修 士 論 文

日本語テキストとテキスト画像の自動分類に関する研究

平成 22 年度修了
三重大学大学院工学研究科
博士前期課程 情報工学専攻

粉川　昭宏

はじめに

テキスト自動分類とは，コンピュータを用いてテキストをカテゴリー別に分類することである．近年，携帯電話やパソコンの普及により，ネットワーク上で利用できる情報の量が格段に増加した．この情報の量の増加により，利用者は必要としている情報を迅速かつ適切に抽出したい．そのためテキスト自動分類 [1] の技術の重要性が高まっている．テキスト自動分類技術は，例えば，ネットワーク上の情報を収集・整理する段階で利用されることで，適切な情報抽出の実現に寄与する．従来のテキストでは，主に人手で行われていた．しかし大量のテキストを分類するには多くの時間と費用がかかり，カテゴリ付与の一貫性を保つことが困難である．自動分類を用いることで，テキストのカテゴリ化が自動化・客観化されるので，カテゴリ化の一貫性が保たれ，テキストを効率的に検索できる．本研究では日本語新聞記事を用いて，日本語テキストの自動分類の研究を行った．

紙媒体上のテキスト自動分類では，スキャナによってテキスト画像をコンピュタ内に取り込み，光学文字認識（Optical Character Recognition：OCR）によって文字コードに変換する必要がある．OCR による文字認識の研究 [2][3] の歴史は古いが文字認識率は 100%ではないため，認識エラー（誤読）が発生する．本研究では，日本語テキスト分類の実用化に向けて，認識エラーがテキスト自動分類にどのような影響を及ぼすかを明らかにする．

日本語テキスト自動分類ではテキストから文字や文字列または単語を抽出し，それらの出現度数を特徴量として分類を行う．日本語のテキスト自動分類では様々な特徴抽出方法が報告されている．しかし，これまでの報告では分類で使用するテキスト，カテゴリ数や分類器が異なるため，どの特徴抽出方法が望ましいか判断できない．本研究では，漢字 N グラムと品詞の出現頻度を特徴ベクトルとする手法を提案し，その有効性を比較実験によって明らかにする．

実験には，学習用記事 1000 件，評価用記事 500 件を用いた．実験の結果，実験には特徴として漢字モノグラム＋全品詞を用いるとテキスト分類率は 80.6%と最も良い結果が得られた．また漢字モノグラムがテキスト分類に最も寄与している特徴であることがわかった．また，あらかじめ文字コード化された記事の分類率は 80.6%，紙媒体の記事を OCR

で文字コード化した場合のテキスト分類率は 80.9%であり，OCR による日本語テキスト画像の自動分類が実用的であることがわかった．

今後の課題として，実験で用いる記事数を増加させテキスト分類実験を行うことがあげられる．また日本語や英語以外の他の言語に対しテキスト分類を行い，結果を比較・検討することが残されている．

目次

第 1 章

序論

1.1 研究の背景

近年，通信技術を含む計算機ハードウェアの急速な進歩により，ネットワーク上で利用できる情報の量が格段に増加した．この情報の量の増加により，利用者が必要としている情報を迅速かつ適切に抽出する技術の重要性が高まっている．テキスト自動分類は，自動索引付け，文書の組織化，テキストフィルタリングや Web ページの階層化分類と幅広い領域で応用が可能である．自動分類を用いることで，テキストのカテゴリ化が自動化・客観化されるので，カテゴリ化の一貫性が保たれ，テキストを効率的に利用することができる。

日本語のテキスト自動分類では，漢字の度数分布を利用する方法と，品詞（単語）の度数分布を利用する方法がある．渡辺 [4] は漢字，呉ら [5] は名詞，平ら [6] は 5 種類の品詞の度数分布を特徴ベクトルとして利用している．これらの文献は，分類で使用するテキスト，カテゴリー数や分類方法が異なるため，どの特徴抽出方法が望ましいが判断できない．

従来のテキスト自動分類はコンピュータで行われており，分類させるテキストも電子化（文字コード化）テキストを扱っていた．しかし日常生活においてテキスト分類させたいテキストは，本，書類，新聞のように印刷された紙媒体が多く存在する．紙媒体を文字コード化テキストに変換するためには紙媒体を，スキャナでデジタル画像として取り込み，そのデジタル画像を光学文字認識（Optical Character Recognition：OCR）技術を用いて，文字コードに変換する．OCR における文字認識の研究 [2][3] の歴史は古いが文字認識率は 100%ではない．そのため文字コード変換の際に認識エラー（誤読）が発生する．

1.2　本研究の目的

本研究では，漢字，漢字 N グラム，品詞の出現頻度数を特徴ベクトルとして抽出しテキスト分類実験を行い，どの特徴ベクトルが良いか検討する．また，漢字 N グラムと品詞の特徴ベクトルの併用に関する比較・検討を行って，テキスト分類の精度向上を目指す．

日本語テキスト分類の実用化に向けて，紙媒体のテキスト自動分類実験を行う．英文のテキスト自動分類 [7][8] では，紙媒体から OCR を用いてテキスト分類実験を行い，テキスト分類性能への影響が調査され報告されている．本研究では，日本語テキストの場合に OCR の文字認識誤りがテキスト分類の性能にどのような影響及ぼすかを明らかにする．

1.3　論文の構成

本論文の第 2 章，第 3 章，第 4 章では, 本研究におけるテキスト分類の流れ，各処理の詳細について述べる．さらに第 5 章では, 実験の結果と考察を述べる．最後に第 6 章で, 本研究のまとめと今後の課題について述べる．

