意味辞書を利用するための形態素変換規則の自動獲得

森田 勝　　白井 清昭
北陸先端科学技術大学院大学 情報科学研究科

1 はじめに

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。

本研究は、ある品詞体系に基づく品詞タグつきコーパスの品詞タグを別の品詞体系に基づく品詞タグに変換する研究 [6][7][9] と関連が深い。但し、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。また、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[3]、茶筌 [1] を、意味辞書として岩波国語辞典 [5]、分類語彙表 [8]、日本語語彙体系 [2]、EDR 日本語単語辞書 [4] を用いた。

2 提案手法

本研究では次の2つの規則を獲得する。

- 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。例を (1) に挙げる。

  (輪なげ, わなげ, 名詞) → (輪投げ, わなげ, 名詞)　(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 の規則と呼び、2.1 項で獲得方法を述べる。
- 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。例を (2) に挙げる。

  (大量消費, たいりょうしょうひ, 名詞) →
  (大量, たいりょう, 名詞) + (消費, しょうひ, 名詞)　(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。

  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) →
  (経済成長, けいざいせいちょう, 名詞)　(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:多の規則の獲得方法は 2.2.1 に、多:1 の規則の獲得方法は 2.2.2 に述べる。

2.1 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。

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

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。

1:1 の規則の一般形を (4) に示す

$$(h_m, y_m, p_m) \rightarrow (h_d, y_d, p_d) \quad (4)$$

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、(4) 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)=(会う, あう, 動詞)、(h_d,y_d,p_d)=(合う, あう, 動詞) のときには条件 1,2 を満たすが、規則 (5) を 1:1 の規則として獲得することは不適切である

   (会う, あう, 動詞) → (合う, あう, 動詞) (5)

   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。ところが、異なる漢字が含まれていても規則として獲得できる場合もある。例えば (h_m,y_m,p_m)=(あい挽き, あいびき, 名詞)、(h_d,y_d,p_d)= (合びき, あいびき, 名詞) のとき、(6) は適切な規則である。

   (あい挽き, あいびき, 名詞) → (合びき, あいびき, 名詞) (6)

   そこで、同じ文字は互いにマッチする、任意のひらがな列は漢字 1 文字とマッチする、という 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規則として獲得する。規則 (6) の場合、

   あい 挽 き
   ↕ ↕ ↕
   合 び き

   のように DP マッチングが成功し、条件を満たす。

こ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。

(相うち, あいうち, 名詞) → (相撃ち, あいうち, 名詞) (7)
(相うち, あいうち, 名詞) → (相打ち, あいうち, 名詞) (8)
(相うち, あいうち, 名詞) → (相討ち, あいうち, 名詞) (9)

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 (7)(8)(9) 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。

2.2 形態素区切りを修正する規則の獲得

2.2.1 1:多の規則

2.1 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。また、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。1:多の規則の一般形を (10) に示す。

$$(h_m, y_m, p_m) \rightarrow (h_{d1}, y_{d1}, p_{d1}) + \cdots + (h_{dn}, y_{dn}, p_{dn}) \quad (10)$$

M と D の中から以下の条件を満たす形態素の組を探し、(10) 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。

   (あん黒街, あんこくがい, 名詞) →
   (安, あん, 名詞) + (黒, こく, 名詞) + (街, がい, 名詞) (11)
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。

   (猪武者, いのししむしゃ, 名詞) →
   (猪, いのしし, 名詞) + (武, む, 名詞) + (者, しゃ, 名詞) (12)

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 (13) 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。

$$\prod_i P(h_{di}) \quad (13)$$

式 (13) における $P(h_{di})$ は、形態素 h_{di} の生成確率であり、これは式 (14) の文字 bi-gram によって推定

表 1: ツールと辞書の形態素数

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

	意味辞書			
	I	B	N	E
形態素数	68,385	88,004	457,902	280,465

する。

$$P(h_{di}) \quad = P(C_1)\prod_{i=2}^{l} P(C_i|C_{i-1}) \qquad (14)$$

式 (14) 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。

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。

(寝不足, ねぶそく, 名詞) →
(寝, ね, 名詞) + (不足, ふそく, 名詞) (15)

(二文字, ふたもじ, 名詞) →
(二, に, 名詞) + (文字, もじ, 名詞) (16)

