同義語，多義語の考慮による文書分類の精度向上

MIURA, Takao / 塩谷, 勇 / UEJIMA, Hiroshi / 三浦, 孝夫 / SHIOYA, Isamu / 上嶋, 宏

(出版者 / Publisher)
電子情報通信学会情報・システムソサイエティ

(雑誌名 / Journal or Publication Title)
電子情報通信学会論文誌. D-I, 情報・システム, I-情報処理 / The transactions of the Institute of Electronics, Information and Communication Engineers. D-I

(号 / Number)
2

(開始ページ / Start Page)
137

(終了ページ / End Page)
144

(発行年 / Year)
2004-02-01

同義語，多義語の考慮による文書分類の精度向上

上嶋　宏[†]　三浦　孝夫[†]　塩谷　勇[††]

Improving Text Categorization by Synonym and Polysemy

Hiroshi UEJIMA[†], Takao MIURA[†], and Isamu SHIOYA[††]

あらまし　本論文では，同義語，多義語を用い，単語のもつ意味のあいまい性を考慮した文書分類を提案する．本論文での文書分類は，シソーラスと単語がもつ複数の意味の使用頻度を用いる．これらを考慮することにより単語のもつ意味のあいまい性を排除し，分類精度を向上させる．本論文ではワードネットを用いて実験を行い，82%を超える高い分類正解率を得たことを示す．

キーワード　テキスト分類，テキストマイニング，単純ベイズ法

1. まえがき

本論文では，ベイズ学習に基づく文書分類法を提案する．ここでは同義語，多義語を利用し，単語の意味を考慮して，分類の一貫性の向上や分類精度の向上を目的としている．

文書分類（テキストカテゴライゼーション）とは，文書をそれが属するカテゴリーへ割り当てることをいう．この文書分類では，構造化されていないテキストデータを扱う．そのためには，ある単位での情報の抽出が必要である．文書分類では，“set of words” や “bag of words” と呼ばれる文書を単語の集まりとして考えるのが一般的である [7]．文書 x は重み $x_1, \ldots, x_d$ をもった単語の連続として，ベクトル $x = (x_1, \ldots, x_d)$ と表現される．ここで d は文書集合内で出現した単語の数である．句単位での文書分類は単語単位に比べて良い性能は示さない [3], [7]．通常の文書分類では，単語のもつ意味などは考慮せず，単語を単に記号的に扱う．また，文書内での単語の出現順序は分類に重要な意味をもたない．

通常，文書内には同じ意味をもつ複数の単語（同義語）や，複数意味をもつ単語（多義語）が存在する．同義語や多義語を含む文書に対して文書分類を行うと，異なるベクトル表現として処理される結果，分類の一貫性の低下や分類精度の低下が生じる可能性が高い [9]．

本論文では，ベイズ学習を用いた文書分類法を提案する．一般に，ベイズ手法に基づく場合，高次元の入力により過学習が起こりやすいとされているが，本論文で用いる次元縮小法はこの問題を顕在化させない．本論文の目的は同義語，多義語を考慮する文書分類法を提案することにある．ワードネットを用いた実験を行い，その有効性を示す．

ワードネットを用いた関連研究としては福本ら [10] が代表的である．ここでは，ワードネットにより，同義語クラスの上位関係を利用する手法を提案している．ワードネット内で，名詞は “entity” や “location” などの 25 種類の意味クラスに分類され，意味クラスごとに階層構造を形成している．各ノードは synset と呼ばれる同義語の集合である．この手法では，意味クラスごとに文書中の単語が unique beginner と呼ばれる root ノードからどの深さまでの synset を用いて表現するかのパラメータを設定し，各カテゴリーの分類規則ごとにそれぞれの最適なパラメータを求め分類を行っている．しかし，単語が多義の場合は，各意味を示す階層構造のうち，unique beginner から単語までの深さが最大である階層構造を使用している．このため，多義語による精度低下の問題を解決していない．

また，Rodriguez らは，文書分類を Rocchio 法（関連フィードバック）や Widrow-Hoff アルゴリズムによる機械学習法で行うときに，ワードネットを利用す

[†] 法政大学工学研究科電気工学専攻，小金井市
Dept. of Elect. & Elect. Engr., Hosei University, 3–7–2 Kajino-cho, Koganei-shi, 184–8584 Japan
[††] 産能大学経営情報学部，伊勢原市
Department of Management and Information Science, Sanno University, 1573 Kamikasuya, Isehara-shi, 259–1197 Japan

