同義語、多義語の考慮によるテキストカテゴライゼーションの精度向上

上嶋　宏† 三浦 孝夫† 塩谷 勇††

† 法政大学 工学部 電気電子工学科 〒 184–8584 東京都小金井市梶野町 3-7-2
†† 産能大学 経営情報学部 〒 259–1197 神奈川県伊勢原市上粕屋 1573
E-mail: †{c9943013,miurat}@k.hosei.ac.jp, ††shioya@mi.sanno.ac.jp

あらまし 本稿では、同義語、多義語を用いることによるテキストカテゴライゼーション (文書分類) を提案している. 本稿での文書分類は、シソーラスと意味の使用頻度用いる. これにより得られる同義語, 多義語を利用して単語の重み付けをすることにより精度を向上させる. シソーラス辞書にはワードネットを用い, 実験により検証する.
キーワード テキスト分類, テキストカテゴライゼーション, テキストマイニング, データマイニング

The peformance improvement of Text Categorization by synonym and polysemy

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

† Dept.of Elect.& Elect. Engr., HOSEI University 3-7-2, KajinoCho, Koganei, Tokyo, 184–8584 Japan
†† Department of Management and Information Science, SANNO University 1573, Kamikasuya, Isehara city, Kanagawa 259–1197 Japan
E-mail: †{c9943013,miurat}@k.hosei.ac.jp, ††shioya@mi.sanno.ac.jp

Abstract We propose Text Categorization with consideration of synonym and polysemy.This TC uses the weight of the meaning of synonym and polysemy change by the thesaurus and the frequency.We improve the performance of TC by this method. And we estimate the performance of Text Categorization by experiment.
Key words Text Categorization,Data Mining,Text Mining

1. 前 書 き

テキストカテゴライゼーション（文書分類）とは, 文書をそれが属するカテゴリ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. 構造化されていないデータを扱うには, 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, 単語単位により情報を抽出する [3].

通常の文書分類では, 単語の持つ意味などは考慮せず, 単語を単に記号的に扱う. 通常, 文書内には複数意味を持つ（多義）単語が存在したり, 複数の単語が同じ意味を持つ場合（同義）がある. 通常の文書分類のように単語の同義性, 多義性を考慮せずに文書分類を行うと, 多義語や同義語が多く存在する文書では分類の一貫性の低下や分類精度の低下が考えられる [5]. 故に, 文書分類を同義語, 多義語を利用し, 単語の意味を考慮して行うことにより, 分類の一貫性の向上や, 意味を考慮しなければ識別できない文書の分類など, 分類精度の向上を考えることができる.

本稿では, 単語の意味を単に記号的に扱うのではなく, 同義語と, 多義語を考慮した文書分類を提案する.

2 章ではベイズ確率論による文書分類について述べる.3 章ではワードネットについてと, その中での多義語, 同義語の扱い方について述べる.4 章同義語, 多義語を考慮した文書分類について述べ,5 章では実験結果を示し,6 章で結びとする. また、実験コーパスには Reuter21578 を用いる.

2. ベイズ確率論による文書分類

2.1 文 書 分 類

文書分類とは文書をその内容に応じて, あらかじめ決められたいくつかのカテゴリに自動で分類することである.
文書分類は,

・Web ディレクトリ構造の作成
・電子メールの自動フィルタリング
・大量の文書の整理

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、コストを削減できる.

2.2 教師付学習による文書分類

ベイズルールやベイズ公式と呼ばれる方法が分類手段として, 幅広く知られている. このベイズルールとは, 教師付学習を行う際に広く使用されている手法である. 教師付学習とは, 人手によりクラスを割り当てられたクラスが既知のデータ（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, その分類規則を用いて未知のデータを予測するものである.

教師付学習による文書分類の一般的な手順は

1:訓練データの作成

2:訓練データから分類規則の学習

3:分類規則の適用

となる.

2.3 ベイズ的学習

ベイズルールの基本的な考えは「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いる. そして, 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。」というものである. このベイズルールの特徴は,

・最終的な仮定成立の確率計算に事前知識を使用する.

という点である. この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, 各仮説が成り立つとしたときに訓練データである仮説が成り立つとする条件付確率の 2 つである.

また, ベイズルールの利点は

1 : 仮説が成立する確率というものを明示的に扱う

2 : 確率的な仮説の表現を許す. すなわち出力結果が真か偽ではない.