第２章

特徴抽出

テキスト自動分類では記事から文字や文字列または単語を抽出し，それらの出現度数を特徴ベクトルとして用いて分類を行う（図 2.1）．本研究では漢字 N グラムを用い，単語として種々の品詞を用いる特徴ベクトルとそれらの組み合わせによる特徴抽出方法について説明する．

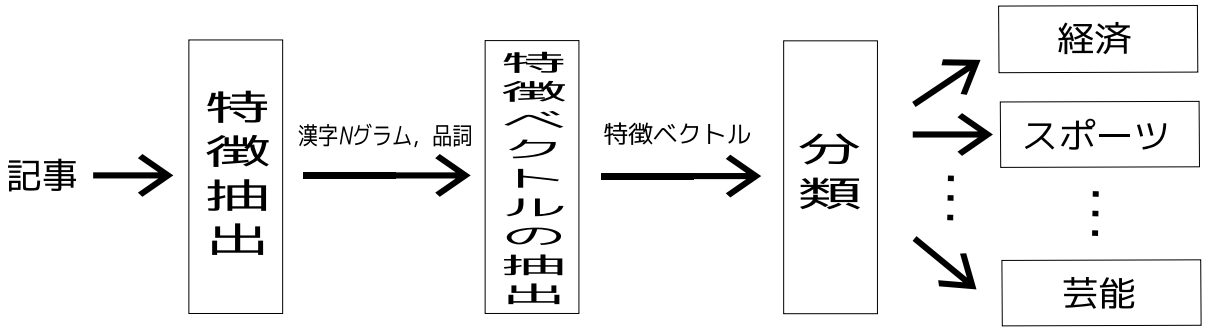


図 2.1　テキスト自動分類の流れ

例（１）政府がアイルランドの民族自決権を認めた。
（２）新政府として民族団結政府が発足した。

図 2.2　例文

2.1 漢字 N グラム

漢字 N グラムとは，文章中で連続する N 個の漢字文字列のことである．漢字には，平仮名やカタカナ，句読点や括弧等の記号，アルファベット，数字（漢数字を含む）を含めないものとする．$N = 1$ の場合を漢字モノグラム（漢字），$N = 2$ の場合を漢字バイグラムと呼び本研究ではこれら 2 つの場合について検討する．また N を漢字列の長さと呼ぶ．図 2.2 の例文に対し漢字モノグラム，漢字バイグラムを抽出した結果を図 2.3，図 2.4 に示す．

（１）政 府 民 族 自 決 権 認
（２）新 政 府 民 族 団 結 政 府 発 足

図 2.3 漢字モノグラムの抽出結果

（１）政府 民族 族自 自決 決権
（２）新政 政府 民族 族団 団結 結政 政府 発足

図 2.4 漢字バイグラムの抽出結果

2.2　品詞

品詞を抽出する場合は形態素解析ツール Chasen[9] によって文書中の品詞を決定し特定の品詞を抽出する．「政府がアイルランドの民族自決権を認めた。」を形態素解析すると，「政府（一般名詞）」，「が（助詞）」，「アイルランド（固有名詞）」，「の（助詞）」，「民族（一般名詞）」，「自決（サ変接続名詞）」，「権（接尾名詞）」，「を（助詞）」，「認め（動詞）」，「た（助動詞）」，「。（記号）」となる．品詞の抽出では形態素解析された記事から特定の品詞を抽出する．全品詞（POS）を抽出する場合，本研究では一般名詞，サ変接続名詞，固有名詞，形容詞，副詞，形容動詞，動詞の 7 種類の品詞を抽出する．図 2.2 の例文に対し，全品詞の抽出した結果を図 2.5 に示す．本論文では，漢字 N グラムと品詞のことを特徴と呼ぶことにする．

2.3　特徴の組み合わせ

特徴を組み合わせることで特徴量が増え，テキスト分類精度の向上するが期待できる．図 2.2 の例文に対する，漢字モノグラムと全品詞を組み合わせた結果（漢字モノグラム＋全品詞）を図 2.6 に示す．

（１）政府 アイルランド 民族 自決 認め
（２）政府 民族 団結 政府 発足 し

図 2.5　全品詞の抽出結果

（１）政 府 民 族 自 決 権 認 政府 アイルランド 民族 自決 認め
（２）新 政 府 民 族 団 結 政 府 発 足 政府 民族 団結 政府 発足 し

図 2.6　漢字モノグラム＋全品詞の抽出結果

第 3 章

分類方法

この章では抽出された文字列，品詞から特徴ベクトルを抽出してテキスト分類するまでの方法と，分類性能の評価と検定手法について説明する．テキスト分類の処理の流れを図 3.1 に示す．あらかじめ学習用記事から抽出された文字列，品詞の集合を特徴辞書として作成する．その特徴辞書を用いて個々の記事における特徴の出現度数を求めて特徴ベクトルを作成する．作成された特徴ベクトルに対して変数変換，次元削減を行った後，サポートベクターマシン（SVM）で分類する．これらの詳細について次節以降で述べる．

