名詞概念との共起関係を用いた用言概念の分類

Clustering verb, adjective, adjective-verb concepts using co-occurrence relation with noun concepts

藤本太郎　　菅野道夫
東京工業大学 システム科学専攻
tarof@fz.dis.titech.ac.jp

Abstract :
本研究では実言語の知識を反映することにより、時間的、分野的なテクストの変化に対応できる頑健な自然言語処理を可能にするため、実言語からの知識を用いて単語概念を分類する手法を提案する。
まず、動詞、形容詞、形容動詞概念を共起する名詞概念を用いて定義した類似度とその閾値（α-カット）によって分類を行なった。さらに、頻度情報を反映させるために、頻度が低い用言組の類似度を下げるファジィフィルタを提案し、修正を加えた。この結果、動詞概念のように十分に標本数がある場合にはフィルタの効果があることが確認された。
対照実験としてEDR概念辞書の体系を用い、コーパスから抜き出した用言をクラスタリングし、比較した。提案した手法では動詞では100から451、形容詞では15から41、形容動詞では18から105に分割され、EDR概念辞書では得られない用言概念間の微妙な結び付きを表現することができ、従来の構造的立場に因われない用言の分類を行なうことができた。また、その有効性を文書分類に応用することで確かめた。

1　はじめに

EDR概念辞書 [1] などのシソーラスは人間が主観によって階層化した知識をデータベース化したものであり、これを用いればある程度の自然言語処理を行なうことができる。しかし、言語は刻々と変化し、また分野によって使われる単語もその結び付きも変化する。そのため単語どうしの結び付きもその時刻、分野によって変化するため、ある一定の文章のまとまり（コーパス）から意味体系が抽出できることが望ましい。これは自己組織化的自然言語処理 [2] の一領域に属する。与えられたコーパスからその文書に関する情報を得ることにより、辞書に未登録な単語、意味体系が現われても頑健なシステムを構築することが可能である [2]。

本研究では用言間の関係をファジィ類似関係 [3][4] (3.1節) を用いて抽出し、コーパスから得られる頻度情報を反映させるために、ファジィフィルタ (3.2節) を用いることを提案する。

対照実験として、コーパスから得られた用言概念関係をEDR概念辞書 [1] から得られる概念関係を用いた分類結果を用いた。さらに、本研究の有効性を確かめるために、コーパス中の文書を得られた用言で分類を行なった。

2　名詞・用言関係

本研究では、同じ名詞を主語 (Theme) にとる用言どうしは機能的な意味も近いと考え、用言概念を共起する名詞概念によって分類した。選択体系機能文法 [5] における日本語の Theme-Rheme 構造の定義 [5][6] を用いて、EDR日本語コーパス [1] より「名詞 + 助詞"は" + 用言（終止形）。」の構造を持つ文を抽出し、名詞、用言それぞれに概念識別子を付与されているものを実験データとして用いる[1]。

3　ファジィ類似関係を用いた分類

本研究では用言を分類する指標として概念と概念の間の類似度を定義し、任意の閾値以上の類似度を持つものを関係ありとするファジィ類似関係を用いる。また、得られた関係にコーパスの頻度

[1] 概念識別子を用いた理由は多義性によって正しく分類できない可能性があるためである [7]。

情報を付与する手法として、平均以下の頻度を持つ単語に対して補正を加えるファジィフィルタを構成した。

3.1 ファジィ類似関係

ファジィ関係とは、要素間のあいまいいさを含む関係を量的に記述するための概念である。ファジィ類似関係とはファジィ二項関係のうち反射律、対象律を満足するものである。[1, 1] 内の任意の閾値（α-カット）によって適当な分類を得ることができる [3][4]。

用言間の類似度を式 1に示すものを用いる。この類似度の求め方はファジィシソーラス構築の際に用いる関連度 [8] の定義を単純化したものである。名詞 $X\{y_1, y_2, \cdots\}$ と用言 $Y\{x_1, x_2, \cdots\}$ の関係において、x_iと共起する名詞の集合 R_{x_i}、x_jと共起する名詞の集合を R_{x_j}とし、x_iと x_jの間の関係の強さ Sを以下の式 1のように定義する。

$$S_{x_i x_j} = \frac{|R_{x_i} \cap R_{x_j}|}{min\{|R_{x_i}|, |R_{x_j}|\}} \qquad (1)$$

3.2 ファジィフィルタ

出現頻度はコーパスからの知識抽出の際に得られる重要な情報である。そこで出現頻度を知識に反映する様々な手法 [2][7][9][10] があるが、本研究では出現頻度をかけ算する手法と、閾値を用いて切り捨てる方法の折衷としてファジィフィルタを提案する。

