共起情報を用いた同表記異義の処理

1N－8

藤崎 博也[1] 阿部 賢司[1] 片見 憲次[1] 武田 和也[1] 白井 克彦[2]

[1] 東京理科大学 [2] 早稲田大学

1. はじめに

概念が異なる複数の語が表記のレベルで縮退する現象を我々は同表記異義の現象と呼ぶ。この現象は語の誤認識をもたらし、特に、情報検索においては不要な検索の要因ともなる [1-3]。したがって、同表記異義の関係にある語の概念を特定することの必要性は極めて高い。このような観点から、我々は、まず同表記異義の現象を定量的に把握することを目的とし、特に、学術情報検索における同表記異義の現象に着目して、学術論文のキーワードから同表記異義の実例を収集した。

本報では、これらの同表記異義を処理するための一方法として、予め概念を特定したキーワードを含む論文を学習サンプルとして用い、その共起情報から、同表記異義の関係にあるキーワードの概念を推定する方法について検討した結果を述べる。

2. 同表記異義の実例の収集と分類

同表記異義の現象を定量的に把握するため、学術情報センターから提供される情報検索システム評価用テストコレクション 1 [4] に含まれる学術論文の日本語キーワード (英略語も含む) の中から同表記異義の実例を収集した。その結果、その全キーワード 132,464 語のうち 0.36%のキーワードに同表記異義の現象が存在することを確認した。また、それらのキーワードを、表記に着目して以下の 4 種類に分類し、それらの出現頻度を調べた結果を表 1に示す。

(1) 英略語：例．PC
　概念 1：Personal Computer
　概念 2：Programmable Controller
(2) 平仮名：例．あそび
　概念 1：遊戯 (play)
　概念 2：機械のゆとり (clearance)
(3) 片仮名：例．アルバイト
　概念 1：曹長石 (albite)
　概念 2：アルバイト (part-time work)
(4) 漢字：例．株
　概念 1：切り株 (stump)
　概念 2:(会社の資本の一部としての) 株式 (stock)

表 1 同表記異義の分類と出現率

	出現率 [%]
(1) 英略語	90.0
(2) 平仮名	1.2
(3) 片仮名	4.0
(4) 漢字	4.8

表 1からも明らかなように、同表記異義の現象は英略語に最も多くみられる。このことから、学術論文に関しては英略語の同表記異義を処理する必要性が最も高いといえる。

3. 同表記異義の処理方法

論文中のキーワードに同表記異義の現象が存在する場合、共起キーワードに着目することにより、その概念を推定することができる。従って本報では、同表記異義を処理するための一方法として、概念を予め特定したキーワードを含む論文を学習サンプルとして用い、その共起情報に基づいて、着目するキーワードの概念を推定する方法を提案する。以下、収集した同表記異義の実例のうち、特に "PC" を例にとり、その処理方法を具体的に説明する。

まず PC という表記に対しては、"Personal Computer" という概念を持つ語と、"Programmable Controller" という概念を持つ語が存在するため、ここでは簡単のため前者を PC_1、後者を PC_2 と記述する。また、PC_1, PC_2 を含む論文を学習サンプルとし、それらの集合を G_1, G_2 と呼ぶ。さらに、それらの集合に属する要素数 (学習サンプル数) をそれぞれ N_1, N_2 とする。

ここで、キーワード：PC を含む一つの論文に着目し、その論文の PC 以外のキーワードを K_1, K_2, …, K_i, …, K_N とする。また、K_i が学習サンプル群 G_1, G_2 に出現する総数を n_{i1}, n_{i2} とする。このとき、G_1, G_2 中の学習サンプル 1 個あたりの K_i の出現頻度 f_{i1}, f_{i2} は以下のように表される。

$$f_{i1} = \frac{n_{i1}}{N_1}, \qquad f_{i2} = \frac{n_{i2}}{N_2}$$

Processing of polysemy using collocation information
Hiroya Fujisaki[1], Kenji Abe[1], Kenji Katami[1], Kazunari Taketa[1] and Katsuhiko Shirai[2]
[1]Science University of Tokyo, 2641 Yamazaki, Noda, 278-8501
[2]Waseda University, 3-4-1 Okubo, Shinjuku, 169-8555

また、K_iが PC_1, PC_2に関連して用いられる割合 w_{i1}, w_{i2}はそれぞれ以下のように表される。

$$w_{i1} = \frac{f_{i1}}{f_{i1} + f_{i2}}, \quad w_{i2} = \frac{f_{i2}}{f_{i1} + f_{i2}} = 1 - w_{i1}$$

