自然言語処理 118－16
（1997．3．22）

共起情報を利用した文書の自動分類

藤井洋一、鈴木克志、今村誠、高山泰博
三菱電機株式会社　情報技術総合研究所　音声・言語インタフェース技術部
E-mail　{fujii, suzuki, imamura, takayama}@isl.melco.co.jp

従来、文書の自動分類の精度向上において問題となっていた多義語の曖昧性を解消し、分類精度を向上するために、以下の二つの点を特徴とする文書自動分類方式を提案する。

1．単語出現頻度と文書分類項目との間の χ^2 統計による重み付けにおいて複数の分類項目で重要度が高い語を分類多義語と定義する。

2．分類多義語と同一段落内で共起する語のみから構成する共起単語の共起ベクトルと入力文書の共起単語の共起ベクトルとの類似度を計算することで入力文書の分類多義語の出現頻度を補正する。

新聞記事 65,447 記事における 735 分類項目への分類実験の結果、再現率が従来方式による場合の 46.4%から最大で 5.6 ポイント改善された 52.0%になり、適合率が 44.4%から 5.1 ポイント改善された 49.5%になった。

An Automatic Document Classification Using Lexical Co-occurrences

Youichi Fujii, Katsushi Suzuki, Makoto Imamura, Hiroyasu Takayama
Human Media Technology Dept., Information Technology R&D Center,
Mitsubishi Electric Corporation

This paper describes an automatic document classification using lexical co-occurrences. In the conventional documents classification methods, there is a problem of miss-classification caused by ambiguous words. In order to resolve the ambiguity depending on the class categories, we use the lexical co-occurrences with the ambiguous word. The method has two main ideas as follows:

1. We define 'the ambiguous word in classification' as the word that characterizes in plural classes in χ^2 method between the word frequency and document classes.
2. And we adjust a word frequency for the ambiguity in classification using the words co-occurred with the ambiguous word in the same paragraphs.

We classify 65,447 newspaper articles into 735 classes in the experiment. The best result shows the recall improvement of 5.6 points from 46.4% of the conventional method to 52.0%, and the precision improvement of 5.1 points from 44.4% to 49.5%.

1　はじめに

近年大量のテキスト情報がインターネットなどを通じてアクセス可能となるにつれて、文書の自動ファイリングやユーザの要求を満足させる自動配信の要求が高まっている。

我々は、大量文書中からユーザの要求を満足させる文書の自動配信を考えた。しかし、従来の全文検索のようなアプローチでは検索ゴミが多すぎる。そこで、我々は配信する文書をある程度細かい分類(以下、設定された分類を分類項目と呼ぶ) にあらかじめ自動分類しておき、ユーザが目的とする分類を選択できるようにするとともに、ユーザが分類をより細かく設定できる必要があると考えた。また、大量の文書

を扱うためにはシステムとして自動学習可能なことが必要不可欠である。そこで、自動学習が可能であるベクトル空間モデルに注目した。

ベクトル空間モデルによる自動分類は文書中に出現した単語などの情報の特徴量をベクトルで表現し、文書と分類項目間のベクトルの類似度によって自動分類先を決定する。ベクトル空間モデルによる自動分類では、精度向上のために様々な手法が提案されている。例えば、河合[1]は単語の意味属性を用いて分類精度が向上することを示し、湯浅[2]は単語共起による単語のベクトルを利用した分類方式を提案した。しかし、これらの実験はいずれも 10 程度の分類項目でしか評価されていない。

