自然言語処理 154－13
（2003．3．7）

多義性を軽減した素性セットによるテキスト分類方式

荒木 淳[†] 中村文隆[††] 中山雅哉[††]

本稿では、テキスト分類において素性選択法によって選出された素性セットの多義性を軽減する方式について述べる。素性選択は相互情報量などの評価尺度に基づき、その値が高い上位単語を用いる手法が行われてきた。しかしながら、単語には複数の意味を持つ多義語も含まれるため、そういった多義語が素性として選出された場合は意図したカテゴリと異なるカテゴリの文書にも含まれるので分類上問題がある。素性選択の尺度として相互情報量を用いた場合に、素性セットの多義性の問題を解決する手段として単語とカテゴリの共起回数を利用した多義性軽減手法を提案し、Reuters-21578コーパスを用いた分類実験により提案手法の有効性について評価する。

Text Classification with a Polysemy Condidered Feature Set

JUN ARAKI,[†] FUMITAKA NAKAMURA[††] and MASAYA NAKAYAMA[††]

This paper proposes a method for considering the polysemy of a feature set in text classification. Feature selection has been performed selecting words with higher evaluation value such as mutual information and so on. However, polysemous words, which have bad influence on classification accuracy, can be selected as features even if we use the evaluation. We propose the text classification method using the number of times of coincidence of a word and a category. In the experiment we test Reuters-21578 corpus and verify our method.

1. はじめに

大量の電子化されたテキストの蓄積や流通に伴い、これらの大量のテキスト情報の中から有用な情報をいかに効率的に抽出するかが重要な課題となっている。テキストを予め定められたカテゴリに分類することは上述の課題に対処する一つの手段であると考えることができる。というのは、多くの情報を予め定められたカテゴリに沿って分類されているならば、その中から必要な情報を探し出すことは比較的容易である。このことから、情報を分類することは情報へのアクセスを支援する一つの方法と考えることができるからである。

テキスト分類技術の実現例として、Webページが予めツリー上に分類されているディレクトリ型と呼ばれるインターネット上の検索エンジンがある。その他に、送られてくる電子メールに対して「普通のメール」と「スパム」のカテゴリに振り分けることで、不特定多数に一方的に送りつけられる広告や勧誘等のスパムメールをシャットアウトするなどといった応用例が考えられる。

テキスト分類技術の歴史的な変遷を振り返ると、1980年代後半までは知識工学的アプローチ、すなわち人手で分類規則を書く手法が主流であった。しかし、1990年代に入ると、大量のテキストデータが利用可能になったことやコンピュータの性能が大幅に向上したことから、機械学習的なアプローチ、すなわち人手によりカテゴリラベルを付与したテキストデータから自動的に分類器を作成する手法が、省力性や保守性の面で優れているために主流となった。さらに近年では、Support Vector Machine(SVM)やAdaBoost等の最先端の機械学習アルゴリズムが次々とテキスト分類に適用された結果、様々な学習理論の実用性を共通のベンチマークに基づいて比較検討することが可能となった。

機械学習アルゴリズムをテキスト分類に適用する場合、ベクトル空間モデル(VSM: Vector Spage Model)という概念を導入する。VSMでは文書を複数のキーワード(素性)に基づく多次元ベクトルとして表現する。文書ベクトルの各要素はある単語がその文書に出現するか否かという2値の場合もあれば、TF-IDF等のような適当な方法で重み付けした実数値の場合もある。多様な文書を高精度に分類するためには、できる

† 東京大学大学院工学系研究科
School of Engineering, University of Tokyo
†† 東京大学情報基盤センター
Information Technology Center, University of Tokyo

だけ多く (数万以上) の素性を使用することが望ましい。しかし、学習を行う際の過学習や計算時間の問題から、多くの分類器は数百から数千程度に素性を削減する必要がある[8]。そのため、単語出現頻度、文書頻度、相互情報量などの様々な評価基準を用いた素性選択法が提案されている[7]。

上述のような記号処理的なアプローチによる自然言語理解には未だに多くの困難が存在する。その一つに単語の多義性の問題がある。多義語は利用される分野や文脈により意味が異なってくる。こうした多義語を含む文書の分類については既に研究がなされている。例えば、単語の共起関係に立脚して直接単語の共起分布を利用する手法[10]や概念空間と呼ばれる全く別の縮退した空間に写像する手法[1]などが考えられている。

提案手法では、素性選択法として相互情報量を用いた場合に、多義語を分類を行う上での危険因子的な素性と考え、素性単語が危険因子であるかどうかについての条件判定を行い、危険因子と判断された素性を除去することにより素性セットから多義性を解消する。分類対象として Reuters-21578 コーパスを用いてこの方式と従来方式による分類結果を比較し、その有効性を評価する。