などがあげられ, ベイズルールの欠点は,

1 : 事前確率分布のために多くの初期知識を必要とする

2 : 計算コストが大きい

等である. しかし, 本実験においては, あらかじめ分類するためのデータはすべて揃っており, 量も十分にあるため前者の欠点に関しては特に問題はない.

2.4 ベイズによる文書分類

ベイズルールによる文書分類の場合, 以下の 4 つを考える [1]

1:$P(c_k)$　訓練例において, 文書がカテゴリ $c_k \in C$ 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:$P(x|c_k)$　訓練例において, カテゴリ 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

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 c_k に属するという事後確率

このときベイズルールは以下のようになる

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

カテゴリ集合 C の要素 c_k で、$P(c_k|x)$ の値が最も大きいものを、最大事後確率仮説と呼ぶ. ベイズルールでは, 最大事後確率仮説を取るカテゴリ c_k を文書 x が属するカテゴリ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.

ベイズルールでの分類規則の作成とは訓練データから $P(x|c_k)$,$P(c_k)$,$P(x)$ の値を求めることである.

2.5 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 $x = (x_1, ..., x_d)$ はほぼすべての文書で異なり, 天文学的な数がある可能性がある, そのため $P(x|c_k)$ や $P(x)$ の見積もりが問題になる. 単純 (naive) ベイズ分類では, ベクトル x をすべての c_k に対して以下の形式に分解して考える. [2]

$P(x|c_k) = \prod_{j=1}^{d} P(x_j|c_k)$　　(2)

単純ベイズ分類では, ドキュメントでの x_j の発生は統計的に他の $x_{j'}$ とは独立であると想定する. 単純ベイズ分類の使用により $P(x_j|c_k)$ を比較的少ないパラメタで構成することができ, (1) 式は以下のようになる.

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

本研究では, この単純ベイズ分類の使用により $P(c_k|x)$ を求める, そして文書 x を $P(c_k|x)$ が最大になるような 1 つのカテゴリ c_k に割り当てる, ドキュメント主導のシングルラベル文書分類を行う.

2.6 文書の表現

文書は自然言語表現で書かれた文字の並びとして保存されており, 文書分類アルゴリズムにより直接解釈することはできない. このため文書分類では, 文書を意味のある単位により認識することで, その内容をコンパクトに表現する必要がある. 文書分類で使用する文書の認識の単位は通常単語単位で,"set of words"や"bag of words"と呼ばれる [3]. この方法により文書を認識することで, 文書 x は単語の重みベクトル $x = (x_1, ..., x_d)$ と表現される. 本研究でも文書を単語単位により認識する. また文書分類においてフレーズ単位による認識はあまりよい有効性を示さないとされている [1] [3].

2.7 バイナリ独立モデル

上記のように, 文書分類では一般にドキュメントの表現を単語の重みベクトル $x = (x_1, ..., x_d)$ と表現する. バイナリ独立モデルとは, 単語重みベクトル $x = (x_1, ..., x_d)$ のすべての値を 0 か 1 で表現するもので [2], 単語の重み x_d の値を文書内で単語 x_d が現れた時は 1, 現れなかったときは 0 とするものである. バイナリ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.

$P(x_j|c_k) = p_{jk}^{x_j}(1 - p_{jk})^{1-x_j}$ (4)

$= (\frac{p_{jk}}{1-p_{jk}})^{x_j}(1 - p_{jk})$ (5)

ここで $p_{jk} = P(x_j = 1|c_k)$ である. これと式 (1) により以下の式を得る.

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

このバイナリ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. バイナリ独立モデルでは, 文書内での単語の出現回数とその確率を考える必要はなく, 文書内で単語が出現する確率, しない確率を計算すればよい.

2.8 分類規則の次元縮小

文書分類において高い次元の用語スペースは計算量や訓練データのオーバーフィットのため問題である. 通常, 文書分類では, この問題解決のために用語スペースを $|d|$ から $|\hat{d}| \ll |d|$ に次元縮小する試みを行う [3]. 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, 訓練データのオーバーフィット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るというリスクがある, そのため注意して次元縮小プロセスを行わなくてはならない. 本研究では用語選択によるローカル次元縮小を使用する. ローカル次元縮小とは, それぞれのカテゴリ c_i 毎に異なった用語セット $\hat{d}$ を作成することである. 用語選択による次元縮小の方法としては, 文書集合の中でもっとも高い頻度で現れている用語が文書分類にとってもっとも重要であると考え, それぞれのカテゴリ毎での出現頻度が高い用語を選択する. 本研究での用語選択による次元縮小の方法には,"DIA asociation factor" [3] 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ローカル次元縮小の場合 $10 \leqq |\hat{d}| \leqq 50$ が一般的とされている [3]. 本実験でもこの範囲の値を使用する. トピック毎の,$P(c_k|x_j)$ の値の大きな単語とその値を表 2 に示す.