る方法を提案している [4]．しかし，本論文の提案と異なり，“カテゴリーを構成する単語” に対して同義語を展開し，テストに使う Reuters 記事そのものは同義語展開していない．

2. ではベイズ学習と文書分類について述べる．**3.** ではワードネットについて同義語，多義語に重点を置いて述べる．**4.** で同義語，多義語を考慮した文書分類について述べ，**5.** では実験結果を示し，**6.** では結論を示す．

2. ベイズ学習による文書分類

2.1 文書分類

文書分類とは，文書をその内容に応じて，あらかじめ与えられたいくつかのカテゴリーに自動的に分類することである．この技術は，文書検索，分類，ルーチング，クラスタリング，Web ページのディレクトリ構造の作成，電子メールなどのフィルタリング等に有用で，これらの作業を人手により行う場合に比べ，はるかに時間やコストを削減できる．

文書分類には，本来自然言語で書かれていることによるあいまいさ，高次元データゆえの粗データ表現と計算量，個々の文書固有の表現方法の差異など多くの問題がある．特に，単語が同義語，多義語のようなあいまい性をもっており，実際に単語がどの意味として働いているかを理解することは困難で，分類精度を下げている要因と考えられる．本論文では教師付きベイズ学習を用いる．ベイズ学習は分類手段として広く使用されている手法である．教師付き学習とは，人手によりカテゴリーを割り当てられたカテゴリーが既知のデータ（訓練集合）から訓練集合の中に潜むパターンを学習し，そのパターンを用いて分類規則を作成し，その分類規則により未知のデータが属するカテゴリーを予測するものである．

2.2 ベイズ学習

ベイズ学習は文書分類に有用な手段の一つである．ベイズ学習は「興味の対象となっている物理量は確率分布によって支配されており，最適な意志決定は，訓練データの確率を推論することで達成される」という考え方に基づいており，特徴は「最終的な仮定成立の確率計算に事前知識を使用する」という点にある．事前知識とは訓練データにおいての，仮説が成り立つかどうかの事前確率，訓練データにおいて，ある仮説が成り立つとした場合の，他の仮説が成り立つとする条件付確率の二つからなる．

ベイズ学習の利点は，仮説が成立する確率を明示的に扱い，確率的な仮説の表現を許すことにある．すなわち出力結果が真か偽ではない点が挙げられる．他方，欠点としては，事前確率分布のために多くの初期知識を必要とし，計算コストが大きいことが挙げられる．しかし，文書分類を行うとき，あらかじめ分類するためのデータはそろっており，量も十分にある．また，計算機の性能も計算量に対し十分であることが多い．

教師付きベイズ学習による文書分類の場合，訓練データから以下の四つを求める [3]．

1: $P(c_k)$　文書がカテゴリー $c_k \in C$ に属する事前確率

2: $P(x)$　何の事前知識をもたないときに，訓練例において文書ベクトル x を観測する確率

3: $P(x|c_k)$　カテゴリー c_k に属するときに，文書ベクトル x を観測する条件付確率（ゆう度）

4: $P(c_k|x)$　文章ベクトル x が観測されたという条件下でカテゴリー c_k に属するという事後確率

ゆう度 $(P(x|c_k))$ と事前確率 $(P(c_k))$ から結合確率 $P(c_k)P(x|c_k)$ を得る．結合確率を正規化し，以下のようなベイズルールを得る．

$$P(c_k|x) = P(c_k) \times \frac{P(x|c_k)}{P(x)} \tag{1}$$

カテゴリー集合 C の要素 c_k で，$P(c_k|x)$ の値が最も大きいもの（最大事後確率を与えるカテゴリー）は，以下のように示される．

$$\begin{aligned} c &= MaxArg_{c_k} P(c_k|x) \\ &= MaxArg_{c_k} P(c_k) \times \frac{P(x|c_k)}{P(x)} \end{aligned} \tag{2}$$

ベイズルールでは，最大事後確率を取るカテゴリー c_k を文書 x が属するカテゴリーとすることで予想される分類エラーの数が最少になると考える．したがって，ベイズルールでの分類規則の作成は，訓練データから $P(x|c_k)$，$P(c_k)$，$P(x)$ の値を求めることである．

2.3 単純ベイズ分類

ベイズルールでは $P(c_k)$，$P(x)$，$P(x|c_k)$ の確率を組み合わせ $P(c_k|x)$ を求める．$P(x)$，$P(x|c_k)$ で出現する文書ベクトル $x = (x_1, \ldots, x_d)$ は，ほぼすべての文書で異なり，極めて多数になることを想定する必要がある．このため本論文では単純（naive）ベイズ分類を用いる．単純ベイズ分類では，ベクトル x を，すべての c_k に対して各要素 x_j を（同時的でなく）独立事象とみなすことで，計算量の削減を行う [5]．

表 1 $P(c_k|x_j)$ の上位の単語とその値
Table 1 $P(c_k|x_j)$ and words.