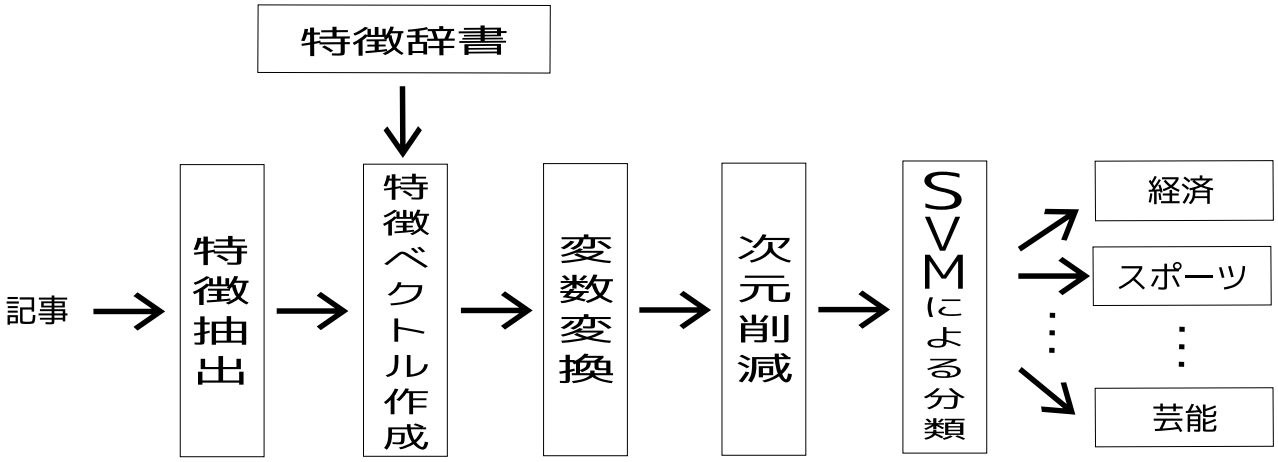


図 3.1　特徴ベクトル，分類処理の流れ

3.1　特徴辞書

特徴辞書は以下のように作成される．まず学習用記事集合に出現する全ての特徴の辞書を作成する．次に辞書中の各特徴の出現度数を求める．テキスト自動分類は，用いる特徴の増加に伴って，特徴辞書の特徴数（次元数）が増大する問題が発生する．特徴ベクトルの次元数が大きいほど，後の処理に要する処理時間が増加する．特徴ベクトルの次元数が不十分だと，分類精度が低下する．そのため分類精度の低下を招かない範囲で特徴辞書の次元数を削減する必要がある．特徴辞書中の特徴数を決める方法は 2 つある．1 つは学習用記事での出現数が多い上位 n 個の特徴を選択をする方法である．本研究では n =5000 とする．他の 1 つは学習用記事での出現数が m 以下の特徴を辞書から削除する方法である．

3.2　特徴ベクトル作成

テキストの自動分類では，1 つの記事を 1 つの特徴ベクトルで表す．特徴ベクトルの要素は，辞書に存在する特徴が記事の中に出現する度数である．特徴ベクトルを次式で表す．

例（１）政府 アイルランド 民族 自決 認め
（２）政府 民族 団結 政府 発足 し

図 3.2　特徴（全品詞）例

{ 政府 アイルランド 民族 自決 認め 団結 発足 し }

図 3.3　特徴辞書の例

（１）[１１１１１０００]
（２）[２０１００１１１]

図 3.4　特徴ベクトルの例

$$X = [x_1 x_2 \cdots x_n]^T. \tag{3.1}$$

ここで n は，特徴ベクトルの次元数である．学習記事集合として図 2.2 に示す例文（図 3.2 の特徴例）を用いた場合の特徴辞書を図 3.3 に示す．また，この辞書を用いて例文（1）（2）から抽出される特徴ベクトルを図 3.4 に示す．

3.3 変数変換

分類精度を向上させるため，変数変換を行う．変数変換とは特徴ベクトルの各要素の値を変換する処理である．本研究では相対度数への変換とベキ変換とを行う．それぞれの変数変換について以下に述べる．

3.3.1 相対度数（RF）への変換

相対度数（RF）への変換とは特徴ベクトルの各要素の値を全要素の和で割る処理である．絶対度数（AF）を x_i とすると，相対度数 y_i は次式により示される．

$$y_i = \frac{x_i}{\sum_{i=1}^{n} x_i}. \tag{3.2}$$

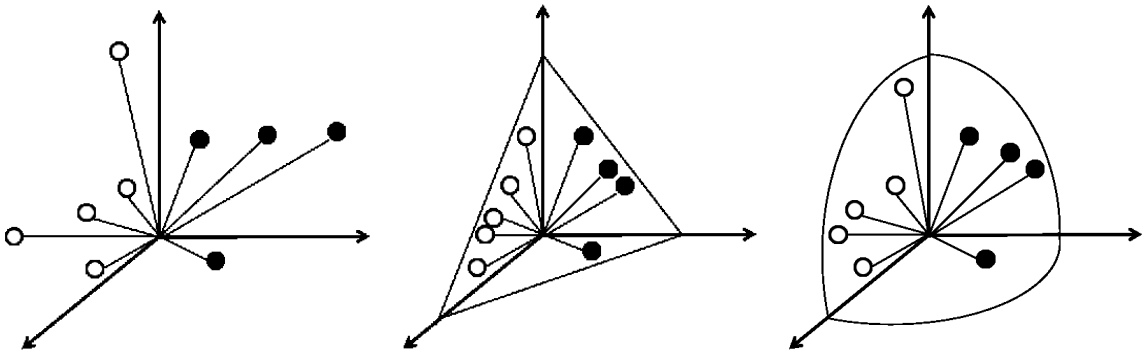
これにより，テキスト長に対して不変な特徴ベクトルが得られ，テキスト長の変動による分類性能の劣化を防ぐことができる．