ファジィフィルタは「平均以下の値は小さい」という一般的な知識を If-then ルールで「If x value is smaller than avarage, then x is small.」と解釈し、さらにファジィを用いて表現し、補正 $Freq_x$ は以下のように表したものである。

$$Freq_x = \begin{cases} 1 & (If \ \ x \geq \bar{x}) \\ \frac{x}{\bar{x}} & (If \ \ x \geq \bar{x}) \end{cases} \qquad (2)$$

類似度に関係する 2 つの概念の頻度を考慮し、2次元のファジィフィルタをかけ算で合成すると、図 1に示されるような 2 次元のフィルタを構成できる[2]

[2]かけ算で合成するだけではなく、他のファジィt-norm 演算 [11] を用いることも可能である。

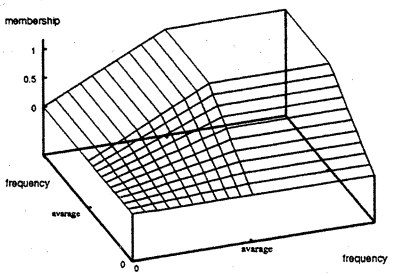

図 1: 2次元のファジィフィルタ

4 EDR 概念辞書を用いた分類

コーパスより抽出された用言概念データに関して、EDR 概念辞書の概念体系より図 2に示す parent、brother、uncle、cousin のラベルを持つ関係を抽出した。

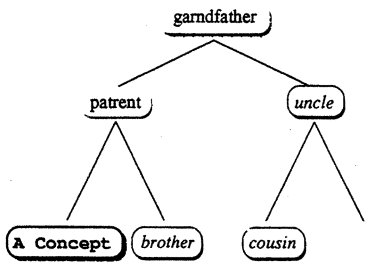

図 2: 概念関係

5 実験

実験は以下の手順で行なった。

1. EDR 日本語コーパス [1] から日本語 Theme-Rheme 構造を持つものを抽出し、Rheme に動詞、形容詞、形容動詞を持つものをそれぞれ分類し、概念 ID を付与されていないものを棄却、整理（結果は 5.1節）。

2. 動詞概念について類似度（3.1節）を求め、それぞれの類似度を持つ関係の数を求める。

3. 類似度をαカットし、閾値以上の類似度を持つ関係によって構成されるグループの数、グループ化されなかった概念の数、これらの合計である分割数をそれぞれ求める（結果は5.2節）。

4. 類似度にファジィフィルタ（3.2節）をかける。

5. ファジィフィルタをかけた類似度をαカットし、関係を持つ対の数、分割数をそれぞれ求める

6. 2から5までの処理を形容詞、形容動詞についても行なう。

7. 1で求めた用言データに関してEDR概念辞書の概念体系を用いて関係を持つ対の数、分割数をそれぞれ求める。

5.1 実験に用いたデータ

実験データには主に朝日新聞、日経新聞、アエラの記事から構成されるEDR日本語コーパス208157文から抽出した名詞–用言データのうち、助詞「は」によって結合している109295文を用いた。

5.2 ファジィ類似度を用いた分類結果

表1に示すように、ファジィフィルタをかける前と後ではαカットされた類似度による類似度が0でない対の数、分割数に変化が見られる。ファジィフィルタでは動詞、形容詞、形容動詞の平均頻度7.5、19.4、24.1をフィルタに反映させて類似度を修正した。

具体的に動詞の分割数を図3に示す。フィルタなしの場合、図3の＋で示されるように100($\alpha = 0.00$)から196($\alpha = 1.0$)のように段階的に分割数が増加するのに対し、フィルタをかけた場合は図3の×で示されるように100($\alpha = 0.00$)から451($\alpha = 1.0$)までをより広い範囲を滑らかに増加していく。

5.3 EDR概念辞書を用いた分類結果

動詞の場合は分割数が13,52,76,426,490の5種類、形容詞の場合は18,42の2種類、形容動詞の場合は13,44 78 130の4種類の分割ができた。いずれの分割の場合もグループ化されないか、グループ化されると単一のグループに収束する。

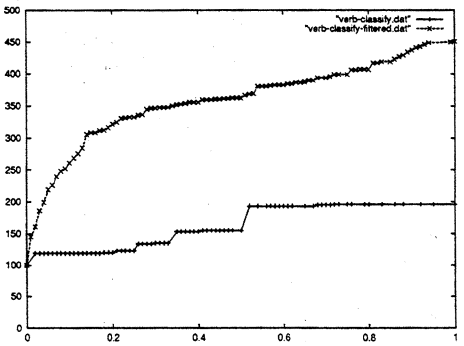

図3: 類似度のみと類似度とファジィフィルタを用いた動詞の分割数

5.4 考察

ファジィフィルタの効果 ファジィフィルタは十分な類似度を持つ対の数が多い動詞概念関係についてはコーパスの頻度情報を反映する手法として有効であると考えられる。

動詞でフィルタをかけない場合、（引き返す–復興する）、（破裂する–反発する）など類似度が1でもどういう関係かが分からないものがあるが、フィルタをかけた場合はこれらのものが取り除かれる。しかし、形容詞の場合はフィルタをかけても（ない–広い）が類似度1で出てきており効果は薄い。ファジィフィルタは十分な概念の要素数があった場合には有効であると考えられるが、そうでない場合の結果は良くないと結論付けられる。