本稿の構成は以下の通りである。次の第 2 節で関連研究について触れ、第 3 節で従来から用いられてきた相互情報量による素性選択について述べる。第 4 節で本実験で分類器として用いた Support Vector Machine(SVM) を解説し、第 5 節で単語とカテゴリの共起回数を利用した多義性軽減方式を提案する。第 6 節で本研究に用いた実験方法や実験結果について述べ、第 7 節で考察を行う。最後に第 8 節で本稿をまとめる。

2. 関連研究

テキスト分類において、単語の多義性の問題に対して既にいくつかの手法が提案されている。湯浅らは、ある一つの新聞記事中で共起している名詞の出現頻度を名詞間共起分布とし、これに基づく名詞間共起分布ベクトルから各分野の基準ベクトルを作成し、その分野基準ベクトルと未知の記事を比較することで分類を行う手法を提案している[10]。記事を 5 分野に分類する実験を行い、平均して 80%以上の正解率 (再現率) を得たと報告している。

その他に、特異値分解 (SVD: Singular Value Decomposition) という多変量解析手法を用いて、文書を概念空間に配置する潜在的意味インデキシング (LSI: Latent Semantic Indexing)[1] と呼ばれる手法もある。VSM では空間は単語を軸とした空間で構成されるのに対し、LSI ではベクトル空間の軸を SVD によって合成した、より少数の軸で意味空間を構成する。SVD を用いることで出現回数の多い共起単語の組を共有する文書は、潜在的意味空間において類似関係を持つことになる。

3. 相互情報量を用いた素性選択

本節では、テキスト分類において従来から行われてきた相互情報量を用いた素性選択法について述べる。

第 1 節で述べたようにテキスト分類では過学習と計算時間の問題から、特徴空間の次元削減のために、より分類を際立たせるような素性だけを予め選ぶ素性選択が用いられる。素性選択に用いられる様々な評価基準として文書頻度、相互情報量、情報利得、カイ 2 乗検定などがある[7]。本研究では素性選択法としてカテゴリと単語間の相互情報量を用いた。

相互情報量 (MI: Mutual Information) はカテゴリと単語の関連性の強さを表す指標である。文書集合における単語 T とカテゴリ C の相互情報量 MI は次式で定義される。

$$MI(T,C) = \sum_{t\in\{T,\bar{T}\}}\sum_{c\in\{C,\bar{C}\}} P(t,c)\log\frac{P(t,c)}{P(t)P(c)} \quad (1)$$

ここで、$\bar{T}$ は T でない単語集合、$\bar{C}$ は C でないカテゴリ集合を表し、

- $P(t)$: 全記事中で単語集合 t を含む記事の割合
- $P(c)$: 全記事中でカテゴリ集合 c に属する記事の割合
- $P(t,c)$: 全記事中で単語集合 t を含み、かつカテゴリ集合 c に属する記事の割合

である。

表 1 単語 T とカテゴリ C の共起回数

	C occurs	C doesn't occur
T occurs	a	b
T doesn't occur	c	d

表 1 に示したのは単語 T とカテゴリ C の共起回数に関する表である。例えば a は単語 T とカテゴリ C が同時に発生した回数を示しており、これは即ち単語 T を含み、かつカテゴリ C に所属する記事の数が a ということである。記事の総数を $N = a+b+c+d$ とすると、相互情報量 MI は次式のように表される。

$$MI(T,C) = \frac{a}{N}\log\frac{aN}{(a+b)(a+c)}$$

$$+\frac{b}{N}\log\frac{bN}{(a+b)(b+d)} +\frac{c}{N}\log\frac{cN}{(c+d)(a+c)} +\frac{d}{N}\log\frac{dN}{(c+d)(b+d)} \quad (2)$$

式 (2) で示された相互情報量は正規化されており 0 から 1 までの値をとる。相互情報量は、単語 T の出現頻度がある一つのカテゴリ C と他のカテゴリの間で偏りがないときに小さな値をとり、偏りがあるときに大きな値をとる。

4. Support Vector Machine によるテキスト分類

Support Vector Machine(SVM) は、訓練データを正例と負例に分け、かつ正負例間のマージンが最大になるような超平面を求め、それによって分割を行う 2 値分類器である。本研究で分類器として用いたので、本節で SVM について説明する。

