

このうち階層が最も上の品詞を用いる。これにより、以下の品詞からなる品詞体系に変換できる。

動詞	接続詞
形容動詞	接頭語
副詞	接尾語
助詞	名詞
感動詞	その他
形容詞	連体詞

例えば「3秒ルール」という形態素があるが、この形態素の品詞は「名詞 普通名詞」である。このとき品詞は「名詞」にする。

さらに意味辞書として分類語彙表を用いて規則を獲得するとき、

- 名詞 → 体の相
- 動詞 → 用の相
- 形容詞 → 相の類
- 指示詞 → 相の類
- 形容動詞 → 相の類
- その他の品詞は全て “その他”

のように品詞を変換する。また意味辞書として日本語語彙体系を用いて規則を獲得するときは、

- 名詞 → 名詞
- 動詞 → 用言
- 形容詞 → 用言
- 形容動詞 → 用言

のように品詞を変換する。

茶筌

茶筌の品詞体系は以下のように階層構造をもつ。

名詞　　一般
動詞　　自立　　五段・マ行　　未然形

このうち階層が最も上の品詞を用いる。これにより以下の品詞からなる品詞体系に変換できる。

動詞	接続詞
形容動詞	接頭語
副詞	接尾語
助詞	名詞
感動詞	その他
形容詞	連体詞

例えば「お調子者」は品詞が「名詞 一般」である。このとき品詞は「名詞」とする。さらに意味辞書として分類語彙表を用いて規則を獲得するとき、

- 名詞 → 体の相
- 動詞 → 用の相
- 形容詞 → 相の類
- 指示詞 → 相の類
- 形容動詞 → 相の類
- その他の品詞は全て “その他”

のように品詞を変換する。また意味辞書として日本語語彙体系を用いて規則を獲得するときは、

- 名詞 → 名詞
- 動詞 → 用言
- 形容詞 → 用言
- 形容動詞 → 用言

のように品詞を変換する。