カテゴリー	"gas"		"gold"		"fuel"	
1	0.825	gasoline	1.0	gold	1.0	fuel
2	0.6	oil	0.5	ounces	0.6923	pct
3	0.6	mln	0.4681	mine	0.6923	oil
4	0.5	crude	0.4468	pct	0.6154	dlrs
5	0.475	pct	0.4042	ton	0.5385	petroleum
6	0.425	year	0.4043	ounce	0.5385	corp
7	0.425	petroleum	0.3936	year	0.5385	barrel
8	0.35	fuel	0.3936	company	0.4615	prices
9	0.325	distillate	0.3723	mln	0.4615	crude
10	0.3	energy	0.3194	ore	0.4615	barrels
11	0.3	barrel	0.3085	production	0.3846	mln
12	0.275	stok	0.3085	mining	0.3846	heavy
13	0.275	product	0.3085	dlrs	0.3846	energy
14	0.25	refinery	0.2872	short	0.3846	effective

$$P(x|c_k) = \prod_{j=1}^{d} P(x_j|c_k) \tag{3}$$

この仮定により $P(x_j|c_k)$ は比較的少ないパラメータで構成され，$P(c_k|x)$ は以下のように表すことができる．

$$P(c_k|x) = P(c_k) \times \frac{\prod_{j'=1}^{d} P(x_j|c_k)}{P(x)} \tag{4}$$

また，本論文では 2 項独立モデル（BIM）を用いる．2 項独立モデルとは，文書ベクトル $x = (x_1, \ldots, x_d)$ のすべての単語の重み x_d の値を文書内で単語 x_d が現れたときは 1，現れなかったときは 0 とするものである [5]．2 項独立モデルを使うことにより，$P(x_j|c_k)$ は以下のように表すことができる．

$$P(x_j|c_k) = p_{jk}^{x_j}(1-p_{jk})^{1-x_j} = \left(\frac{p_{jk}}{1-p_{jk}}\right)^{x_j}(1-p_{jk}) \tag{5}$$

ここで $p_{jk} = P(x_j = 1|c_k)$ である．式 (1) と式 (5) より対数をとることにより，以下の式を得る．

$$\log P(c_k|x) = \log P(c_k) + \sum_{j=1}^{d} x_j \log \frac{p_{jk}}{1-p_{jk}} + \sum_{j=1}^{d} \log(1-p_{jk}) - \log P(x) \tag{6}$$

この 2 項独立モデルを利用することで，単純化されたベイズルールにより文書分類を行うことができる．2 項独立モデルでは，文書内での単語の出現回数や文書を占める割合を考える必要はなく，単語の出現の有無のみを調べればよい．

2.4 分類規則の次元縮小

文書分類において，高い次元 d をもつ文書ベクトル x は，粗な表現を生むことが多く，また多量の計算量を発生させるため問題である．通常，文書分類では，この問題解決のために x の次元 $|d|$ のサイズを縮小する試みを行う [7]．すなわち次元縮小とは，分類規則の単語を減らし，分類に使用する単語を限定したベクトルを生成することである．次元縮小は計算量の減少や，訓練データの過学習を避けるのに有用である．しかし次元縮小は用語を削除するので，潜在的に有用な情報が削除され，精度を低下させる可能性がある．次元縮小には，局所的次元縮小と大域的次元縮小の二つの方法が提案されている [7]．局所的次元縮小とは，それぞれのカテゴリー c ごとに異なった用語セット $\hat{d}_c$ を使用し，分類を行う．大域的次元縮小とは，すべてのカテゴリー上で同一の用語セット $\hat{d}$ を使用する．