図 1 に SVM の概念図を示す。ここでは 2 値分類問題を考え、入力を文書ベクトル $\mathbf{x}$、出力を正例・負例のクラスラベル $y = 1$ または -1 とする。図 1 に示されるように、正例と負例のデータは 2 つの超平面

$$\begin{cases} \mathbf{w}\cdot\mathbf{x_i} \geq \mathbf{1} & \text{if } y_i = 1 \\ \mathbf{w}\cdot\mathbf{x_i} \leq \mathbf{-1} & \text{if } y_i = -1 \end{cases} \quad (3)$$

によって分類される。最も負例寄りの正例側の境界面 $\mathbf{w}\cdot\mathbf{x}+b=1$ と最も正例寄りの負例側の境界面 $\mathbf{w}\cdot\mathbf{x}+b=-1$ の間の距離をマージンと呼ぶ。このとき求める分離超平面は、$\mathbf{w}\cdot\mathbf{x}+b=0$ である。また、$\mathbf{w}\cdot\mathbf{x}+\mathrm{b}=\pm 1$ 上の訓練データをサポートベクタ (support vector) と呼ぶ。

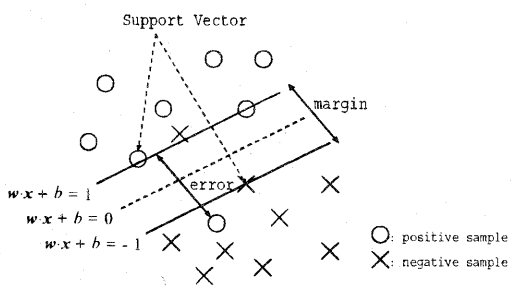


図 1 SVM の概念図

分離超平面から点 $\mathbf{x}$ までの距離 d は

$$d(\mathbf{w},\mathrm{b};\mathbf{x}) = \frac{|\mathbf{w}\cdot\mathbf{x}+\mathrm{b}|}{||\mathbf{w}||} \quad (4)$$

と表されるので、マージン ρ は

$$\rho = \min_{\mathbf{x_i};y_i=1}\frac{|\mathbf{w}\cdot\mathbf{x_i}+\mathrm{b}|}{||\mathbf{w}||} + \min_{\mathbf{x_i};y_i=-1}\frac{|\mathbf{w}\cdot\mathbf{x_i}+\mathrm{b}|}{||\mathbf{w}||} = \frac{2}{||\mathbf{w}||} \quad (5)$$

となる。よって、マージン ρ を最大化することは $||\mathbf{w}||$ を最小化することであり、式 (7) の制約条件の下で、式 (6) の目的関数の最小化問題を解くことと等価である。

$$\text{目的関数: } L(\mathbf{w}) = \frac{1}{2}||\mathbf{w}||^2 \rightarrow \text{最小} \quad (6)$$

$$\text{制約条件: } y_i[(\mathbf{w}\cdots\mathbf{x_i}+\mathrm{b})] \geq 1(\mathrm{i}=1,\ldots,\mathrm{l}) \quad (7)$$

この 2 次計画問題を数値計算で求めた解を $\bar{\alpha_i}$ とし、正極側のサポートベクタを $\mathbf{x_p}$、正極側のサポートベクタを $\mathbf{x_n}$ とすると、以下のようにして $\bar{\mathbf{w}}$ と $\bar{b}$ を決定できる。

$$\bar{\mathbf{w}} = \sum_{i=1}^{l}\alpha_i y_i \mathbf{x_i} \quad (8)$$

$$\bar{b} = -\frac{1}{2}(\mathbf{w}\cdot\mathbf{x_p} + \mathbf{w}\cdot\mathbf{x_n}) \quad (9)$$

最終的に正例、負例を判別する関数は、

$$f(\mathbf{x}) = \mathrm{sgn}(\bar{\mathbf{w}}\cdot\mathbf{x}+\bar{\mathrm{b}}) \quad (10)$$

で与えられる。ただし、$sgn(z)$ は引数 z の符号を表す関数で、

$$sgn(z) = \begin{cases} 1 & \text{if } z \geq 0 \\ -1 & \text{if } z < 0 \end{cases} \quad (11)$$

である。

5. 多義性軽減方式

5.1 素性の二面性

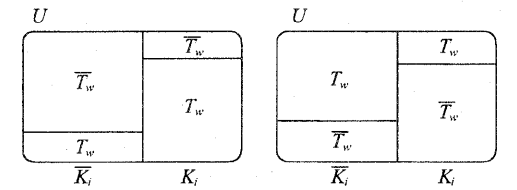
カテゴリを特徴付ける素性には、正に特徴付けるものと負に特徴付けるものがある。カテゴリを正に特徴付ける素性は、その素性を含めば文書はそのカテゴリに属す、あるいはその素性を含まなければ文書はそのカテゴリに属さないという性質を持つ素性である。逆に、カテゴリを負に特徴付ける素性は、その素性を含めば文書はそのカテゴリに属さない、あるいはその素性を含まなければそのカテゴリに属すという性質を持つ素性である。正に特徴付ける素性、負に特徴付ける素性はどちらもカテゴリを特徴付ける点で有用である。