3.3.2 ベキ変換（PT）

ベキ変換とは，特徴ベクトルの各要素をべき乗（v 乗）して，各要素の分布を正規分布に近づける処理である．ベキ変換は次式により示される．

$$z_i = x_i^v, \ (0 < v < 1). \tag{3.3}$$

本研究では $v = 0.5$ とする．



（1）絶対度数（AF）　（2）相対度数（RF）　（3）相対度数の 0.5 乗（RFPT）

図 3.5　変数変換による分布の変化

3.3.3　変数変換による分布の変化

変数変換による分布の変化を図で示す．白と黒の 2 クラスの場合の絶対度数（AF）の特徴ベクトルを図 3.5（1）に示す．それら特徴ベクトルを，相対度数化した結果（RF）を図 3.5（2）に示す．相対度数化した特徴ベクトルに対し，ベキ変換した結果（RFPT）を図 3.5（3）に示す．相対度数化すると特徴ベクトルが平面上に投影され，固有次元数が 1 次元減少する．相対度数を 0.5 乗すると特徴ベクトルが球面上に投影される．

3.4　特徴変換による次元削減

特徴ベクトルを分類器で学習・分類する際，特徴ベクトルの次元数が高いと計算量や記憶容量の計算資源が大量に必要となる．そのため特徴ベクトルの次元数を減少する必要がある．本研究では主成分分析（PCA）と正準判別分析（CDA）を使用し，次元削減を行う．この章では，それぞれの方法について詳しく述べる．

3.4.1 主成分分析（PCA）

PCA とは複数の変数間の共分散を除いて少数の合成変数に変換する処理である．手順を以下に示す.

まず，標本平均ベクトル M と標本の全共分散行列 S_t を次式で求める.

$$M = \frac{1}{N}\sum_{j=1}^{N} X_j, \tag{3.4}$$

$$S_t = \frac{1}{N}\sum_{j=1}^{N} (X_j - M)(X_j - M)^T. \tag{3.5}$$

そして全共分散行列に対して，その固有値 λ_i と固有ベクトル Φ_i を次式により求める.

$$S_t\Phi_i = \lambda_i\Phi_i, \ (i = 1, 2, \cdots, n). \tag{3.6}$$

求めた固有ベクトル Φ_i を基底ベクトルとする正規直行変換によって，主成分 z_i を次式によって求める．これを新たな特徴ベクトルとし次元削減を行う $(m \leq n)$.

$$z_i = \Phi_i^T X, \ (i = 1, 2, \cdots, m). \tag{3.7}$$

3.4.2 正準判別分析（CDA）

CDA はカテゴリ間の分離度（判別比）を最大化する基底ベクトルを求めて正準判別変量に変換する処理である．手順を以下に示す．

まず式 (3.4) より，標本平均ベクトル M を求め，級内共分散行列 S_w を次式により求める．但し，L はクラス数であり，N_l は各クラスのサンプル数，M_l はクラス l の平均ベクトルとする．

$$S_l = \frac{1}{N_l}\sum_{j=1}^{N_l}(X_j - M_l)(X_j - M_l)^T, \tag{3.8}$$

$$S_w = \frac{\sum_{l=1}^{L} N_l S_l}{\sum_{l=1}^{L} N_l}. \tag{3.9}$$

次に級間共分散行列 S_b を次式により求める．但し，M は全体の平均ベクトルとする．

$$S_b = \frac{\sum_{l=1}^{L} N_l (X_j - M)(X_j - M)^T}{\sum_{l=1}^{L} N_l}. \tag{3.10}$$

次に固有ベクトル行列 Φ'_i と固有値行列 λ'_i を求める．

$$S_b\Phi' = S_w\Phi'\lambda'_i, \ (i = 1, 2, \cdots, n). \tag{3.11}$$

z' を次式によって求め，これを新たな特徴ベクトルとし次元削減を行う $(m \leq n)$.

$$z'_i = \Phi_i'^T X, \ (i = 1, 2, \cdots, m). \tag{3.12}$$

$rank\{S_b\} \leq L - 1$ となるので $m \leq L - 1$ とする．

3.5　サポートベクタマシン (SVM)

本研究では，サポートベクタマシン（SVM）[10] を採用し分類する．SVM は線形識別器の 1 つであり，2 つのカテゴリの学習サンプル間のマージンが最大となる超平面（決定境界）を求める学習器である．そのために，超平面と学習データとの最小距離を評価関数として用い，これを最大にするように超平面を決定する．

SVM は本来線形識別器であるが，カーネル関数と組み合わせることによって，学習の容易さを失うことなく，非線形に容易に拡張できる．カーネル関数として線形カーネル（Linear），多項式カーネル（Polynomial），RBF カーネル（RBF）を用いる．

SVM のしきい値の変化による再現率 R と精度 P の関係を図 3.7 に示す．本研究では，F 値が最大となるしきい値で分類を行う．しかし，マクネマー検定の場合は同条件で実験する必要があるため，F 値が $R = P$ となるしきい値で分類を行う．再現率 R と精度 P と F 値，マクネマー検定の詳細は次節以降で述べる．