(ただし、$f_{i1} = f_{i2} = 0$ のとき $w_{i1} = w_{i2} = 0.5$ とする)

さらに、全てのキーワード K_1, …, K_Nが PC_1, PC_2に関連して用いられる割合の合計は以下のように表される。

$$W_1 = \sum_{i=1}^{N} w_{i1}, \quad W_2 = \sum_{i=1}^{N} w_{i2} = N - W_1$$

従って、着目するキーワード：PCが PC_1および PC_2の概念で用いられる割合 R_1,R_2は次式で求められ、これらの値を比較することにより、キーワード：PCが PC_1, PC_2のどちらの概念で用いられているのかを推定することができる。

$$R_1 = \frac{W_1}{N}, \quad R_2 = \frac{W_2}{N}$$

なお、ここでは、1つの表記に2つの概念が対応する場合の処理方法について説明したが、この方法は1つの表記に3つ以上の概念が対応する場合に関しても容易に拡張することができる。

4. 同表記異義の処理実験

前節の方法に従い、収集した同表記異義のうち出現頻度が最も高い英略語に着目し、その概念を推定する実験を行った。なお、この方法では学習サンプルが必要となるが、英略語の場合、省略前の語が記載されている場合が多く、それに基づいて英略語の概念を自動的に特定する事ができるため、本研究では、省略前の語から概念を特定できる英略語を含む論文を学習サンプルとした。

収集した英略語による同表記異義のうち、学習サンプル数が適切で、かつ概念の候補数が2つのもの(134件)をテストサンプルとし、その概念を区別する実験を行った。ここで、判別率および正解率を次式で定義したときの実験結果を表2に示す。

$$\text{判別率} = \frac{\text{概念を区別できたものの総数}}{\text{テストサンプル数}}$$

$$\text{正解率} = \frac{\text{区別した結果が正しかったものの総数}}{\text{概念を区別できたものの総数}}$$

表2 判別率と正解率

判別率	71.6% (96/134)
正解率	100% (96/96)

また、学習サンプル数が比較的多い30件について、学習サンプル数を変化させた場合の実験を行い、学習サンプル数と判別率の関係を調べた結果を図1に示す。

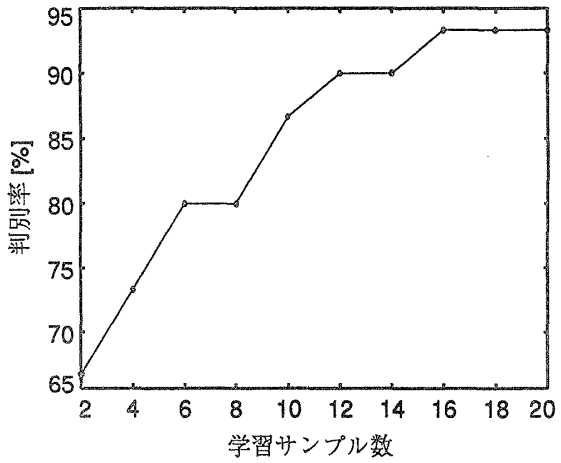

図1. 学習サンプル数と判別率の関係

5. おわりに

本報では、同表記異義の関係にあるキーワードの概念を共起情報を用いて推定する手法を提案した。また、その手法に基づいて、出現頻度の最も高い英略語の同表記異義を処理することにより提案する手法の有効性を検証した。

参考文献

[1] H. Fujisaki, H. Kameda, S. Ohno, T. Ito, K. Tajima and K. Abe : "An intelligent system for information retrieval over the Internet through spoken dialogue," *PROCEEDINGS of EUROSPEECH'97*, vol. 3, pp. 1675-1678 (1997).

[2] 劉軼, 戸井田和重, 八杉大輔, 阿部賢司, 大野澄雄, 藤崎博也, 久保村千明, 亀田弘之 : "学術情報検索における異表記同義・同表記異義の分類・分析および処理," 言語処理学会第4回年次大会発表論文集, pp. 108-111 (1998).

[3] 藤崎博也, 大野澄雄, 阿部賢司, 片見憲次, 飯島岐勇, 鈴木匡芳 : "キー概念に基づく情報検索方式の高度化(2)-キーワードの同表記異義の処理-," 情報処理学会第57回全国大会講演論文集, vol. 3, pp. 239-240, (1998).

[4] http://www.nacsis.ac.jp/nacsis.index.html