いま U を文書集合全体、K_i を分類対象のカテゴリに含まれる文書の集合、$\overline{K_i}$ を分類対象以外のカテゴリに含まれる文書の集合の和とし、ある単語 w を含

む文書の集合を T_w、含まない文書の集合を $\overline{T_w}$ と表すことにすると、w がカテゴリを正に特徴付ける素性である場合とカテゴリを負に特徴付ける素性である場合に T_w と $\overline{T_w}$ が K_i と $\overline{K_i}$ にどのように分布するかを模式的に描いた図を**図 2** に示す。

w がカテゴリを正に特徴付ける場合は T_w が K_i に多く、$\overline{T_w}$ が $\overline{K_i}$ に多いという傾向にあるのに対して、負に特徴付ける場合は T_w が $\overline{K_i}$ に多く、$\overline{T_w}$ が K_i に多いという傾向にある。

カテゴリを正に特徴付ける場合は第 3 節で述べたカテゴリと単語の共起回数で表現すると、表 1 中の記号で a と d の値が大きければ大きいほど正に特徴付けると言うことができるので、この両者の和をとって正に特徴付ける素性は $a+d$ の値が大きいと統一化して表現できる。同様に、負に特徴付ける素性は $b+c$ の値が大きいと表現できる。



(a) the case of a feature characterizing to positive　(b) the case of a feature characterizing to negative

図 2　正に特徴付ける素性と負に特徴付ける素性を含む文書集合

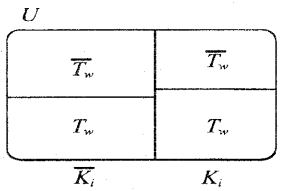


図 3　多義的な素性を含む文書集合

5.2　単語とカテゴリの共起回数による多義性の定義

正と負の特徴付けという素性の二面性を考慮すると、多義的な素性はその多義性故に正にも負にも特徴付けることができない素性であると考えることができる。従って、w が多義的な素性である場合は、T_w も $\overline{T_w}$ もカテゴリ K_i と $\overline{K_i}$ の両方に多く存在するので、多義性が強ければ強いほど、**図 3** に示すような分布に近づくと考えられる。よって、多義的な素性 w は、T_w と $\overline{T_w}$ がともにカテゴリ K_i と $\overline{K_i}$ の両方に多く存在するという性質を持つ。

このような性質を持つ単語は、上述の議論から $a+d$ の値も $b+c$ の値も偏って大きくないので、$N=a+b+c+d$ という制約条件を考慮すると、$a+d=b+c=N/2$ の場合が最も多義性が強いと考えられる。これを踏まえて、こういった多義的な単語を定量的に判定するために、次のような 2 次関数を用意することによって多義性の判定条件を定式化した。

定義 (多義性の判定条件)

分類対象のカテゴリを K_i とし、K_i と単語 w の共起回数 (表 1 参照) を a_i、b_i、c_i、d_i とする。このとき、w が多義的な素性となるのは閾値 ϵ に対して次の条件式を満たすときであると定める。

$$s=\sqrt{\frac{a_i+d_i}{N}\,\frac{b_i+c_i}{N}}>\epsilon \qquad (12)$$

ただし、$N=a_i+b_i+c_i+d_i$, $0\leq\epsilon\leq 1/2$ である。

なお、a_i+d_i, b_i+c_i の範囲は $0\leq a_i+d_i\leq N$, $0\leq b_i+c_i\leq N$ で、s は $a_i+d_i=0$ または $a_i+d_i=N$ のとき最小値 $s_{min}=0$ をとり、$a_i+d_i=b_i+c_i=N/2$ のとき最大値 $s_{max}=1/2$ をとる。よって、s の範囲は、$0\leq s\leq 1/2$ である。なお、s を a_i+d_i の関数としてグラフ表示すると**図 4** に示すような半楕円になる。

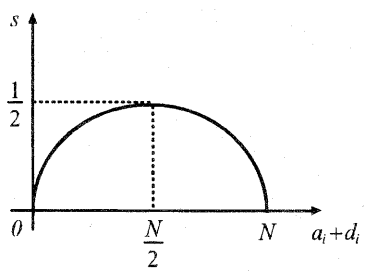


図 4　a_i+d_i の関数としての s

5.3　多義性軽減方式の提案

第 3 節で述べた相互情報量による素性選択法を行ったとしても、各素性の意味カテゴリ (単語が関連性の深い意味的なカテゴリ) の数は一つとは限らない。分類モデルに与える特徴として、特定のカテゴリに絞らずに多義性があるまま全ての意味カテゴリを利用すると、訓練文書中のノイズが増え、分類精度を悪化させる可能性がある。