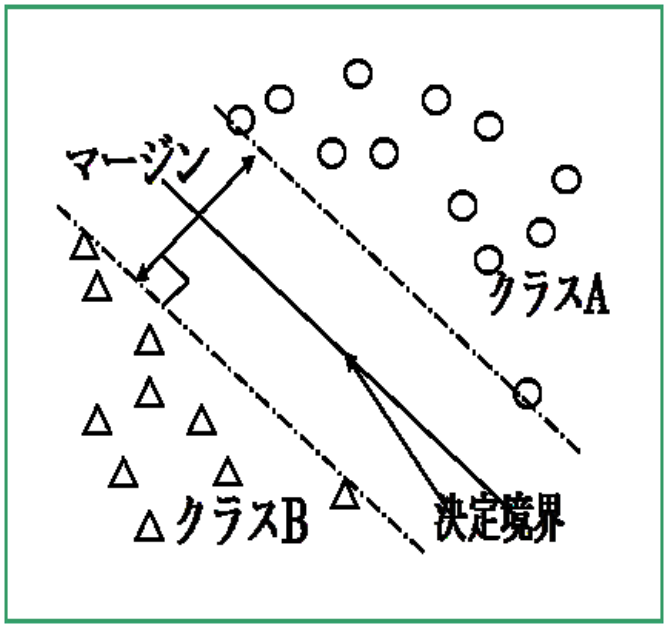


図 3.6　2 クラスの問題の決定境界とマージンの例

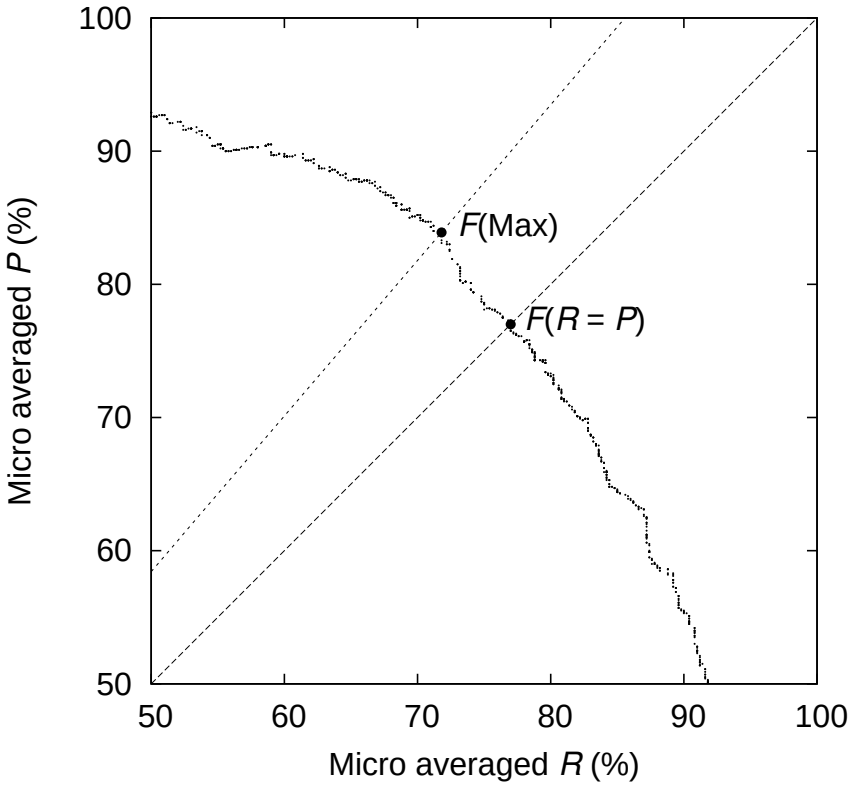


図 3.7 しきい値の変化による再現率 R と精度 P の関係

3.6 性能評価方法

性能評価方法は，以下の通りである．任意のカテゴリに対し，入力された記事がそのカテゴリに含まれるか否かを SVM により判別する．その判別結果を表 3.8 のように分類し，a から d を求める．同様の判別を全てのカテゴリで行い，a から d のそれぞれの合計を求める．次式で定義される再現率 R，精度 P，F 値を求め，F 値を評価指標とする．

$$R = \frac{a}{a+b} \times 100(\%), \tag{3.13}$$

$$P = \frac{a}{a+c} \times 100(\%), \tag{3.14}$$

$$F = \frac{2RP}{R+P}(\%). \tag{3.15}$$

表 3.8　判別結果の分類

		判別結果	
		正	誤
正解	正	a	b
	誤	c	d

3.7　統計学検定

統計学検定として本研究では，マクネマー検定を行う．マクネマー検定とは，対応のある 2 つの方法の結果に有意差があるかどうかを検定する．方法 1 と方法 2 の 2 つの分類方法でテキスト分類した場合，方法 1 や方法 2 で誤って分類されたサンプルの数を表 3.9 に従って $\hat{a}$ から $\hat{d}$ のそれぞれの合計を求め（$\hat{b} < \hat{c}$ とする），z の統計値を計算する．

$$z = \frac{\hat{c} - \hat{b}}{\sqrt{\hat{b} + \hat{c}}}. \tag{3.16}$$

2 つの方法の誤り確率に差はないという帰無仮説と方法 2 では方法 1 より優れているという対立仮説を立てるならば，z の分布は標準正規分布で近似できる．このとき $z > t = 1.28$ なら有意水準 $\alpha = 0.10$（10 %）で帰無仮説が棄却され，方法 2 は方法 1 より優れている．また $z > t = 1.65$ なら有意水準 $\alpha = 0.05$（5 %）で方法 2 は方法 1 より優れているといえる．

