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 (110119)

北陸先端科学技術大学院大学 情報科学研究科

2003 年 2 月 14 日

キーワード: 形態素解析ツール, 意味辞書, 形態素変換規則, 表記, 形態素区切り.

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。