また、河合[1]では、分類を誤る原因の一つとして、一つの単語が複数の意味を持つ多義性単語の問題が指摘されている。このような従来の言語処理で問題になるような多義性の解消は少ない個数の分類では有効である。一方、分類数が多くかつ細かくなった時には、同じ意味を持つ単語でも文脈上で果たす役割の違いで異なる分類項目に分類する必要が出てくる。そこで、我々は単語が文脈上で果たす役割の違いを捕らえられれば分類精度が向上するであろうと予測し、分類に重要である単語のうちで特に複数分類項目にて重要度が高い単語に対して、文脈上で果たす役割の違いを捕らえる方法として共起を考えた。例えば、「大統領」という単語が＜政治＞（以下、単語と区別するため、分類項目名は＜＞で挟む）や＜国際＞といった複数の分類項目で頻繁に出現するように、複数分類項目で重要度が高い単語を「**分類多義語**」と定義し、文書の「分類多義語」が出現した段落と同一の段落に出現（**段落内共起**と呼ぶ）した単語の頻度情報を「分類多義語」の多義性解消に利用する方法を試みた。すなわち、「分類多義語」に対して分類項目別に学習情報の頻度と重み付けを再学習し、分類対象文書中の「分類多義語」に対して段落内共起単語の頻度情報を使って分類多義語の頻度を補正し、再学習した結果との類似度計算によって分類先を決定する。

2 分類多義語

2.1 分類多義語の定義

単語出現頻度と文書分類項目との間の従来の χ^2 統計による重み付けにおいて複数の分類項目で重要度が高い単語を分類多義語と定義する（また、この時の複数の分類項目を**重要分類項目**と呼ぶ）。

すなわち、単語「大統領」が＜政治＞という分類項目にのみ多く現れたとすると、「大統領」は、＜政治＞へ分類の有効な単語となり、分類多義語とならない。ところが、＜首相＞や＜大統領選挙＞、＜地方行政一般＞、＜外交関係＞といった分類項目で単語「大統領」が頻繁に出現したとすると、単語「大統領」はどの分類項目に対しても重要度が高く、分類多義語である。

我々は、分類多義語が分類項目によって異なる単語とともに使用される可能性が高いと考えた。たとえば、単語「大統領」は、＜首相＞では、「最高責任者」、「発言」といった単語とともに現れ、＜大統領選挙＞では、「選挙」、「当選」といった単語とともに現れ、＜外交関係＞では、「来日」、「対談」といった単語とともに現れる可能性が高くなる。

2.2 分類多義語による語の分割

分類多義語に重要分類項目をつけたものをあたかも単語であるかのように扱うこととする。すなわち、＜首相＞および＜大統領選挙＞、＜地方行政一般＞、＜外交関係＞で頻繁に現われる単語「大統領」に対しては「大統領$_{<首相>}$」と「大統領$_{<大統領選挙>}$」、「大統領$_{<地方行政一般>}$」、「大統領$_{<外交関係>}$」という拡張単語（以下、**分類多義拡張単語**と呼ぶ）を定義し、これをベクトル計算の基底単語と呼ぶことにする。

自動分類対象の入力文書中の分類多義語の多義性を解消して分類多義拡張単語に変換し、分類多義拡張単語を頻度計算の単位とすれば、

分類多義語を1単語のまま扱うよりも分類精度が向上すると考えられる。

3 自動分類方式

自動分類の方式として河合[1]で用いられているχ^2統計を応用した方式に本手法を適用して自動分類を行った。但し、以下では意味属性を使わない単語のみの自動分類方式を差すものとする。

3.1 従来の自動分類方式

従来の自動分類方式として河合[1]の方式を説明する。

学習フェーズでは、全ての学習文書に対して単語を抽出し、正解の分類先から各分類項目毎に単語の頻度学習を行い、χ^2統計による重み付けを学習して、学習テーブルができる。すなわち、分類項目を$C_i(i=1,\ldots,N)$、単語を$w_j(j=1,\ldots,L)$としたとき、分類項目C_iに単語w_jが出現した回数（出現頻度）をF_{ij}とすると、単語w_jが分類項目C_iに理論的に出現する回数（理論頻度）M_{ij}と、単語w_jが分類項目C_iへの寄与する度合い（重み付け）Y_{ij}は、それぞれ(1)、(2)で計算される。

$$M_{ij}=\sum_{i=1}^{N}F_{ij}\cdot\sum_{j=1}^{L}F_{ij}\Big/\sum_{j=1}^{L}\Big(\sum_{i=1}^{N}F_{ij}\Big) \quad (1)$$