表 3.9　2 つの方法の性能対応表

		方法 2	
		正	誤
方法 1	正	$\hat{a}$	$\hat{b}$
	誤	$\hat{c}$	$\hat{d}$

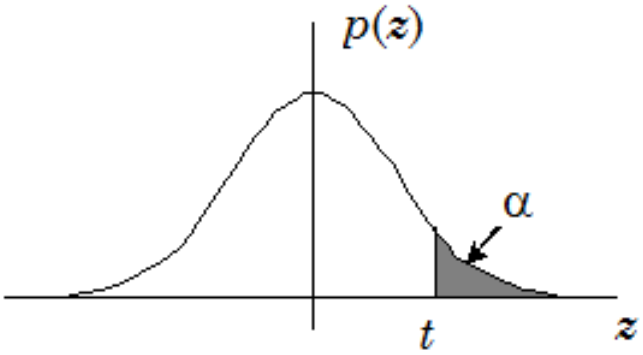


図 3.10　標準正規分布（$\alpha = 0.10,\ t = 1.28$）

第 4 章

紙媒体の文字コード化

紙媒体を文字コードに変換する方法の流れを図 4.1 に示す．紙媒体をスキャナを用いてテキスト画像として取り込み，そのデジタル画像を OCR ソフトウェアを用いて文字コードに変換する．

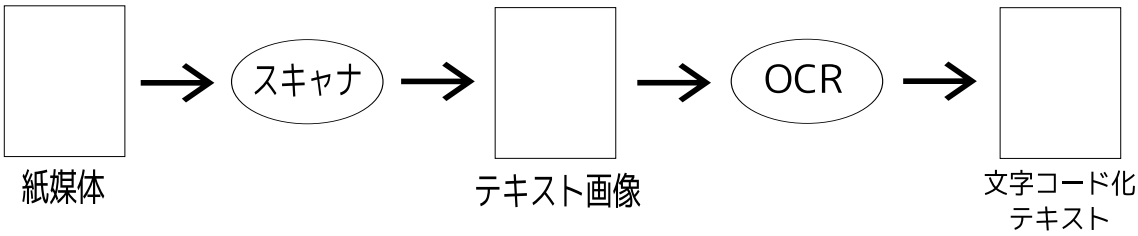


図 4.1: 紙媒体を文字コードに変換する処理の流れ

4.1　スキャナによる電子化

紙媒体を文字コードに変換するために，その紙媒体の文字をコンピュータ内に取り込まなくてはならない．そこで，紙媒体をスキャナを用いてテキスト画像としてコンピュータ内に取り入れる．OCR の認識率とテキスト自動分類率の関係を調べるために本研究では，記事をフォント MS 明朝，フォントサイズ 10point で統一し，紙媒体に印刷する．（図 4.2）また，OCR の認識率を独立変数として組織的に変化させるため印刷した紙媒体をスキャナの解像度を変化させ，テキスト画像を作成する．

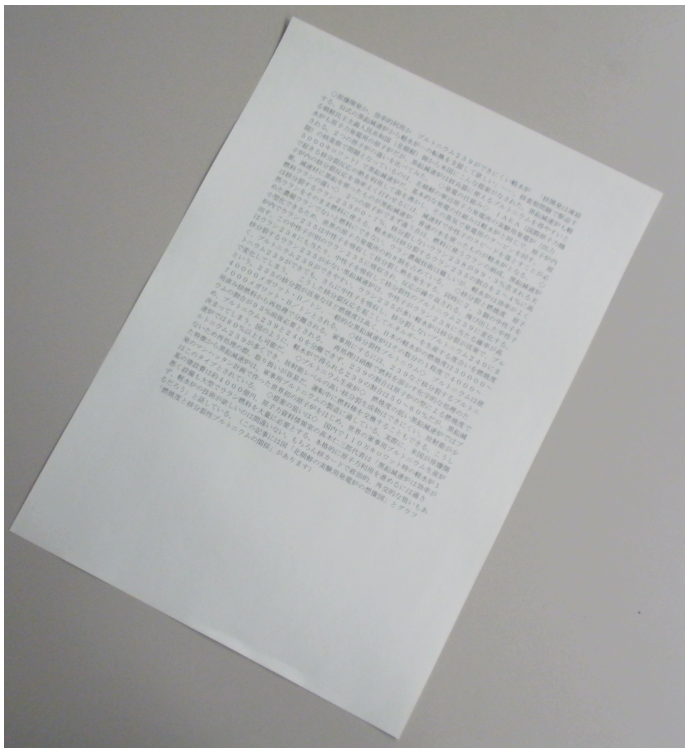


図 4.2: 記事を印刷した紙媒体

4.2　OCR による文字コード変換

スキャナにより生成されたテキスト画像を日本語 OCR により文字コードへ変換する．本研究では日本語 OCR のソフトに e.Typist v.12.0 を用いる．

OCR における文字認識は 100%ではないため，OCR の文字認識率を算出する．算出方法は OCR に読み込み前の記事と比較し，文字認識の F 値を次式により算出する．

$$F = \frac{2RP}{R+P}(\%). \tag{4.1}$$

$$R = \frac{a}{a+b} \times 100(\%), \tag{4.2}$$

$$P = \frac{a}{a+c} \times 100(\%), \tag{4.3}$$

a，b，c はそれぞれ，OCR の入出力の両方に存在する文字の数，OCR のエラーにより消失した文字の数，OCR のエラーにより新たに出現した文字の数である．

第 5 章

実験

5.1　実験で用いる記事

テキスト分類手法の学習と評価・比較のために，あらかじめ正解カテゴリが付与されたテキスト集合が必要である．本研究では，人手によって 10 カテゴリに分類されている．1994 年から 1999 年に発行された毎日新聞の記事（150 件/カテゴリ計 1500 件）を用いた．学習用記事として 100 件/カテゴリ計 1000 件，評価用記事として 50 件/カテゴリ計 500 件を用いた．