本論文では "DIA asociation factor" [2], [7] と呼ばれる用語選択による局所的次元縮小を用いる．この方法は $P(c_k|x_j)$ の値の大きな x_j だけを分類規則に用いる．文書分類に使用する用語数としては局所的次元縮小の場合，$\hat{d}_c$ のサイズは 10 ないし 50 程度がよいとされている [7]．Reuter コーパス（新聞記事）における "gas" など三つのカテゴリーごとの $P(c_k|x_j)$ の大きな単語とその値を表 1 に示す．

3. ワードネットによる同義語，多義語の利用

本論文では語彙参照のためワードネットを用いる [1]．ワードネットはオンライン形式で語彙を参照することが可能であり「英語用語彙データベース（lexical

The noun human has 2 senses (first 2 from tagged texts)

1. (7) person, individual, someone, somebody, mortal, **human**, soul -- (a human being; "there was too much for one person to do")
2. (5) homo, man, human being, **human** -- (any living or extinct member of the family Hominidae)

The adj human has 3 senses (first 3 from tagged texts)

1. (47) **human** -- (characteristic of humanity; "human nature")
2. (20) **human** -- (relating to a person; "the experiment was conducted on 6 monkeys and 2 human subjects")
3. (15) **human** -- (having human form or attributes as opposed to those of animals or divine beings; "human beings"; "the human body"; "human kindness"; "human frailty")

図 1 ワードネットによる検索結果（human）
Fig. 1 Examining a word "human" to WordNet.

database for the English language)」とも呼ばれる．

ワードネットは特定の分野の知識をもたず，一般的な知識をもつ．本論文では，特定のカテゴリーに特化して優れた性能を示すのではなく，すべてのトピックに関して均等な性能を示す文書分類を対象としており，ワードネットを使用することは妥当である．

ワードネットはシソーラス辞書とは異なり，辞書を構成する基本単位として synset（synonym set）を用いる．synset とは同義語の集合であり，これより単語の意味や概念の記述及び分類を行う．また synset には反義語，上位語，下位語，部分語，全体語などが定義されている．

本論文では，複数の単語が同じ概念を定義しているとき，これを同義語という．例えば "lofty" は "high" の同義語である．多義語は複数の意味をもつ単語をいう．例えば，"picture" は "movie"，"figure"，"photo"，"illustration" を意味する．当該語がどの意味を表現するかを考慮しないならば，明らかに同義語や多義語は検索や分類性能を低下させる．

図 1 に "human" をワードネットで検索した結果を示す．これより頻度や意味，同義語の単語の集合などを得る．またワードネットは品詞ごとに結果を出力する．例えば，"human" は名詞と形容詞の両方に使われ，名詞としては "person" と "human" の二つの意味をもつ．図 2 では "person" が "human"，"body"，"grammatical category" の意味として使われることを示す．通し番号の後ろの () 内の数字は，ワードネットの作成に使われた文書集合で使用された意味の数を表し，意味の使用頻度として扱うことができる．例えば，図 1 より "human" は文書集合内で $7+5+\ldots=94$ 回出現しており，そのうち 7 回は "person" という意味で使用されていたことを示す．また，使用頻度の低い意味は頻度が付けられていない．

The noun person has 3 senses (first 2 from tagged texts)

1. (7229) **person**, individual, someone, somebody, mortal, human, soul -- (a human being; "there was too much for one person to do")
2. (11) **person** -- (a person's body (usually including their clothing); "a weapon was hidden on his person")
3. **person** -- (a grammatical category of pronouns and verb forms; "stop talking about yourself in the third person")

図 2 ワードネットによる検索結果（person）
Fig. 2 Examining a word "person" to WordNet.

表 2 単語がもつ意味番号
Table 2 Synset identifiers.

	"human"	"person"
名詞	1:5303	1:5303
	2:2130996	2:4465544
形容詞	1:324678454	
	2:2634237	
	3:2634331	

これらの例が示すように，synset は複数の単語を含む場合があり，それらの単語は同義語である．また，複数の意味をもつ単語は，図 1 の単語 "human" のように複数の synset をもつ．また synset は固有の識別番号をもつ．

"human"，"person" はそれぞれ，synset 番号 5303 をもつ．この synset 番号は図 1，図 2 での第 1 番目の意味に対応している．すなわち，"human" と "person" は 5303 という同じ意味をもつ．これらの単語がもつ識別番号を表 2 に示す．