トピック	gas	trade	fuel
1	0.825 gasoline	0.930667 trade	1 fuel
2	0.6 oil	0.586667 year	0.692308 pct
3	0.6 mln	0.477333 billion	0.692308 oil
4	0.5 crude	0.469333 export	0.615385 dlrs
5	0.475 pct	0.453333 import	0.538462 petroleum

表 1　$P(c_k|x_j)$ の上位とその値の例

3. ワードネットによる同義語, 多義語の利用

3.1 ワードネットとは

ワードネットとは Princeton 大学でのプロジェクトで「英語用語彙データベース (lexical database for the English language)」とも呼ばれる. ワードネットは, 同義語の集合を同義語セットまたは"synset"と呼ばれるグループにまとめることで、単語と概念の記述および分類を行うシステムである.

3.2 なぜワードネットを使用するか

本研究では, ワードネットは特定の分野の知識を持たず, 一般的な知識を持つと考える, また同様に, 今回実験に使用する Reuter（新聞記事）も比較的一般的な文書であると考える. 本研究では, 特定のカテゴリ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考えるため, シソーラス辞書としてワードネットを使用する.

3.3 同義語, 多義語と頻度の利用

図 1, 図 2 は"human","person"をワードネットで検索した場合の検索結果である."human","person"は複数の品詞の複数の意味（多義）を持つ. また検索結果は"human"の持つ意味の一覧が"名詞","動詞","形容詞","副詞"の順に出力される [表 3].

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)

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)

名詞
1.（頻度）意味、同義語（synset）
2. ·····
動詞
1. ·····
形容詞
副詞

表 2　ワードネットの出力

ワードネットでは, 単語の意味に意味番号が割り振られおり, 同じ意味を持つ単語は同じ意味番号を持っている. 例えば"human","person"は共に意味番号 5303（図 1, 図 2 のそれぞれ一番上の意味) という同じ意味 (同義) を持つ [表 3]. 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で, 意味の使用頻度として使うことができる. また, 頻度の低い意味は頻度がつけられていない.

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

表 3　単語が持つ意味番号

4. 単語の意味を考慮した文書分類

4.1 通常の文書分類

通常の文書分類では単語の意味を考慮せず, 単語を単に記号的に解釈する. 文書内には複数の意味も持つ単語が存在したり, 複数の単語が同じ意味を示すこともある. 例えば,"bank"が

「銀行」や「堤防」という複数の意味（多義）で使われることや,”student”と”pupil”が両方出現する場合（同義）もある. 単語を単に記号的に扱う通常の文書分類では表 4 のように”bank”の出現は「銀行」という意味か「堤防」という意味かはまったく考慮しない. また, 同じ文書で”student”と”pupil”が出現する場合, これらの単語は違うものとして扱う.

実験ごとに, 訓練データ, テストデータの選択方法や件数,”stopword”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通常のベイズによる文書分類による Reuter コーパスの分類精度は約 74 %〜79.5 %となっている [3], また本実験と同じ条件での通常のベイズによる文書分類での分類精度（正解率）は 79.26 %であった. 詳しくは 5 章で述べる.

文書 x1	カテゴリ：金融
The bankruptcy of the bank.	
文書 x2	カテゴリ：工事
The construction of the bank.	
↓	
金融カテゴリの分類規則に”bank”がある場合	
文書 x_2 の金融カテゴリに属する確率が増加	
↓	
精度の低下が考えられる	

表 4　多義語を考慮しないことによる精度低下

お互いに重ならないカテゴリ”school”に属する文書集合	
文書集合 X_1 student が出現	文書 X_2 pupil が出現
$P(”student”\|”school”) = 0.2$	$P(”pupil”\|”school”) = 0.2$
2 つの単語を同じと考えると $P(”student, pupil”\|”school”)$=0.4 重みの増加	

表 5　同義語の考慮による精度向上

4.2　同義語を考慮した文書分類

同義語を考慮した文書分類の研究も行われている [5]. 表 5 のように,”student”と”pupil”が”school”というカテゴリの文書集合で出現していた場合, この 2 つの単語を同じものと考えることで, これらの単語の重要性を増すことができる. このように同義語を考慮することで, 精度の高い分類が行えると考えられる. しかし, 同義語だけを考慮して文書分類を行った場合, 使用される頻度が非常に低い意味の同義語を同義語をして扱った時に頻度の低下が考えられる. 例えば 6 のように文書内で”man”という多義の単語が出てきたとき, 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”という意味も持つことから”piece”という単語を同義語として扱ったとき, これは全く違う意味であり, 分類精度の低下が考えられる, 同様に”男(gentleman)”を同義語として扱った場合も意味は近いが, 問題である.