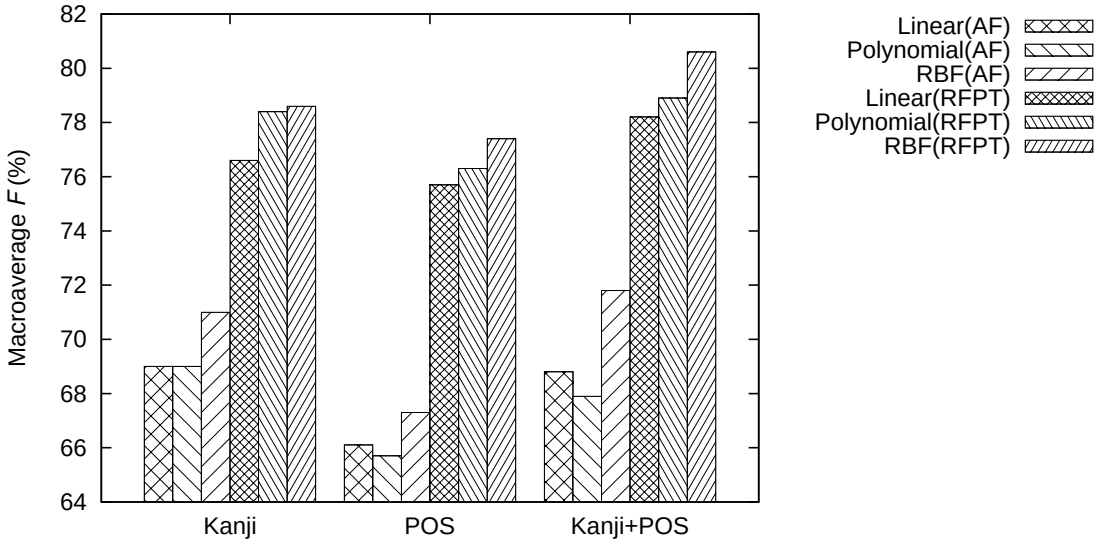
5.2　用いる特徴とテキスト分類率の関係

テキスト分類実験により，どの特徴が分類に有効か比較・検討する実験を行った．特徴としては，漢字，漢字 N グラム及び品詞とその組み合わせを用いた．

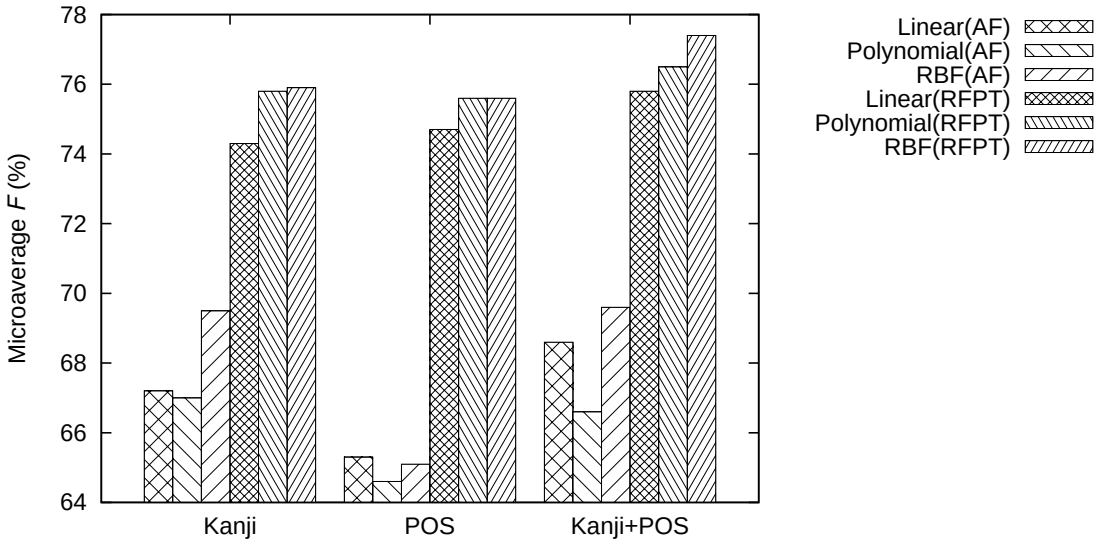
5.2.1　漢字，品詞による実験

漢字，全品詞，漢字＋全品詞の 3 通りを検討した．また変数変換を AF，RFPT の 2 通り，SVM のカーネルに linear，Polynomial，RBF の 3 通りのカーネルを用いた場合について検討する．実験結果を図 5.1 に示す．

特徴を組み合わせることでテキスト分類率が向上していることがわかる．最も良い結果は特徴として漢字＋全品詞を用いた場合の 80.6%であった．漢字と品詞を組み合わせることでテキスト分類率が向上することがわかる．漢字と品詞を組み合わせた場合と，漢字，全品詞を単独で用いた場合の有意差を統計学検定した結果を表 5.2 に示す．検定結果より有意確率は漢字＋全品詞と漢字が 2.33%，漢字＋全品詞と全品詞が 5.48%である．よって漢字＋全品詞は有意水準 5%で漢字より良いといえる．また有意水準 10%で全品詞より良いといえる．