4. 単語の意味を考慮した文書分類

4.1 同義語を考慮した文書分類

通常の文書分類では，単語の意味を考慮せず，単に記号的に解釈する．実験ごとに，訓練データ，テストデータの選択方法や件数，ストップワード（停止語）の選択等が多少異なるので，分類精度は一概には比較できないが，同義語，多義語を考慮しないベイズルールでの Reuter コーパスの分類精度は約 74～79.5%であることが知られている [7]．実際，筆者らが行った実験と同じ条件でも 79.26%を得た．

同義語を考慮した文書分類の研究も行われている [9]．例えば，「生徒」という意味の単語 "student" と "pupil" の両方が "school" というカテゴリーの文書集合で出現していた場合，この二つの単語を同じとみなすことで，これらの単語の "school" カテゴリーでの出現確率を増やすことができ，重要性を増すことができる．

表 3　分類規則にワードネットを適用
Table 3　$P(c_k|x_j)$ applied to WordNet.

カテゴリー	gas			trade		
1	0.825　gasoline			0.930667　trade		
	意味番号	出現数	頻度	意味番号	出現数	頻度
	12402140	1	1	831317	394	0.502
				6962272	124	0.158
				844555	107	0.136
				454930	94	0.12
2	0.6　oil			0.586667　year		
	12653557	11	0.647	12867473	832	0.962
	3347615	5	0.294		·	
		·			·	
		·			·	
3	0.6　mln			0.477333　billion		
	mln	0	1	11604853	1	1

このように，意味を考慮して重みを与えることで，カテゴリーに関係が深い単語の重要度が増加する．また「生徒」を “pupil” と表現する文書が少なく，単語 “student” のみが “school” カテゴリーで重要性をもっている場合も，「生徒」を pupil と表現している文書に対してより多くの重要性を付与することができる．

しかし，同義語だけを考慮して文書分類を行った場合，使用頻度が低い同義語を扱うときに分類精度が低下することがある．例えば，文書内で “man” という多義の単語が出現したとき，この “man” は主に “人間 (human) ” という意味で使われているが，“チェスの駒” という意味ももつ．このため “piece” という単語を同義語として扱い，この単語に “man” と同様の重要性を与えてしまう．通常これらの単語は全く違う意味で使われているので，分類精度の低下が生じる．

4.2　同義語，多義語と頻度を考慮した文書分類

本研究では単語の同義性だけでなく，単語がもつ多義性と，その複数の意味の使用頻度を利用した文書分類を提案する．ここで使用頻度とは，ワードネットにより得ることのできる意味の出現回数の総和に対する比率である（図 1，図 2）．2 項独立モデルでは，特徴ベクトル $x = (x_1, \ldots, x_d)$ の値を，すべて 0 か 1 で表現する．しかし，本研究では前述の多義語の使用頻度の低い意味による問題を解決するために，単語の意味の使用頻度を利用して文書ベクトル $x' = (x'_1 \ldots, x'_d)$ の重み x'_j を決定する $(0.0 \leq x'_j \leq 1.0)$．これにより式 (5) を以下のように変更する．

$$P(x_j|c_k) = p_{jk}^{x'_j}(1 - p_{jk})^{1-x'_j} \tag{7}$$

$$= \left(\frac{p_{jk}}{1 - p_{jk}}\right)^{x'_j}(1 - p_{jk}) \tag{8}$$

表 4　最大の出現頻度をもつ synset
Table 4　Maximum frequency synset.

oil	12653557
gasoline	12402140
trade	831317
craft	454930
student	8734996
pupil	8734996

このため本論文での単語の重みは，もはやバイナリー値ではないが，文書内での単語の出現回数に基づくものでもない．重み付けの具体的な方法を以下に示す．

本論文では分類規則にワードネットを適用し，意味の出現頻度により文書ベクトルの重みを与える．しかし，一つの単語は複数の意味をもつので，ワードネットにより多義語をその意味に分解することで，文書ベクトルの次元の増加が考えられる．このため本研究では，“DIA asociation factor” による局所的次元縮小を行った分類規則（表 1）にワードネットを適用する（表 3）．これにより単に多義語を意味に分解して分類規則を作る場合に比べ，次元の増加を抑制することができる．また，通常の文書分類と同じ分類規則を利用する．更に本論文では，テストデータにもワードネットを適用する．テストデータ内の単語を最大出現頻度をもつ synset へ置き換え，それぞれの単語の主要な意味のみを扱う（表 4）．特に，意味の頻度が分散している単語の分類への重要度は低いとみなし，最大頻度が低い単語の重要性は減少させる．