$$Y_{ij}=(F_{ij}-M_{ij})\cdot\left|F_{ij}-M_{ij}\right|\Big/M_{ij} \quad (2)$$

一方、自動分類時には、分類対象文書中の単語w_jの出現頻度をd_jとし、文書ベクトル$\mathbf{D}=(d_1,\ldots,d_L)$を計算する。上記重み付け$Y_{ij}$に対して単語$w_j$の重みベクトル$\mathbf{S}_j$を$\mathbf{S}_j=(Y_{1j},\ldots,Y_{Nj})$と定義する。また、重みベクトル$\mathbf{S}_j$に単語$w_j$の出現頻度$d_j$を掛けて、

$$\sum_{j=1}^{L}(\mathbf{S}_j\cdot d_j)=(s_1,\ldots,s_N) \quad (3)$$

とおく。この時、(3)で表されるs_iを使って、分類対象文書の分類項目C_iへの類似度を$s_i\Big/\left(\sum_{i=1}^{N}s_i\right)$と定義する。そして、類似度の値が大きい分類項目を分類先とする。

3.2 分類多義語を用いた自動分類方式

本手法における自動分類は、図1の流れに従って行う。

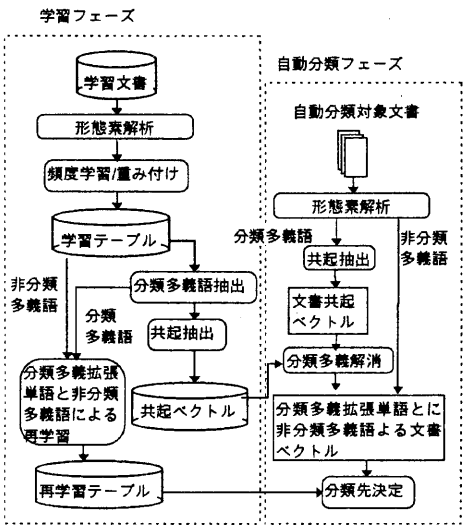

図1 自動分類の流れ

まず、学習フェーズでは、学習文書を形態素解析する。各学習文書から単語出現頻度を分類項目毎に計算して重み付けを行い、学習テーブルを作成する。ここまでは河合[1]と同様である。次に、学習テーブルから分類多義語を抽出し、各分類多義語に対して、**共起ベクトル**を作成する。一方で多義語分割された分類多義拡張単語と非分類多義語を使って頻度計算および重み付けを再学習し、再学習テーブルを生成する。（詳細は3.3で述べる。）

学習フェーズが終了すると、自動分類フェーズの処理が可能となる。自動分類フェーズでは、分類対象の文書の形態素解析を行い、分類多義語に対しては分類対象文書中の分類多義語と段落内共起するすべての単語の頻度を抽出し、分類多義語の**文書共起ベクトル**とする。各分類多義語に対して学習結果の共起ベクトルと、分類対象文書の文書共起ベクトルとから分類多義の解消をして、分類多義拡張単語の頻度を求める。最後に、分類多義拡張単語と非分類多義語の頻度から作成した文書ベクトルと、再学習テーブルとの類似度を計算して分類先を決定する。（詳細は3.4で述べる。）

3.3 分類多義による学習結果の再学習法

本手法では、分類多義語を抽出しその分類多

義語に対して再学習を行う。また分類対象文書の分類多義語の頻度を分類多義拡張単語の頻度に分配するために共起ベクトルを作成する。

分類多義語は、河合[1]の方式で学習した重み付け情報(2)の中で、以下の(4)の条件を満たす単語 $w_j (j=1,\ldots,L)$ とした。

$$\#\{Y_{ij} \mid \max_{1\le k\le N}(Y_{kj})\cdot V \le Y_{ij}\} \ge 2 \ (V\text{:閾値}) \qquad (4)$$