4.3　同義語, 多義語と頻度を考慮した文書分類

上記問題を解決するため, 本研究では単語の同義性, 多義性と,

互いに重ならない同量の文書集合 X_1, X_2
X_1 で単語”man”がカテゴリ”science”の 9 割で出現 $P(”man”\|”science”) = 0.9$ →分類に重要
X_2 で単語”piece”がカテゴリ”science”の 1 割で出現 $P(”piece”\|”science”) = 0.1$ →分類にあまり重要でない
science”カテゴリの分類規則に”man”があった場合 man”と”piece”の単語を同義語と考えると $P(”man”, ”piece”\|”science”) = 0.5$ science”カテゴリにおける”man”の重要性の低下 ↓ 分類精度の低下

表 6　同義語だけを考慮することによる精度の低下

その意味の使用頻度を利用する.2-7 で述べたバイナリ独立モデルでは, 特徴ベクトル $x = (x_1, ..., x_d)$ の値を, すべて 0 か 1 で表現する. 本研究では表 4 や表 6 のような多義語の使用頻度の低い意味の同義語を的確に扱うことを考える. このため本研究では, 文書ベクトルの値を単語の意味の頻度から決定する.

このため, 本研究の同義語, 多義語を考慮した文書分類はバイナリ独立モデルではないといえるかもしれないが, 文書内での単語の出現回数では重み付けを行わない. 重み付けの具体的な方法については 4-5 で述べる.

4.4　ワードネットの利用での問題と解決

1 つの単語は複数の意味を持つので, ワードネットの利用により次元や計算量の増加が考えられる. 本研究では通常の文書分類の手順で作成した分類規則 [(2-5）表 2] にワードネットを適用し頻度により重み付けを行う [表 7]. これにより、ワードネットの利用による次元の増加を比較的少なくすることができる. また, 通常の文書分類と同じ分類規則を利用することができる. 本研究では, 分類するテストデータにもワードネットを適用するが表 6 のような同義語考慮による精度の低下を防止するために, 頻度による重みは考慮せず, 最大の頻度で使われる意味だけを考慮する [表 8]. これにより本研究ではテストデータ内の単語の意味を最大の頻度で使われる意味として扱う. 本研究では, このようにワードネットを適用することにより, 次元の増加や使用頻度の少ない意味による精度の低下を防ぐ.

4.5　頻度による重み付け

上記のように, 本研究では分類規則とテストデータにワードネットを適用し, テストデータの単語は最大の頻度で使われる意味だけを考慮する. この方法により本研究では, 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考え, 最大の頻度が低い単語の重要性は減少させる. 本研究での重み付けは, 表 8 の文を表 7 の”gas”カテゴリの分類規則に当てはめた場合, バイナリ独立モデルでの 0 か 1 とは異なり,oil が 0.647 回,gasoline が 1 回出現したと考える. ワードネットの使用により oil の出現が 1 から 0.647 に減少した. ほとんどの用語は 1 つの意味が頻度の大部分を占めている.

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 8 のように最大頻度の意味しか考慮しないので, テ

ストデータでこの 2 つの単語が出現した場合は全く同じものとして考える. 本研究ではこの方法により, 表 5 の精度の向上や表 6 の同義語の問題を扱う.