2.2.2 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 (17) に示す。

$$(h_{d1}, y_{d1}, p_{d1})+\cdots+(h_{dn}, y_{dn}, p_{dn}) \rightarrow (h_m, y_m, p_m) \quad (17)$$

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。

3 評価実験

1 節に挙げた 2 つの形態素解析ツールと 4 つの意味辞書、計 8 通りの組み合わせについて、2 節で述べた手法により変換規則を獲得する実験を行った。以下、JUMAN を “J”、茶筌を “C”、岩波国語辞典を “I”、分類語彙表を “B”、日本語語彙体系を “N”、EDR 日本語単語辞書を “E” 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 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 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 2: 1:1 の規則の獲得

	ツール	規則数	正解率	集約率
(J,I)	506,944	138,934 (1,003)	99.45% (76.24%)	69.41%
(J,B)	516,533	134,714 (3,506)	97.58% (93.14%)	73.25%
(J,N)	463,458	176,361 (3,686)	98.08% (85.29%)	45.76%
(J,E)	381,631	292,104 (14,327)	95.29% (96.04%)	42.89%
(C,I)	171,158	14,635 (30)	99.87% (63.33%)	69.32%
(C,B)	185,241	18,112 (128)	99.36% (90.38%)	73.81%
(C,N)	102,408	11,888 (70)	99.53% (80.00%)	38.00%
(C,E)	153,120	21,439 (274)	98.86% (89.09%)	62.20%

表 2 は、獲得された 1:1 の規則の詳細である。1:1 の規則のうち、h_m と h_d に異なる漢字が含まれていないときは全て正しい規則であるとみなした。また、異なる漢字が含まれる規則については、それらが正しい規則かどうかをランダムに選んで人手で調べた。表 2 の下段は、h_m と h_d に異なる漢字が含まれている規則数とその正解率 (正しい規則が獲得された割合) を表している。また、上段の正解率は、それ以外の規則は全て正しいとみなしたときの正解率である。

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、読みが同じで表記が全く異なるエントリや、品詞が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、2.1 項で述べたように、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 2 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 (18) の値の平均である。

$$\frac{1:1\text{の規則によって得られる意味辞書のエントリ}}{\text{読みが一致する意味辞書のエントリ}} \qquad (18)$$

表 2 から、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。

獲得された 1:多および多:1 の規則の数を表 3 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

表 3: 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
(J,I)	17,667	84.31%	4,047
(J,B)	7,655	90.20%	5,308
(J,N)	18,350	88.24%	13,201
(J,E)	92	88.24%	29,004
(C,I)	1,345	84.62%	1,099
(C,B)	1,096	92.31%	3,932
(C,N)	1,534	82.35%	14,255
(C,E)	193	82.05%	15,067

表 4: 新聞記事への適用実験

	辞書	1:1	1:多	多:1
(J,I)	61.96%	9.69%	1.73%	0.30%
(J,B)	69.70%	5.74%	0.60%	1.17%
(J,N)	77.05%	5.72%	0.88%	1.44%
(J,E)	85.14%	2.02%	0.03%	1.86%
(C,I)	61.87%	9.26%	0.47%	0.04%
(C,B)	66.32%	4.68%	0.36%	0.82%
(C,N)	73.67%	5.56%	0.27%	3.35%
(C,E)	77.43%	2.08%	0.10%	0.46%

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 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。

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数がどれだけ増加したかを調べた。結果を表 4 に示す。JUMAN を用いて形態素解析を行った場合、未知語を除く自立語の数は 1,478,714 であった。また、茶筌を用いた場合は 1,939,378 であった。表 4 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、変換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”、“1:多”、“多:1” は、それぞれの規則によって表記や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である。1:1 の規則については、意味辞書のエントリを取り出すことのできた形態素の数はおよそ 2~10%増加した。これに対し、1:多、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。

4 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、その結果を変換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 (19) のように品詞の異なる形態素を連結しても正しい変換規則が得られる場合もある。

(阿呆, あほ, 名詞) + (くさい, くさい, 形容詞) → (阿呆くさい, あほくさい, 形容詞) (19)

しかし、全ての品詞の組を連結可能とすれば不適切な規則が誤って獲得されることが多くなると予想されることから、連結可能な品詞の組をよ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ときがある。しかし、本研究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。未知語へ対応も重要な課題の一つである。

参考文献

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

[2] 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.

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

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 国立国語研究所. 分類語彙表, 増補版, 1996.

[9] 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.