例えば、単語"plant" は「植物」の意味で"grain" カ

テゴリと関連性が深く、「工場」の意味で"oilseed" カテゴリと関連性が深い多義語である。「植物」の意味の"plant" を含む文書は"grain" カテゴリと無関係であり、「工場」の意味の"plant" を含む文書は"oilseed" カテゴリと無関係であるので、単純に全てのカテゴリで利用すると学習モデルに無関係な特徴を与えることになる。

よって、我々は多義語を分類上の危険因子的な素性と考え、これを素性セットから除去することにより多義性を軽減する方式を提案する。具体的には相互情報量で選択した各素性に対して、前節で定義した多義性の判定条件式を用いて多義語の判定を行い、多義語と判定されたものは素性セットから除外するという処理を行う。

6. 実　　験

6.1 実験設定

テキストコーパスには Reuters-21578☆を用いた。Reuters-21578 の訓練記事とテスト記事の振り分け方法は、"The Modified Apte(ModApte) Split"(付録 A.1 参照) に従った。この方法に用いると訓練記事は 9603 個、テスト記事は 3299 個に振り分けられる。本実験ではこれらを訓練記事、テスト記事として用いた。

本実験で分類処理に用いたカテゴリを**表 2** に示す。表 2 には分類実験に用いた各々のカテゴリに属する訓練記事とテスト記事の数も示した。

これらの記事に対して、tree-tager☆☆という品詞付与ツールを用いて禁止語の除去、複数形・語形変化への対応、数詞の除去を行った。禁止語の除去では接続詞、冠詞等を除去し、複数形・語形変化への対応では名詞の複数形を単数形にし、形容詞の比較級を原形にするなどといった処理を行った。この処理の詳細は付録 A.2 に記述した。この前処理によって訓練記事の総単語数は 45558 個から 17526 個に絞られた。

6.2 評価方法

分類精度の評価尺度として F 値を用いた。

- A: 正解が正例で、分類システムも正例と判断した文書の数
- B: 正解が負例で、分類システムは正例と判断した文書の数
- C: 正解が正例で、分類システムは負例と判断した文書の数

☆ 付録 A.1 参照。
☆☆ 付録 A.2 参照。

表 2 分類処理に用いたカテゴリに属する訓練記事とテスト記事の数

カテゴリ名	訓練記事	テスト記事
acq	1650	719
corn	182	56
crude	389	189
earn	2877	1087
grain	433	149
interest	347	131
money-fx	538	179
ship	197	89
trade	369	118
wheat	212	71

とすると、分類システムの適合率 P と再現率 R は下の式で表される。

$$P = \frac{A}{A+B}, \quad R = \frac{A}{A+C} \tag{13}$$

分類精度を示す F 値は β をパラメータとして、

$$F_\beta = \frac{1+\beta^2}{\frac{1}{P}+\beta^2\frac{1}{R}} \tag{14}$$

と表される。本実験では、$\beta = 1$ とした場合の

$$F_{\beta=1} = \frac{2PR}{P+R} = \frac{2A}{2A+B+C} \tag{15}$$

を評価値として用いた。

6.3 相互情報量による素性選択

分類処理に用いた 10 個のカテゴリにおいて相互情報量の高い上位単語のうち、1 番目、5 番目、10 番目、50 番目、100 番目、500 番目、1000 番目、2000 番目の単語を**表 3** に示す。上位単語はカテゴリに非常に関連性が強い単語が選出されているのが分かる。また、上位 100 単語では名詞が多い傾向にある。

6.4 実験結果

まず従来手法による Reuters-21578 の分類実験を行った結果を**図 5** に示す。素性は相互情報量の上位 1000 単語を用いた。また、SVM は SVM^{light} Ver5.00☆☆☆というフリーのソフトウェアを利用し、線形 SVM によって分類を行った。なお、F 値は 0 から 100 までの値をとるように補正した。この図から、全てのカテゴリに共通することとして、ある程度素性数が多くなると F 値はそれ以上向上しないということが分かる。カテゴリ別には、"acq" や"earn" などのカテゴリで 80〜90 ポイントと F 値が比較的高く、"interest" や"money-fx" などのカテゴリで 50〜60 ポイ