分類多義語に対しては、単語と重要分類項目との組で表現した分類多義拡張単語を基底単語として頻度再学習したあとで、(1)、(2)の式を使って再学習し、再学習テーブルを作成する。図2は再学習テーブルの頻度再学習の方法を示したもので、各分類項目に対して対応する分類項目が付いた分類多義拡張単語がある場合には、対応する分類多義拡張単語に全ての頻度を割り当てる。一方、対応する分類項目が付いた分類多義拡張単語が無い場合には、全ての分類多義拡張単語に均等に頻度を割り当てる。

学習テーブルの頻度学習結果

単語 / 分類項目		大統領	
首相		120	
大統領選挙		50	
裁判		5	
地方行政一般		70	
外交関係		90	
:		4	

<首相>は「大統領<首相>」が分類多義拡張単語なので頻度120をそのまま「大統領<首相>」に与える。<裁判>は大統領の重要分類項目でないので、値を4つの分類多義拡張単語に割り振り、1.25ずつ与える。

再学習テーブルの頻度学習結果

単語 / 分類項目		大統領 <首相>	大統領 <大統領選挙>	大統領 <地方行政一般>	大統領 <外交関係>	
首相		120	0	0	0	
大統領選挙		0	50	0	0	
裁判		1.25	1.25	1.25	1.25	
地方行政一般		0	0	70	0	
外交関係		0	0	0	90	
:		1	1	1	1	

図2　分類多義語の頻度分割例(学習時)

また、図3のように抽出された各分類多義語に対して、それぞれの分類多義拡張単語に付いている分類項目と正解分類先が一致し、かつ分類多義語を含む文書すべてから、分類多義語と段落内共起する単語の頻度を値とするベクトルを**共起頻度ベクトル**とする。共起頻度ベクトルは、4.6.1に示す4種類の方式で処理をして、共起ベクトルとなる。

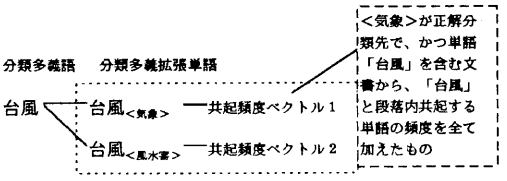

図3　共起頻度ベクトルの作成例

3.4共起情報を利用した分類方式

自動分類時には、分類対象の文書の形態素解析結果から、単語の頻度を計算する。分類多義語に対しては、分類多義語と段落内共起した単語をすべて取り出し、その頻度を値とする文書共起ベクトルを生成する。分類多義語の解消は図4のように、文書共起ベクトルと、分類多義語の共起ベクトルとの内積を計算し、内積値によって分類対象文書中の分類多義語の頻度を分類多義拡張単語の頻度に比例分配する。最後に、再学習テーブルと、分類多義拡張単語と非分類多義語の頻度を値とする文書ベクトルとの類似度を河合[1]と同様に計算し分類先を決定する。

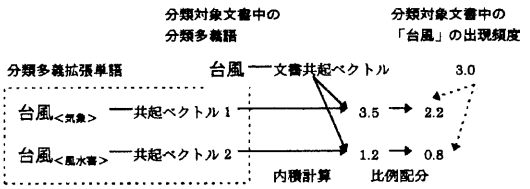

図4　分類多義語の頻度分割例(自動分類時)

4　実験

形態素解析にはJUMAN 2.0[4]+EDR日本語単語辞書1.5版[5]を利用した。実験は、従来方式による自動分類の精度を求める実験と、分類多義が理想的に解消された場合の精度を試算する実験及び、共起情報を利用した自動分類の精度を求める実験の3つを行った。

