3Q-2

係り受け解析を用いた類義語抽出システムの開発

桜井　寛子†　　木村　昌臣†
†芝浦工業大学工学部

1．はじめに

テキストマイニングとは，形態素解析などを適用し，テキスト中から単語の頻度などを取得して，それらに対しデータマイニングを適用して解析する手法である．単語の頻度をもとに解析を行う際に，単語ごとに頻度が計算されるため，同義語であっても頻度は別々に計算されてしまう．

そのため，類似した内容の単語はまとめる必要があり，かつ，膨大な単語の量を解析するため，類似語の自動抽出を行うシステムが必要となってくる．

2．研究内容

類義語とは，同じ意味を持った単語であるが，その種類として，下位語・表記の揺れ・同義語といったものがある．今回は表記の揺れ・同義語を類義語とし，それらを抽出してくる．

同じ意味の単語であるならば，その単語に係る単語または係られる単語が類似したものになると考えられるので，次のような手順で，類義語を抽出する．

2.1　前処理

BLOG の本文などのインフォーマルな文章では，顔文字など内容と直接関係ない記号が用いられることが多い.そのため,テキストを形態素解析し，読点以外の記号を削除する．流れを図１に示す.

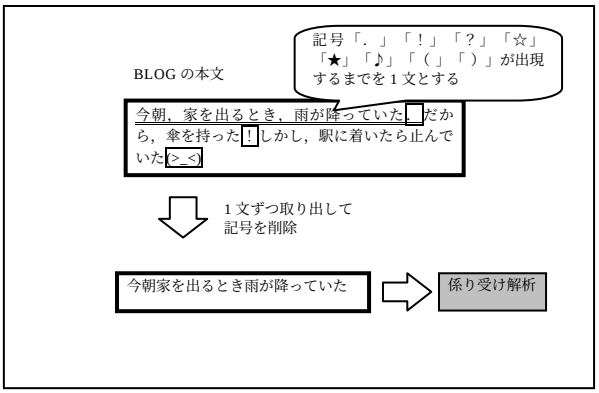

図１　本文から係り受け受け解析までの流れ

まず，本文から「！」「♪」などの記号までの文章を抽出し，それを１文とする．次に，その１文から，読点以外の記号を削除する．そして，その処理を行った 1 文を係り受け解析に適用する．

2.2　係り受け解析の適用

係り受け解析ツール CaboCha を用いて，テキストを係り受け解析する．実際に必要とする単語は，名詞（数・接尾・代名詞・固有名詞－人名・非自立・副詞可能以外）と動詞・形容詞・形容動詞（自立）である．ただし，「する」については直前の名詞（例えば，「静止」）と結合し，サ変動詞（「静止する」）として扱う．

2.3　二部グラフ作成

係る単語と係られる単語を関係づける二部グラフを作成する．

2.4　行列の作成

まず，係られる単語を行，係る単語を列，テキストに出現する係り受け関係の頻度を要素とした行列を作成する．

ここで，低頻度語は出現頻度が少ないために，そこから係り方の出現パターンの情報を得ることは難しい．さらに，低頻度語は，類義語としてまとめても，まとめた後の語の頻度に寄与する割合は低い．そのため，本研究では対象とする単語を中頻度以上出現する単語に限った．中頻度 f は，１度しか現れない単語の数 F を利用して，次のように表すことができる[4].

$$f \geq \frac{\sqrt{8F+1}-1}{2} \qquad (1)$$

上記の方法を利用して，係られる単語と係る単語，それぞれ別々に中頻度を求める．

そして，ある閾値以上の係り受け関係の頻度から行列を作成した後に，各列の平均を求めて全ての要素からそれぞれ対応している列の平均を引く．

2.5　特異値分解の使用

2.3 節で作成した行列 G に特異値分解を適用し，各単語とそれを特徴付けるベクトルとを対応させる．

特異値分解とは，行列 G を式(2)のように，3 つの行列に分解する手法である．

$$G = U \sum V^T \qquad (2)$$

Development of synonym extraction system using dependency analysis.
Hiroko Sakurai†, Masaomi Kimura†
†Department of Engineering, Shibaura Institute of Technology