☆☆☆ http://www.cs.cornell.edu/People/tj/svm_light/からダウンロードすることで利用できる。

表 3　相互情報量によって選択された単語

カテゴリ名	単語							
	1 番目	5 番目	10 番目	50 番目	100 番目	500 番目	1000 番目	2000 番目
acq	acquire	company	say	co	usair	hanson	enlarged	escrow
corn	corn	tonne	bushel	share	moscow	reexport	ii	seller
crude	oil	energy	day	mile	pump	meet	commission	angle
earn	vs	be	shrs	today	add	reduce	former	household
grain	wheat	tonne	barley	subsidized	land	iwc	harmonious	tian
interest	rate	england	discount	yesterday	half	debenture	symposium	socialist
money-fx	currency	central	dealer	iraq	circulation	alert	explosive	competitor
ship	ship	tanker	seaman	aegean	turkish	cruzado	continent	arrange
trade	trade	japan	tariff	product	reduce	ambassador	ally	judgement
wheat	wheat	export	soviet	planting	algeria	harkin	cultivation	colorado

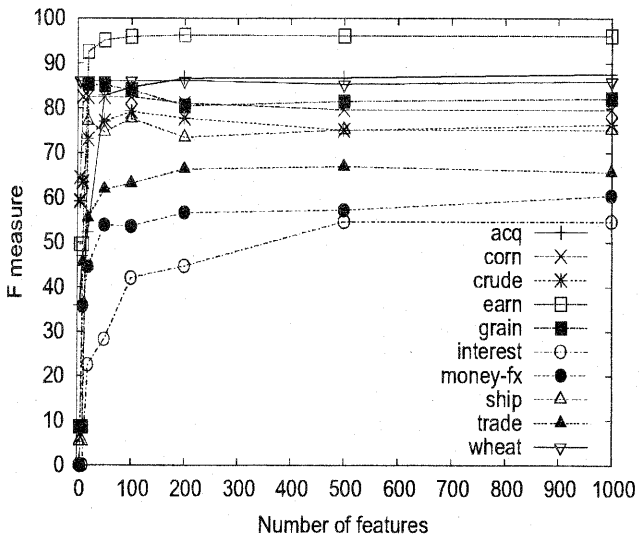


図 5　従来方式による分類結果

ントと F 値が比較的低いことが分かる。

次に、従来手法と第 5 節で述べた多義性軽減方式の比較実験を行った。2 手法の分類結果を比較したものを**図 6**、**図 7**、**図 8** に示す。図中の"the conventional method" は従来手法、"the proposal method" は多義性軽減方式の分類結果を表す。図 6、図 7、図 8 は、実験結果の典型的な 3 パターンとしてそれぞれカテゴリ"interest", "ship", "trade" における結果を示している。

この実験では、多義的な素性の判定条件式 (式 (12)) の閾値 ϵ を 0.300〜0.495 の間で 0.005 刻みで動かした。

図 6 のカテゴリ"interest" では ϵ が下がるについて F 値も減衰している。また、図 7 のカテゴリ"ship" では F 値は ϵ の変動に影響を受けない。今回の実験で唯一 F 値がやや向上したと言えるのが図 8 に示したカテゴリ"trade" である。$\epsilon = 0.330, 0.420$ のときに数ポイントながら従来手法と比較して F 値の向上が確

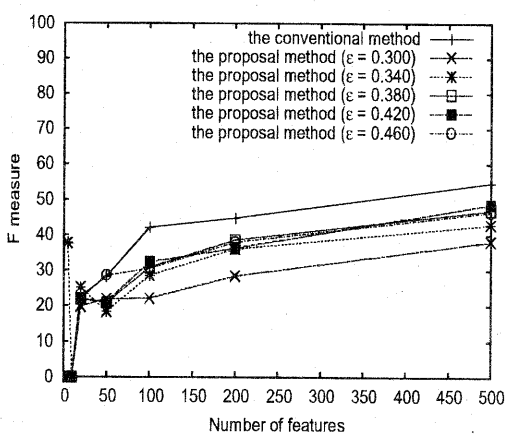


図 6　従来方式と多義性軽減方式の比較
(カテゴリ"interest")

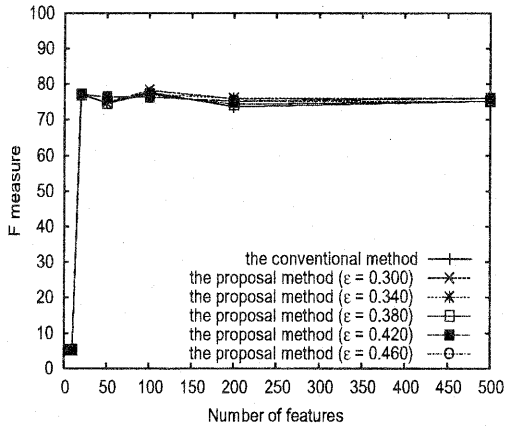


図 7 従来方式と多義性軽減方式の比較
(カテゴリ"ship")