4.1　記事データ

今回の実験に用いた新聞記事（朝日新聞記事1年分：朝日新聞社提供）[3]の「主題分類」は図5に示すように4階層の分類を持っている。

大分類、中分類、小分類、小小分類のうち、各記事には主として小分類、小小分類が複数個付与されており、これを正解として、自動分類

の評価を行った。

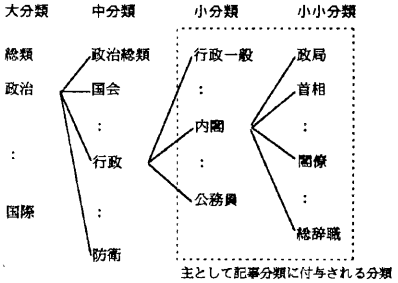

図5　新聞記事分類

4.2　学習方式

従来の単語のみを使った河合[1]の方式による自動分類の実験のため、1）大分類、2）中分類、3）小分類、小小分類の3種類の分類を考えた。4.1に示したように1記事に付与された分類項目は、3）の階層で複数個付与されているので、学習時の1記事中の頻度を表1のように分配した。例えば、<図書>、<読書>及び、<日本文学>という正解分類項目が付与されている記事を考える。3）の分類では、1記事中の頻度を正解分類項目数3で割った頻度を各分類項目で学習した。一方、1）の分類では、上位の大分類項目<総類>と<文化>を正解分類項目とし、学習時には、3）の割合をそのまま加えた大分類項目での「頻度にかける割合」が出現したものとして学習した。2）の分類でも1）と同様の処理を行った。

表1　学習時の正解分類先対応付け方法の例

小分類項目		頻度にかける割合	大分類項目		頻度にかける割合
052	図書	0.33	0	総類	0.66
054	読書	0.33			
412	日本文学	0.33	4	文化	0.33

各分類項目の番号は分類コードを示す

4.3　実験方法

まず、3で示した河合[1]の方式で重み付けした学習テーブル利用して、200 記事と、1000 記事及び、10000 記事で従来方式による分類精度を求めた。次に、実験文書に関しても分類先（正解）が分かっていることを利用して、上記10000 記事を利用して3）の分類で理想条件での分類精度を試算をした。最後に上記 10000 記事と、1年分の新聞記事の2種類を利用して3）の分類で本方式による自動分類を行った。なお、200 記事と、1000 記事及び、10000 記事の実験の際には、河合[1]の実験に等しくなるように、150 文字から 500 文字の記事を古い順番に取り出し、学習記事と実験記事がそれぞれ75%と 25%となるようにランダムサンプリングした。

今回の実験で利用した単語は、名詞及び、固有名詞、未知語、サ変名詞であり、1文字のみからなる単語と記号列のみの未知語は取り除いた。なお、以下で示す自動分類の結果は、次の数値を示している。

- 1st:最も分類先として類似度が高いと判断されたものが、正解に含まれる割合(%)
- ave:類似度の閾値を変化させた場合に適合率と再現率が等しくなる時の正解率(%)

4.4　従来方式による分類結果

従来の河合[1]の手法を使って自動分類を行った結果を表2に示す。

表2　従来の手法による分類精度

分類（分類項目数）	記事数					
	200		1000		10000	
	1st	ave	1st	ave	1st	ave
大分類(10)	54.0	48.4	77.6	63.4	82.4	68.8
中分類(92)	−	−	62.4	50.0	72.8	58.5
小分類(735)	−	−	48.0	37.8	62.1	48.1

ただし、小分類は小小分類を含む

200 記事では河合の実験とほぼ同様の約50%の精度を得ることができ、大分類に限れば記事の数を増やすことで約 70%まで精度向上した。しかし、小分類では大分類よりも約20%精度が悪く、分類が細かくなった時には問題が大きい。

4.5　理想条件での分類結果

理想実験では分類記事に対して、分類多義拡張単語への頻度の分配を図6の様に行い、類似度を計算した。具体的には、分類多義語「台風」(頻度3)に対して記事の正解分類先が3つの場合は、それぞれの分類先に頻度1.0を与え、<運輸交通一般>は単語「台風」の重要分類項目

でないので、「台風_{<気象>}」と、「台風_{<風水害>}」に、均等に 0.5 を与える。一方、＜風水害＞は、分類多義拡張単語「台風_{<風水害>}」の分類項目と一致するので、「台風_{<風水害>}」に＜風水害＞に与えられた頻度 1.0 をそのまま与える。従って、文書ベクトルの頻度は、「台風_{<気象>}」が 1.0、「台風_{<風水害>}」が 2.0 となる。