例えば，“oil is gasoline” という文を考える．“gas” カテゴリーについては，表 3 より “oil”（最大頻度 id 12653557）が 0.647 回，“gasoline”（最大頻度 id

12402140）が1回出現するとみなす．他に分類規則内のsynset idの出現がなかった場合，文書ベクトル(oil,gasoline,...)は(0.647,1,...)となる．表3の“mln”のようにワードネットに生じていない単語が分類規則にある場合，頻度を1とし，通常の文書分類と同様，出現するか否かの0か1により重みづけする．

最大頻度の意味が同じ単語，例えば“student”と“pupil”の最大頻度の意味番号は両方とも8734996で同じである．テストデータは表4のように最大頻度のsynsetに置き換えるので，テストデータでこの二つの単語が出現した場合は全く同じものとみなす．

類似の意味をもつ単語がテストデータ内で出現するとき，例えば文書内に“trade”と“craft”が出現した場合，それぞれの最大頻度の意味番号はtrade:831317, craft:454930である．本論文では表3の“trade”カテゴリーに関して，“trade”の831317により“trade”が0.5019回出現し，また“craft”の454930により同じく“trade”が0.1197回出現するとみなし，“trade”の出現回数x'_{trade}は$x'_{trade} = 0.5019 + 0.1197 = 0.6216$回とする．本論文ではテストデータの単語がもつ最大頻度の意味と分類規則の単語がもつ意味を比較し，同じ意味をもつ場合，分類規則での意味の頻度を単語の類似度とする．この類似度により単語の出現回数を重みづけ，前述の意味のあいまい性を解決し，より精度の高い分類を得ることができる．

分類規則にワードネットを適用するとき，その意味番号と使用頻度を使用する．しかし，そのすべての意味を使用するのではなく，表5のように，頻度の高い上位S個の意味を使用する．これにより，上位頻度の意味の重みが増加し，意味の頻度が分散している単語にも比較的重要性をもたせることができ，前述の多義語の同義語の問題をより的確に扱うことができる．しかし，使用する意味を限定することは，使用頻度がS番目以下の意味を無視することであるため，

表5 S=ALLとS=2の場合の頻度の変化
Table 5 “human” in S=ALL and in S=2.

human		
S=ALL		
意味番号	出現数	頻度
2634237	47	0.5
2634331	20	0.2128
1220852	15	0.1596
5303	7	0.0745
	⋮	

(a)S=ALLの場合の頻度

human		
S=2		
意味番号	出現数	頻度
2634237	47	0.7015
2634331	20	0.2985

(b)S=2の場合の頻度

本実験ではS=1, 2, 3, 5とすべての意味を使用する(S=ALL)計5パターンを比較する．以下でこれをS値と呼ぶ．

5. 実験と評価

5.1 実験に使用するコーパス

本実験にはReuter21578を用い，ApteModに従い訓練文書集合とテスト文書集合を作成する．本実験でのReuterの設定を以下に示す．この設定も以下のように標準的手法に従う[3]．

- Reuterの“TOPICS”を分類するカテゴリー，すなわち分類の答として扱う．
- “TOPICS”をもたない記事は使用しない．
- 訓練集合で出現回数が5回以下の“TOPICS”にしか属さない記事は削除する．

その結果，本実験で使用する訓練文書集合は7907件，テスト文書集合は3081件となり，総カテゴリー数は73となる．またストップワードについてはあらかじめ削除しておく．本論文では，通常のベイズ学習による文書分類（以下，通常の手法）と，提案手法での文書分類による文書ごとの重み付けの変化や精度の変化を比較する．このため，しきい値の設定の必要がない文書主導の単一ラベル文書分類を行う．分類精度は正解率により評価する．

$$\text{正解率} = \frac{\text{正解した文書数}}{\text{全文書数}} = \frac{\text{正解した文書数}}{3081}$$

文書が複数のカテゴリー（答）をもつ場合は，そのうちのどれか一つに当てはまれば正解とする．文書分類では精度と再現率による評価が一般的であるが，本実験では文書を必ず一つのカテゴリーに割り当てているため，精度と再現率，F値(注1)は，正解率と同じになる．

5.2 実験手順

2.1で述べたように，教師付き学習での文書分類の手順は一般に(1)訓練データ，テストデータの作成，(2)訓練データから分類規則の作成，(3)分類規則をテストデータに適用であり，最後に，(4)その結果から性能評価を行う．(1)，(2)は通常の手法，提案手法のどちらの場合も同じである．提案手法では通常の手法により作成された分類規則にワードネットを適用する．また3で使用するテストデータにもワードネット