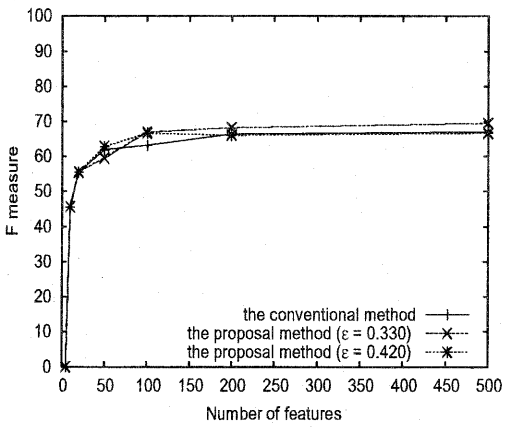


図 8 従来方式と多義性軽減方式の比較
(カテゴリ"trade")

認できた。

7. 考　　察

多義性軽減方式では本実験で分類処理に用いた 10 カテゴリのうち、閾値 ϵ のパラメータサーチによって F 値の向上が確認できたのは"trade" カテゴリのみであり、その他のカテゴリではほとんど F 値に変化がないか、減衰している。"trade" カテゴリにおいても F 値の向上は数ポイント程度であるので、カテゴリ平均で考えると一般的には本提案手法による精度向上は期待できないという結果を示している。

その原因としては、本実験で定義した判定条件式は厳密には本当に複数の意味を持つ多義語を判定できるわけではなく、カテゴリとの共起回数から多義的と予測するに過ぎないため、本当の意味での多義性の解消になっていないということが考えられる。例えば、"grain" カテゴリと"wheat" カテゴリのようにカテゴリ間で意味的に関連性が強いものがある場合、"grain" カテゴリに関連性の強い一義的な単語があったとすると意味的に優秀な素性であるにもかかわらず、"wheat" カテゴリにも関連性が強いためにカテゴリとの共起回数から考えて、多義的とみなされ除去されてしまう。その結果、分類精度に悪影響を与えるものと考えられる。

8. ま と め

本稿では、テキスト分類における単語の多義性の問題に対処すべく、相互情報量によって選択された素性のうち、カテゴリとの共起回数から多義性が強いと考えられるものを除去する多義性軽減方式を提案した。Reuters-21578 コーパスを用いて従来方式と提案手法の比較実験を行ったが、提案手法により精度向上が見られたのは 10 個のカテゴリ中 1 個だけであった。

本手法の拡張の一つの方向として、あらかじめ分類カテゴリ間の関連性を相互情報量などから求めておき、ある素性を含む文書が当該カテゴリ以外のある特定のカテゴリに現れた場合に、分類カテゴリ間の関連度情報を利用して、そのカテゴリと当該カテゴリとの関連度が低い場合のみ多義語であるとみなすといったように、多義性の判定条件式を追加する方式などが考えられる。

参 考 文 献

1) Scott Deer, Susan T. Dumais, George W. Furnas, Thomas K, Laudauer, Richard Harshman, "Indexing by Latend Semantic Analysis", *Journal of the American Society for Information Science*, Vol.41, No.6, pp.391-407, 1990

2) Thorsten Joachims, "Text Categorization with Support Vector Machine: Learning with Many Relevant Features", *Proc. 10th European Conference on Machine Learning*, pp.137-142, 1998

3) Thorsten Joachims, 11 in: "Making Large-Scale SVM Learning Practical", Advances in Kernel Methods - Support Vector Learning, MIT Press, 1999

4) James Tin-Yau Kwok, "Automated Text Categorization Using Support Vector Machine", *Proc. the International Conference on Neural Information Processing*, pp.347-351, 1999

5) Fabrizio Sebastiani, "A Tutorial on Automated Text Categorisation", *Proc. 1st Argentinian Symposium on Artificial Intelligence*, pp.7-35, 1999

6) Vladmir N. Vapnik, "The Nature of Statistical Learning Theory – Second Edition", Springler-Verlag New York, Inc, 2000

7) Yiming Yang, Jan O. Pedersen, "A Comparative

表 4 "ModApte" による訓練データとテストデータの分割

記事の種類	記事数	属性の条件	
訓練記事	9603	LEWISSPLIT="TRAIN"	TOPICS="YES"
テスト記事	3299	LEWISSPLIT="TEST"	TOPICS="YES"
非使用記事	8676	LEWISSPLIT="NOT-USED"	TOPICS="YES"
		LEWISSPLIT="NO"	
		LEWISSPLIT="BYPASS"	