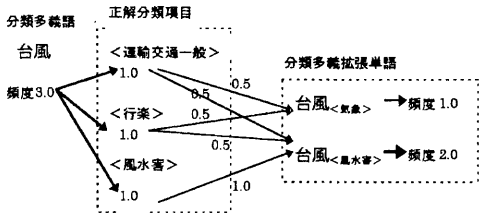

図6　理想条件での頻度分配の例

(4)の閾値を変化させ、自動分類した結果を表3に示す。

表3　理想条件での分類精度

閾値(ν)	0.9	0.7	0.5	0.3	0.1
1st	62.8	64.5	66.5	71.9	82.4
ave	48.3	49.0	50.7	54.7	65.7
分類多義語数	2229	7176	12399	18099	24637

この理想条件の方式は、正解分類先でない分類項目に対して頻度を下げる効果があり、閾値を下げれば下げるほど分類多義語が増え、精度が向上するのは明らかである。本来は各閾値を使って実験するべきであるが、今回は最も良かった 0.1 を選択した。

4.6　共起情報利用による分類結果

共起による分類多義語解消のため、共起頻度ベクトルに対して 4 種類の方式を考え、共起ベクトルをそれぞれ作成して実験した。

(I)　分類多義拡張単語毎に長さ 1 に正規化
(II)　頻度 0 のもの以外は全て 1 にする
(III)　(I)のベクトルを共起単語が出現した分類数で割る
(IV)　(II)のベクトルを共起単語が出現した分類数で割る

(I)は頻度を重視し、(II)は出現の有無を重要視する効果がある。また、(III)、(IV)は多くの重要分類項目で出現した共起単語に対して分類多義語解消の影響を小さくする効果がある。

4.6.1　新聞 10000 記事の分類結果

理想実験の場合と同じ 10000 記事で自動分類を行った結果のうち、従来(orig.)、共起を用いた場合(IV)、理想(id.)の適合率/再現率のグラフを図 7に示す。また分類結果を表 4に示す。

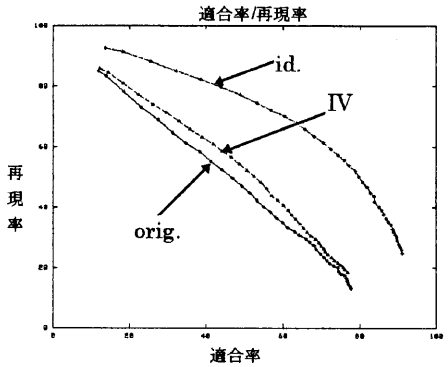

図7　新聞 10000 記事の適合率/再現率

表4　新聞 10000 記事の分類結果

	1st	ave
従来(orig.)	62.1	48.1
共起を用いた場合(I)	62.8	49.6
共起を用いた場合(II)	64.0	50.2
共起を用いた場合(III)	64.0	50.5
共起を用いた場合(IV)	65.1	51.4
理想実験(id.)	82.4	65.7

従来(orig.)と比較して、0.9～3.3 ポイント向上し、1st は(IV)の場合に最大の 3.0 ポイント、ave も(IV)の場合に最大の 3.3 ポイントの向上があった。

4.6.2　新聞 1 年分の記事の分類結果

次に、学習記事数の影響を見るため、新聞記事 1 年分で実験した。学習記事を 61500 記事とし、残りの 3927 記事のうち、記事サイズが 150 文字から 500 文字の 1880 記事を自動分類した。また、この実験では、学習記事数をなるべく多くする目的で、学習記事のサイズ制限はしなかった。一方、分類多義語は閾値 0.1 で抽出したが、ディスクスペースの関係から(5)式の値が大きい上位 10000 分類多義語に制限した。(5)は分類に重要な単語ほど値が大きくなる。

$$\sum_{i=1}^{N}(F_{ij}-M_{ij})^2 / M_{ij} \qquad (5)$$

なお、理想実験において、分類多義語を閾値 0.1 で(5)の値が大きい上位 10000 分類多義語に

制限しても、1st で 82.1%、ave で 65.3%となり、ほとんど精度の低下は見られなかったことから、新聞記事1年分でも分類精度にあまり影響ないと考える。結果を表5に示す。