提案した手法とEDR概念辞書を用いた分類との比較 動詞の場合について考える。提案した手法では分割数が100から451の66通りに分けられるが、EDR概念辞書を用いた分類では13から490の8通りにしか分けられない。さらに、EDR概念辞書を用いた分類では関係がある概念はすべて1つのクラスに収まり、微妙な類似度の程度は出てこない。これは階層的・構造的に人間が考えた構造のためと考えられる。

フィルタ	類似度を持つ対		分割数	
	無し	有り	無し	有り
動詞	1009～6222	42～6222	100～196	100～451
形容詞	10～111	1～111	15～35	15～41
形容動詞	296	148～296	18	18～105

表 1: ファジィフィルタによって補正されたαカットされた類似度による類似度が 0 でない対の数および分割数

5.5 文書分類への応用

本手法で分類された動詞、形容詞、形容動詞によってEDRコーパス[1]の文を分類する。"名詞＋助詞「は」＋用言。"の形を持つ72016文中55228文(76.7%)を133(α-カット値0)から598(α-カット値1)のクラスタに分類できた。

一例を示すと、α-カット値が1のものでは、「変換」と「放射」に関する文書が同じグループになる。これは比較的近い意味の言葉であると考えられる。次にα-カット値が0.5のものでは、「「戸惑う」と「話す」に関する文書が同じグループになる。これは「変換」と「放射」に比べるとやや遠い印象がある。最後にα-カット値が0.01のものは、「入信する」と「臥せる」が同一の分類内に入る。おそらく「祖母」という概念が偶然同じThemeとなり、ファジィフィルタでも全くは排除しないためと考えられる。

6 おわりに

本研究では用言を共起する名詞を類似度を定義して分類し、頻度を基にしたファジィフィルタによってデータベースがもつ頻度情報を反映した。また、評価として文書を分類し、その有効性を確かめた。

なお本研究では名詞–用言間の構造についての再考、動詞の多義性などの課題が残されている。将来的には本研究で提案した手法を分野毎のテクストに当てはめて、分野毎の意味体系やそれらを統合し、一般化した大局的な意味ネットワーク構築を目指したいと考えている。

参考文献

[1] (株) 日本電子化辞書研究所. EDR 電子化辞書仕様説明書. Technical report, (株) 日本電子化辞書研究所, Mar. 1993.

[2] 中渡瀬 秀一. 統計的手法による分かち書き境界の獲得. 信学技法 [言語理解とコミュニケーション] *NLC95-77*, Vol.96, Mar. 1996.

[3] L.A.Zadeh. **Similarity relations and fuzzy ordering**. *Inf.sci*, Vol.3, 1971.

[4] D. Dubois and H. Prade. *FUZZY SETS AND SYSTEMS*. ACADEMIC PRESS, New York, 1980.

[5] M. A. K. Halliday. **AN INTRODUCTION TO FUNCTIONAL GRAMMAR**. Edward Arnold, London, second edition, 1994.

[6] 佐々木 真. **An Analysis of Realization of Theme in Japanese**. 愛知学院短期大学 研究紀要, 第4号, Mar. 1996.

[7] F. Fukumoto and J. Tsujii. **Automatic Recognition of Verval Polysemy**. *Proc. 15th COLING*, 1994.

[8] ファジィ学会 編. 講座 ファジィ9 ファジィデータベースと情報検索. 日刊工業新聞社, Sep. 1993.

[9] 北 小倉 森本 矢野. 仕事量基準を用いたコーパスからの定型表現の自動抽出. 情報処理学会論文誌, Vol.34 No.9, Sep. 1993.

[10] 井佐原 均 新納 浩幸. 疑似Nグラムを用いた助詞的定型表現の自動抽出. 情報処理学会論文誌, Vol.36 No.1, Jan. 1995.

[11] 水本 雅晴. ファジィ集合とファジィ推論. 第3回ファジィシステムシンポジウム, Jun. 1987.

[12] 井ノ上 直己 工藤 育男. コーパスに基づく共起知識の獲得とその応用. 人工知能学会誌, Vol .10, Mar. 1995.