カテゴリ	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	0.5019
				6962272	124	0.1580
				844555	107	0.1363
				454930	94	0.1197
2	0.6 oil			0.586667 year		
	12653557	11	0.647	12867473	832	0.9618
	3347615	5	0.2941		・	
		・			・	
		・			・	
3	0.6 mln			0.477333 billion		
	mln	0	1	11604853	1	1

表 7 分類規則にワードネットを適用

oil is gasoline
最大頻度の意味番号に変換
↓
oil → 12653557
gasoline → 12402140

表 8 テストデータにワードネットを適用

また似た意味の単語がテストデータで出現した場合. 例えば 1 つの文書で"trade"と"craft"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. 本研究では表 7 の"trade"カテゴリに関して,"trade"の 831317 により"trade"が 0.5019 回出現, また"craft"の 454930 により同じく"trade"が 0.1197 回出現したと考える. また, この場合この文章での"trade"の出現回数 x'_j は $x'_j = 0.5019 + 0.1197 = 0.6216$ 回出現しているとする. 本研究ではテストデータの単語が持つ最大頻度の意味と分類規則の単語が持つ意味を比較し, 同じ意味を持つ場合, 分類規則での意味の頻度を単語の類似度とする. 本研究ではこの類似度により単語の出現回数を重み付ける. これにより上記の表 4,6 のような問題を防ぎ, より精度の高い分類が行えると考える.

4.6 使用する意味の限定

本研究では, 分類規則にワードネットを適用し, その意味番号と使用頻度を使用するが, 単語が持つ意味をすべて使用するのではなく, 表 9 のように, 頻度の高い上位 S 個の意味を使用する. このように本研究では, 分類規則で使用する意味を限定する. これにより, 上位の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持たせることができると考えるため, この方法を利用する. 本研究ではこの方法により表 6 のような問題をより的確に扱うことができると考える. しかし, 使用する意味を限定することは, 使用頻度が S 番目以下の意味はつかわれなくなるため, 情報の損失が考えられる, そのため, 本実験では S=1,2,3,5 とすべての意味を使用する計 5 パターンを比較する.

human		
全意味番号と出現数, 頻度		
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
	・	
	・	

→

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

表 9 分類規則での全意味使用と S=2 の場合

4.7 単純ベイズ分類への適用

上記の同義語, 多義語と頻度情報を利用した重み付けを通常の文書分類に適用する.4-5 で述べた頻度を考慮した出現回数 x'_j により, 本研究では 2-7 での (4) 式を以下のようにする.

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

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

(4) 式を (8) 式に変更することにより, 同義語, 多義語と頻度を考慮した文書分類を行う. 通常単語の重みは 0 か 1 である. 本研究では, 同義語, 多義語を的確に考慮するために, この重みを頻度により決定した.

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, [1] の分割に従い訓練文書集合とテスト文書集合を作成する. また本実験での Reuter の設定を以下に示すこの設定も基本的に [1] に従う,

・Reuter の"TOPICS"を分類するカテゴリ, すなわち分類の答えとする

・"TOPICS"を持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"は削除

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, 総トピック数は 73 となった. また, 訓練集合, テスト集合の両方とも stop-word についてはあらかじめ削除しておく.

5.2 評価方法

本実験では通常のベイズによる文書分類とワードネットを適用した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したいため, ドキュメント主導のシングルラベル文書分類で行う（2-5）. 精度の評価方法は,

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

で行い, 文書が複数のカテゴリ（答え）を持つ場合は, そのうちのどれか 1 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 1 つのカテゴリに割り当てているため, 精度と再現率の値は等しく, ここでの正解率と同じ値になる.

5.3 実験手順

2-2 で述べたように, 教師付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, 最後にその結果から性能評価となる. この 1,2 は通常のベイズによる文書分類も今回のワードネットを利用した文書分類も同じである. ワードネットを用いた文書分類では通常のベイズ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を利用し最大頻度の意味のみを考慮する, また 4 の評価方法は 5-2 で述べた通りで同じである.

5.4 実験結果

通常のベイズによる文書分類での実験結果を図 3, 表 10 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単語数は 40 である.

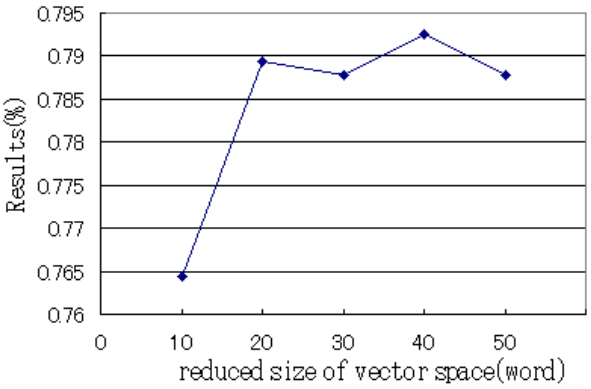

図 3　通常のベイズによる分類結果

分類規則の単語数 (語)	10	20	30	40	50
正解率 (%)	76.43	78.94	78.77	79.26	78.77

表 10　通常のベイズによる分類結果

また次にワードネットにより同義語, 多義語, 頻度を考慮した文書分類の分類結果を図 4, 表 11 に示す. このワードネットを使用した結果では S=5, 分類規則の単語数が 50 のときに 82.31 %の正解率を示した. また,S=2 以上の場合は, ほぼ同じ値を示した.