表5　新聞記事1年分で記事限定した分類結果

	1st	ave
従来(orig.)	64.7	49.0
共起を用いた場合(I)	66.7	51.8
共起を用いた場合(II)	65.1	50.0
共起を用いた場合(III)	67.0	51.7
共起を用いた場合(IV)	66.3	50.6
理想実験(id.)	79.8	62.3

従来(orig.)と比較して、0.4～2.8ポイント向上し、1st は(III)の場合に最大の 2.3 ポイント、ave は(I)の場合に最大の 2.8 ポイントの向上があった。

4.6.3　記事サイズを制限しない分類結果

最後に、分類対象記事サイズの影響を見るため実験記事サイズも制限せず、3927 記事での実験を行った。分類結果を表6に示す。

表6　新聞記事1年分の分類結果

	1st	ave
従来(orig.)	64.3	48.1
共起を用いた場合(I)	66.3	50.3
共起を用いた場合(II)	65.1	48.9
共起を用いた場合(III)	67.0	50.7
共起を用いた場合(IV)	66.2	49.7
理想実験(id.)	79.2	60.9

従来(orig.)と比較して、0.8～2.7ポイント向上し、1st は(III)の場合に最大の 2.7 ポイント、ave も(III)の場合に最大の 2.6 ポイントの向上があった。また、(III)を適合率と再現率で評価すると、それぞれ最大で 5.1%（再現率 52.0%時に適合率が 44.4%から 49.5%へ）、5.6%（適合率 49.5%時に再現率が 46.2%から 52.0%へ）向上した。

10000記事での実験と比較して新聞記事1年分の方が0.5ポイント分類精度が下がっている。

5　考察

5.1　分類精度

共起情報を用いた新聞記事1年分の自動分類の結果は、1st で 2.7%、ave で 2.6%向上した。

分類精度を評価するため、10000 記事での(orig.)と(IV)による自動分類結果のうち、第1解を細かく分類した。結果を表7に示す。

表7　分類結果の分類(1st)

共起を用いた場合(IV)		正解	不正解	合計
ともに正解/不正解	分類多義語でない	392	217	609
	分類多義語	861	339	1200
	その他	287	133	420
正解になった/不正解になった		179	92	271
合計		1719	781	2500

「ともに正解/不正解」のものは(orig.)でも、(IV)でも同じ分類項目が最も類似していると判定されたことを示している。「その他」は最も寄与した単語が(orig.)と(IV)で変化したものを示している。また、「正解になった/不正解になった」は実質的な分類結果の差を示している。そのうち、「分類多義語でない」ものは、特定の 1 つの分類項目にのみ多く出現した単語が分類に最も寄与したこととなり、本手法によって精度向上が望めない部分と考えられる。従って、それ以外の部分が実質的には本手法によって効果を生む部分と考えられる。

「分類多義語でない」部分を向上させるためには、別のアプローチが必要と考える。

5.2　分類多義語の多義解消

分類多義語の多義解消がされているかどうか確認するため新聞記事1年分での(III)の分類結果について、単語「大統領」に着目し頻度分配の割合（平均）を計算した。新聞記事1年分では分類多義語「大統領」は「大統領_{＜大統領選挙＞}」、「大統領_{＜首相＞}」、「大統領_{＜外交関係＞}」及び、「大統領_{＜首脳会談＞}」の4つの分類多義分割単語になる。そこで、「大統領」が出現する全ての分類対象記事に対して、それぞれ、＜大統領選挙＞、＜首相＞、＜外交関係＞及び、＜首脳会談＞が正解分類先に含まれるかどうかを判定し、分類先である場合と、ない場合で別々に集計した。結果を表8に示す。なお、括弧内は理想実験の場合の割合を示している。