(注1)：文書分類の性能評価によく使われる値．精度(r)と再現率(p)から求める$F = \frac{2rp}{r+p}$．

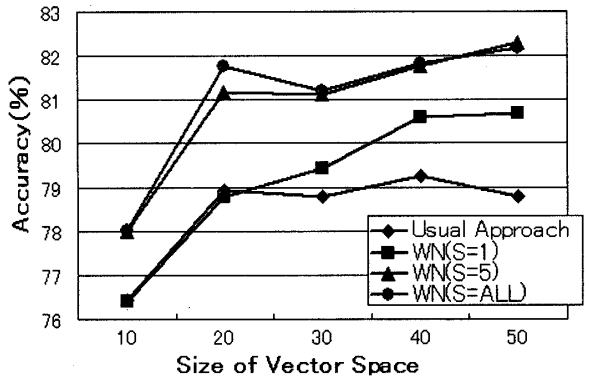


図 3 分類結果
Fig. 3 Results.

表 6 通常の手法による分類結果
Table 6 Results by usual approach.

分類規則の次元（語）	10	20	30	40	50
正解率 (%)	76.43	78.94	78.77	79.26	78.77

表 7 提案手法による分類結果
Table 7 Results by proposed approach.

分類規則の次元（語）	10	20	30	40	50
ワードネット (S=1)	76.40	78.77	79.42	80.59	80.69
ワードネット (S=2)	77.47	80.98	81.08	81.82	81.99
ワードネット (S=3)	77.83	81.17	81.08	81.89	82.02
ワードネット (S=5)	77.99	81.17	81.14	81.76	82.31
ワードネット (S=ALL)	78.03	81.78	81.21	81.82	82.18

を利用し最大頻度の意味のみを考慮する．本実験では縮小した次元（語）数は 10，20，30，40，50 の 5 種類，また S 値は 1，2，3，5 と ALL の 5 種類の分析を行う．

5.3 実験結果

通常の手法による実験結果を図 3，表 6 に示す．結果は **4.1** で述べたように最高で 79.26%の正解率を示した，このときの分類規則の次元数は 40 である．

次に提案手法による分類結果を図 3，表 7 に示す．この結果では S=5，分類規則の次元数が 50 のときに 82.31%の正解率を示した．

通常の手法による分類で最高の正解率を示した分類規則は次元数 40 であった．これをこのままを使用した場合の提案手法での結果は，S=3 のときに 81.89%の正解率を示した．

5.4 考　　察

図 3 及び表 7 から分かるように，提案手法では正解文書数が 2442 件から 2536 件と増加し，正解率が約 3.04%向上する．誤分類された文書数は 639 件から 545 件と約 15%減少している．この結果より，同義語，多義語とその頻度の考慮により，より正確に文書分類を行うことができる．

表 8 通常の手法と提案手法の正解数，間違い数
Table 8 Comparison of usual approach and proposed approach.

		通常の手法	
		正解	間違い
提案手法	正解	2339	197
	間違い	103	472

提案手法では，すべての S 値において，通常の手法の性能を上回った．また，$S \geq 2$ の場合は，ほぼ同じ値を示した．S=1 の場合，多義性を扱っておらず，最大頻度の synset しか考慮しないことに注意したい．この結果，単語の同義性と多義性を同時に考慮することで，より正確な分類が行えるようになる．本実験ではすべての意味を考慮した場合（S=ALL）の場合でも，大きな精度の低下は得られなかった．しかし，S=5 の場合に最大の精度を示したため，使用する意味を限定することは有効である．有効な結果を導くために，意味の使用頻度に対するしきい値として S 値を設定するなど，詳密な限定方法が必要である．

通常の手法と提案手法の正解文書の変化を表 8 に示す．

3081 の文書中 2339 の文書に関しては，双方の手法のどちらでも正確に分類された．また，通常の手法では正確に分類され，提案手法により間違って分類された文書が 103 件あった．この内容を分析すると，提案手法により間違って割り当てられたカテゴリーは，正解カテゴリーに類似したものが多数あることが分かった（例えば正解のカテゴリーが “trade” の文書を “yen” カテゴリーに誤分類）．また正解カテゴリーも必ず上位にランクされていた．この変化の原因としては，上記分析より (1) ワードネット作成に使われたコーパスと Reuter との意味頻度の違いなどコーパス間の相性によるもの (2) 分類精度は手作業により割り当てられたカテゴリーに依存する (3) 手作業によるカテゴリー分類の基準のあいまいさ等の原因が考えられる．