表 5 禁止語の除去、複数形・語形変化への対応、数詞の除去

1.	CC	等位接続詞	→ 除去	24.	SYM	科学・数学記号	
2.	CD	基数	→ 除去	25.	TO	不定詞 to	→ 除去
3.	DT	冠詞	→ 除去	26.	UH	間投詞	
4.	EX	存在の there	→ 除去	27.	VB (VH, VV)	動詞 原形	
5.	FW	外来語		28.	VBD (VHD, VVD)	動詞 過去形	→ VB(VH, VV)
6.	IN	前置詞／従属接続詞	→ 除去	29.	VBG (VHG, VVG)	動詞 動名詞・現在分詞	→ VB(VH, VV)
7.	JJ	形容詞		30.	VBN (VHN, VVN)	動詞 過去分詞	→ VB(VH, VV)
8.	JJR	形容詞 比較級	→ JJ	31.	VBP (VHP, VVP)	動詞 3 人称単数形現在以外	→ VB(VH, VV)
9.	JJS	形容詞 最上級	→ JJ	32.	VBZ (VHZ, VVZ)	動詞 3 人称単数形現在	→ VB(VH, VV)
10.	LS	項目記号	→ 除去	33.	WDT	wh 決定詞	→ 除去
11.	MD	助動詞		34.	WP	wh 代名詞	→ 除去
12.	NN	名詞 単数又は質量		35.	WP$	wh 代名詞 所有格	→ 除去
13.	NNS	名詞 複数	→ NN	36.	WRB	wh 副詞	→ 除去
14.	NP	固有名詞 単数		37.	#	シャープ記号	→ 除去
15.	NPS	固有名詞 複数	→ NP	38.	$	ドル記号	→ 除去
16.	PDT	前決定詞		39.	SENT	ピリオド	→ 除去
17.	POS	所有格語尾	→ 除去	40.	,	カンマ	→ 除去
18.	PP	人称代名詞	→ 除去	41.	:	コロン	→ 除去
19.	PP$	所有代名詞	→ 除去	42.	(	左側括弧	→ 除去
20.	RB	副詞		43.	)	右側括弧	→ 除去
21.	RBR	副詞 比較級	→ RB	44.	”	…重引用符 右 × 2	→ 除去
22.	RBS	副詞 最上級	→ RB	45.	“	…重引用符 左 × 2	→ 除去
23.	RP	不変化詞					

Study on Feature Selection in Text Categorization", *Proc. 14th International Conference on Machine Learning*, pp.412-420, 1997

8) 永田昌明, 平博順, "テキスト分類 — 学習理論の「見本市」(特集 情報論的学習理論とその応用)", 情報処理, Vol.42, No.1, pp.32-37, 2001

9) 平博順, 向内隆文, 春野雅彦, "Support Vector Machine によるテキスト分類", 情報処理学会研究報告, NL-128, pp.173-180, 1998

10) 湯浅夏樹, 上田徹, 外川文雄, "大量の文書データから自動抽出した名詞間共起関係による文書の自動分類", 情報処理学会研究報告, NL-98, pp.81-88, 1993

付　　録

A.1 Reuters-21578 コーパス

Reuters-21578 コーパスはロイター通信社の記事をタグ付けによってカテゴリ分類したテキストコーパスで、インターネット上からダウンロード☆することで利用できる。

本研究では、最も多く使用されている"The Modified Apte(ModApte) Split" と呼ばれる分割方法によって訓練データとテストデータを区別した。この分割方法は、**表 4** に示す <REUTER >タグの属性値によるものである。

A.2 tree-tagger

tree-tagger は入力文を品詞ごとに解析し、品詞のラベル付けを行う品詞付与ツールで、インターネット上からダウンロード☆☆することで利用できる。

第 6.1 節で行った前処理の詳細を**表 5** に示す。表中で「→ 除去」とあるのは禁止語の除去を意味し、「→ (英語略字)」とあるのは単語の複数形・語形変化への対応を意味する。数詞の除去では、表 5 で「CD 基数」という品詞が付与されたものを除去した。

☆ ダウンロード先の URL は、http://www.daviddlewis.com/resources/testcollections/reuter21578/である。

☆☆ ダウンロード先の URL は、http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/DecisionTreeTagger.html である。

正誤表

85 ページ（1 ページ目）英語タイトル

（誤） Text Classification with a Polysemy Condidered Feature Set

↓

（正） Text Classification with a Polysemy Considered Feature Set