＜大統領選挙＞に分類すべき記事に対しては頻度分配の結果として分類対象文書中の単語「大統領」の頻度の 45%が「大統領_{＜大統領選挙＞}」

表8　分類多義語の多義解消の例

分類先かどうか / 分類項目	分類先である	分類先でない
＜大統領選挙＞	45%(60%)	16%(18%)
＜首相＞	30%(46%)	28%(21%)
＜外交関係＞	29%(41%)	25%(19%)
＜首脳会談＞	31%(34%)	23%(17%)

に与えられ、分類すべきでない記事に対しては16%しか与えられなかった。その他の分類多義拡張単語についても、値の差は「大統領＜大統領選挙＞」ほど大きくないが、いずれも分類すべき記事の方に大きな値が分配されている。しかし、理想の場合と比較すると差が小さいため、精度を十分に向上できなかった。

5.3　不正解の分析

朝日新聞の記事の中には以下に示すような特徴の記事があり、これが、従来の分類方式でも、本提案の方式でも自動分類の精度を下げている原因の1つであることが分かった。例えば、

- 記事の種類として「書籍案内」の記事がある。これらの記事には必ず分類項目＜図書＞と＜読書＞が付与され、一部それ以外に、著者の身分を示す分類項目や、書籍の内容に関する分類項目が付与されている。記事の内容からは、学習時に＜図書＞と＜読書＞を特徴付けられず、分類時に＜図書＞と＜読書＞の類似度が上がらず分類精度が低い。(107記事)

	1st	ave
従来(orig.)	35.5	32.6
共起を用いた場合(IV)	28.9	38.1

- 死亡記事では、死亡した人物の身分に関する分類項目が付与されているが、このような記事でよく取り上げられる教授や国会議員の分類項目＜教職員＞と＜国会議員＞が単語「死去」の重要分類項目となるため分類精度が低い。(50記事)

	1st	ave
従来(orig.)	42.0	33.8
共起を用いた場合(IV)	40.0	34.5

などが挙げられる。

また、分類多義語の選択方式から、分類多義語の解消を行ってもうまくいかない例として、

- 『「風の子学園」園長を追起訴　別の園児4人にも体罰』という記事では、正解分類先は＜学校教育＞、＜捜査＞、＜殺人＞および、＜監禁人質＞である。従来(orig.)の分類では、単語「広島」が分類先決定に大きく影響して＜広島長崎＞という分類項目が最も類似していると判定された。分類多義語として、「広島」は、＜プロ野球＞、＜高校野球＞、＜原水禁運動＞および、＜広島長崎＞の分類先しかなく、共起計算によっても＜広島長崎＞に全ての頻度が与えられ、分類項目＜プロ野球＞、＜高校野球＞および、＜原水禁運動＞は候補から消えたが、＜広島長崎＞は残った。

が挙げられる。

5.4　今後の課題

5.3に挙げた不正解の原因の様に、記事の特徴によるものもあるが、それ以外に5.2で示したように分類多義語の解消が十分でなく誤分類していると考えられるものがある。今後は、例えば係り受けのような言語的な情報で共起情報を強めることで、自動分類の精度向上を目指すことが課題である。

6　まとめ

今回、自動分類に際して、700程度の細かい分類を行い、自動分類の精度を確認した。現状ではまだ不十分な点もあり、今後は精度向上のために「今後の課題」で述べた内容について実験を行いたい。

なお、本研究にあたり、朝日新聞社電子電波メディア局の関係者の方々に新聞記事の利用を了解いただいた。

参考文献

[1] 河合:意味属性の学習結果に基づく文書自動分類方式，情報処理学会論文誌，Vol.33，No.9, pp.1112-1122(1992).
[2] 湯浅：大量文書データ中の単語共起を利用した文書分類,情報処理学会論文誌，Vol.36，No.8, pp.1819-1827(1995).
[3] 朝日新聞記事データベース(1991年9月～1992年8月).
[4] 松本他：日本語形態素解析システムJUMAN使用説明書 version 2.0(1994).
[5] EDR電子化辞書 日本語単語辞書1.5版，（株）日本電子化辞書研究所(1995).