双方の手法のどちらでも正しく分類できない 472 件のデータに関しては (1) 他の文章と内容がかけ離れている (2) 人手によるトピック割当の問題 (3) ベイズルールによる分類の限界等の原因が考えられる．

6. むすび

本研究では単語を単に記号的に認識する通常の文書分類とは異なり，同義語，多義語と，その意味の使用頻度を考慮する文書分類を行った．Reuter コーパスを使用した実験により評価した結果，従来の文書分類よりも有用性を示した．実際に，間違って分類される文書数が15%減少し，正解率も82.3%となることから，本手法の有用性を確認した．

今後はベイズ学習以外の学習アルゴリズムや，教師なしのクラスタリングへの本手法の適用を行う予定である．また，単語をワードネットで調べることは非常に大きな計算時間を要するので，ワードネットの高速化や，技術文書などの，より専門的な文書を対象にした性能評価などが今後課題となる．

謝辞 本論文の査読者の皆様から，的確かつ有意義な意見を頂いた．本研究の一部は文部科学省科学研究費補助金（課題番号 14580392）の支援による．

文　献

[1] G.A. Miller, R. Beckwith, C. Fellbaum, D. Gross, and K. Miller, "Introduction to WordNet – An on-line lexical database," J. Lexicography, vol.3, no.4, pp.235–244, 1990 (revised 1993, Princeton University).

[2] N. Fuhr and C. Buckley, "A Probabilistic Learning Approach for Document Indexing," ACM TOIS 9-3, pp.223–248, 1991.

[3] D.D. Lewis, Representation and Learning in Information Retrieval, Ph. D. Thesis, Department of Computer Science, University of Massachusetts, 1992.

[4] M.B. Rodriguezd, J.M. Gomez-Hidalgo, and B. Diaz-Agudo, "Using WordNet to complement training information in text categorization," Proc. Recent Advances in Natural Language Processing, pp.150–157, 1997.

[5] D.D. Lewis, "Naive (Bayes) at forty: The independence assumption in information retrieval," Proc. ECML-98 (10th European Conference on Machine Learning), 1998.

[6] Y.H. Li and K. Jain, "Classification of text documents," Comput. J., vol.41, no.8, pp.537–546, 1998 (revised 1993, Princeton University).

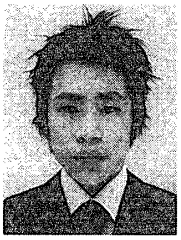
[7] F. Sebastiani, "Machine learning in automated text categorization," Proc. ACM Computing Surveys, vol.34, no.1, pp.1–47, 2002.

[8] 市村由美，長谷川隆明，渡辺　勇，佐藤光弘，"テキストマイニング—事例紹介，" 人工知能誌，vol.16, no.2, pp.192–200, 2001.

[9] 那須川哲哉，河野浩之，有村博紀，"テキストマイニング基盤技術，" 人工知能誌，vol.16, no.2, pp.201–211, 2001.

[10] 福本文代，鈴木良弥，"WordNet の同義語クラスとその上位関係を利用した文書の自動分類，" 情処学論，vol.43, no.6, pp.1852–1865, 2002.

（平成 15 年 6 月 2 日受付，9 月 23 日再受付）

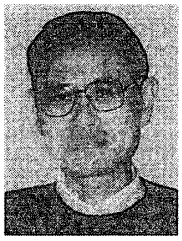


上嶋　宏

法政大学工学研究科電気工学専攻，現在，修士課程在学中．テキストマイニング，トピック検出，TDT，オントロジーなどの分野に興味をもつ．

三浦　孝夫　（正員）

京大・理卒．工博（東大）．現在，法政大学工学部情報電気電子工学科教授．データモデル，知識表現，演繹データベース，複合オブジェクトなどの分野の研究に従事．ACM 会員．著書に "データモデルとデータベース"（全 2 巻，サイエンス社）．



塩谷　勇　（正員）

東京電機大学大学院修士課程了．現在，産能大学経営情報学部教授．時系列モデルの同定，論理プログラミング，グラフ文法，論理データベースの研究に従事．ACM 会員．