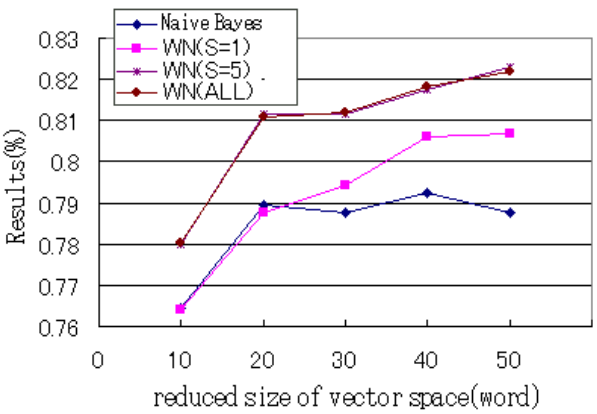

図 4　ワードネット使用による分類結果

分類規則の単語数 (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

表 11　ワードネット使用による分類結果

5.5 考　　察

この図 4, 表 11 より, 本研究のワードネットの使用した文書分類により通常のベイズによる文書分類よりも分類結果の正解文書数が 2442 件→ 2536 件と増加し, 正解率が約 3.04 %向上した. ワードネットを利用した文書分類による正解率の向上の値は小さいが, 誤分類された文書数は 639 件→ 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, より正確にテキストの分類が行えると考えられる.

また通常のベイズによる文書分類とワードネットを使用した文書分類の正解文書の変化を表 12 に示す.

		通常のベイズ	
		正解	間違
ワードネット	正解	2339	197
	間違	103	472

表 12　通常のベイズとワードネットを使用した文書分類の正解数, 間違い数

通常のベイズによる分類では正確に分類され, ワードネットを使用して分類することにより正確に分類されなくなる文書が 103 件あった. この内容も見てみたところ, ワードネットを利用した文書分類により間違って割り当てられたカテゴリは, 正解のカテゴリに似ているものばかりであった. 例えば”trade”が正解のカテゴリの文書を”yen”カテゴリに誤分類したり, 他にも”rice-cotton”,”crude-fuel”等で, これらのカテゴリの分類規則は類似している, また正解のカテゴリ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,

・ワードネット作成に使われたコーパスと Reuter との意味頻度等の違いなど相性

・割り当ての精度は手作業により割り当てられたトピック（カテゴリ）に依存する

・手作業によるカテゴリ分類の基準があいまい

等の原因が考えられる.

また, 通常の文書分類でも本研究の文書分類でも正解できなかった 472 件のデータに関しては

・他の文章と内容がかけ離れている

・人手によるトピック割り当ての問題

などが考えられるが,

・確率論による分類の限界

という原因が考えられる. しかしながらワードネットを使用することにより多義語, 同義語, 頻度を考慮した文書分類を行うことにより誤分類されるドキュメント数が 15 %減少し, 正解率も

82.3 %を示したことから, 文書分類において同義語, 多義語, をその使用頻度を考慮することは有用性を示すことが確認できた.

6. 結　　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.

また, 今後の展開としては

・確率論以外からのアプローチ

・ワードネットの品詞情報を利用

・ワードネット使用による重み付けの”Boosting”

等を行うことにより更なる性能の向上を目指す.

謝辞

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] David D.Lewis: 1992b. ”Representation and Learning in Information Retrieval”. Ph. D. thesis, Department of Computer Science, University of Massachusetts, Amherst, MA.

[2] David D.Lewis: ”Naive (Bayes) at forty: The independence assumption in information retrieval”. In Proceedings of ECML-98, 10th European Conference on Machine Learning (Chemnitz, Germany, 1998), 4・5.

[3] Fabrizio Sebastiani: ”Machine Learning in Automated Text Categorization” , proc.ACM Computing Surveys,Vol.34,No.1,2002 pp.1-47

[4] 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”, 人工知能学会誌,16 巻,2 号,2001

[5] 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”, 人工知能学会誌,16 巻,2 号,2001