Ｕと Ｖは直行化行列，Σは対角成分が固有値の絶対値の平方根，すなわち特異値をとる対角行列である．

ここで，主成分分析と特異値分解を対応させると，平均を引いた後の各単語に係る頻度の共分散行列 S は次のように与えられる．

$$S = \frac{1}{N} G^T G \qquad (3)$$

ここで，N は係られる単語の数である．この共分散行列 S に式(2)を代入すると，

$$S = V\Sigma^2 V^T \qquad (4)$$

となる．そのため，

$$H = GV \qquad (5)$$

と定義し、新しい行列 H を作成すると，この共分散行列は対角行列となる．さらに，式(2)と式(5)から，UΣが求められ，主成分が得られる[5]．ただし，係り受け関係の頻度のばらつきに最も寄与する量は，各単語の出現頻度であると考えられる．このことから，第一主成分は，この単語の出現頻度に依存すると考えられる．本研究では，出現頻度ではなく，係り受けの出現パターンから類義語を抽出するので，この第一主成分以外の成分を取り扱う．

2.6　類似度の計算

単語間の類似度を求める時，単語に対するベクトルのなす角度の余弦を利用する．単語 p，q に対応するベクトルを W_p, W_q とおくと，類似度は式(5)で与えられる．各ベクトルは UΣの行ベクトルに相当する．

$$sim(w_p, w_q) = \frac{\vec{w}_p \cdot \vec{w}_q}{|\vec{w}_p||\vec{w}_q|} \qquad (5)$$

３．実験

映画「ダヴィンチコード」を話題とした BLOG の本文，2000 件を対象データとして提案した類義語の求め方を検証する．

係り受けの頻度は 175027 個あり，中頻度の閾値は係られる単語で 154.045，係る単語で 116.042 である．閾値より大きい頻度の単語数は，係られる単語が 171 個，係る単語が 143 個ある．

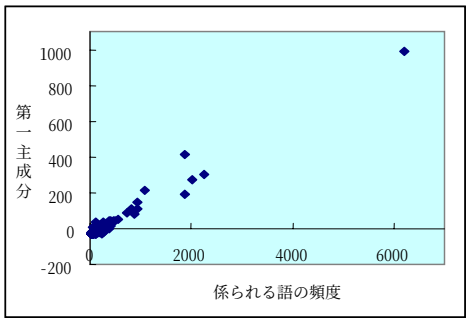

図 2　第一主成分と係られる語の頻度の関係

第一主成分と係られる単語の出現頻度の関係を示したものが図 2 である．この図から，第一主成分は係られる語の出現頻度と高い相関を持つことが見て取れる．

そこで，第一主成分を入れる方法と入れない方法，かつ，類義語同士の語は品詞が同じ単語が多いことから，品詞が同じ単語という制限を加える方法と加えない方法の計４つのパターンから類似度の比較を行う．

４．評価

人の目で見た類義語の組み合わせは 19 組あった．その中でも，表記の揺れが多い．例えば，「みる」「見る」「観る」や「シーン」「場面」の組み合わせなどがある．

3 章で述べた計４つのパターンから類似度を計算し，それぞれ，類似度の順位を計算する．それぞれ上位 16 位までの結果は，品詞の制限を加えて第一主成分を除くパターン(a）では 8 個，第一主成分を除くパターン(b)では 6 個，品詞の制限のみ行うパターン(c)では 2 個，第一主成分を含むパターン(d)では 2 個，類義語の組み合わせをそれぞれ検出した．

そして，19 組の類義語の組み合わせの順位の合計を検証すると，(a)は 15874，(b)は 29011，(c)は 113710，(d)は 50741 となった．二つの結果から，第一主成分を除き，かつ，同じ品詞の単語の組み合わせという制限を加えた(a)のパターンが，より上位に類義語が出現することが分かった．

５．おわりに

今回は，2000 件の BLOG のデータを用いたが，処理時間が長くかかってしまった．より多くのデータを取り扱う必要があるので，処理時間を短縮し，かつ，類義語抽出の精度の向上を目指す．

参考文献

[1]笹原要，稲子望，加藤恒昭“単語の属性空間の表現方法”人工知能学会論文誌（2002）
[2]笹原要，稲子望，加藤恒昭“テキストデータを用いた類義語の自動作成”人工知能学会（2003）
[3]中渡瀬秀一“複合語からの類義語抽出法”情報処理学会（2002）
[4]徳永健伸“情報検索と言語処理”東京大学出版会（1999）
[5]竹村彰通，金明哲，村上征勝，永田昌明，大津起夫，山西健司“言語と心理の統計”岩波書店（2003）