(1) Macro average F



(2) Micro average F

図 5.1　漢字，品詞による実験結果

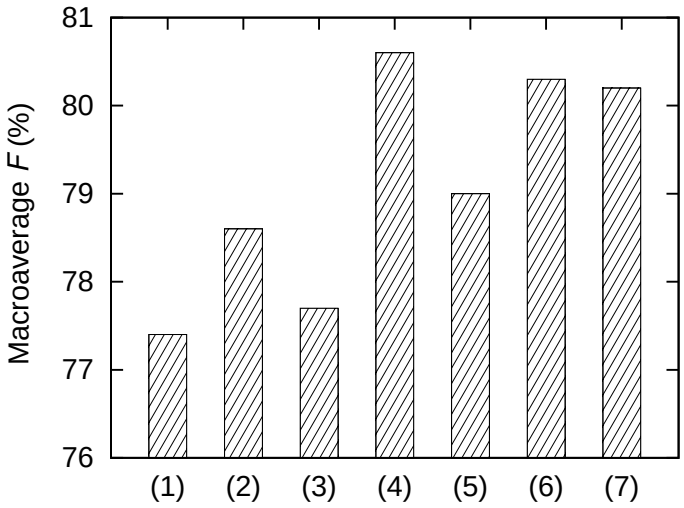
表 5.2　マクネマーの検定結果（RFPT，RBF）

	漢字 vs （漢字＋品詞）	品詞 vs （漢字＋品詞）
$\alpha(\%)$	2.33	5.48

5.2.2 漢字 N グラムと品詞による実験

特徴として，全品詞，漢字モノグラム，漢字バイグラム，漢字モノグラム＋全品詞，漢字バイグラム＋全品詞，漢字モノグラム＋漢字バイグラム，漢字モノグラム＋漢字バイグラム＋全品詞の 7 通りについて比較・検討する．実験結果を図 5.3 に示す．

漢字バイグラムを用いても漢字モノグラムと比較して良い結果が得られなかった．漢字バイグラムは漢字モノグラムと比較して特徴の種類が増加し，学習効果に悪影響を及ぼしていると考えられる．また最も良いテキスト分類率は漢字モノグラム＋全品詞の 80.6%であった．漢字モノグラム＋漢字バイグラム＋全品詞を組み合わせても分類率は向上しなかった．理由として漢字モノグラムと漢字バイグラムの特徴が似ているため，組み合わせの効果が小さいと考えられる．



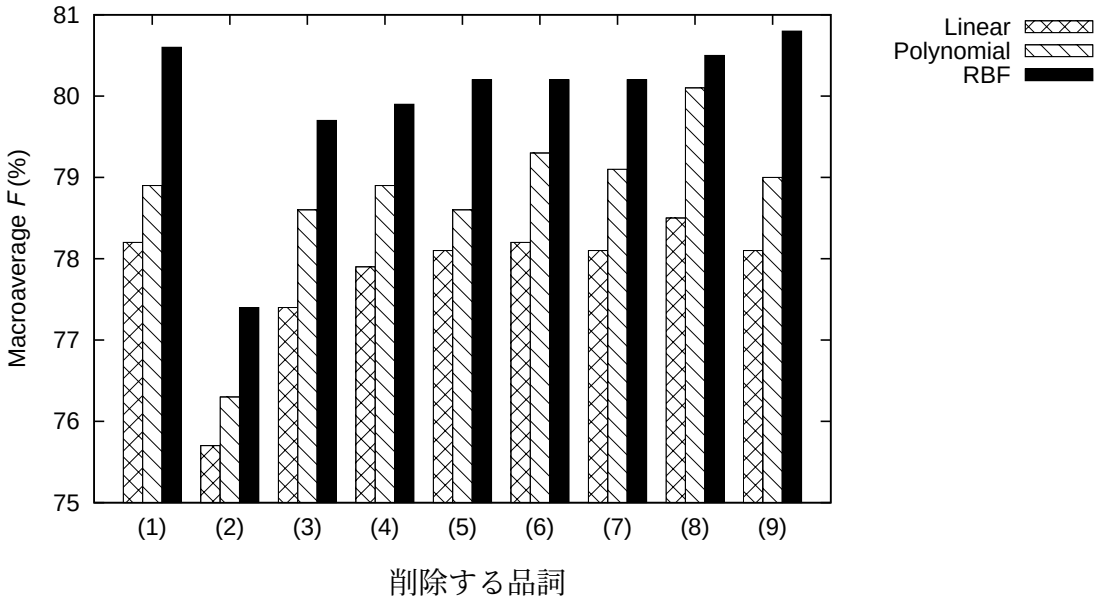
(1) 全品詞 (2) 漢字モノグラム (3) 漢字バイグラム
(4) 漢字モノグラム＋全品詞 (5) 漢字バイグラム＋全品詞
(6) 漢字モノグラム＋漢字バイグラム
(7) 漢字モノグラム＋漢字バイグラム＋全品詞

図 5.3 漢字 N グラムと品詞による実験結果

5.2.3 品詞別実験

品詞別実験では，どの品詞が分類に寄与しているか明らかにする実験を行った．特徴としては漢字+全品詞から 1 つの品詞を削除して実験を行う．SVM のカーネルを linear, Polynomial，RBF の 3 通りのカーネルを用いて比較・検討する．

実験結果を図 5.4 に示す．この結果からテキスト分類に最も寄与している品詞は一般名詞であることがわかる．また漢字モノグラムは一般名詞よりテキスト分類に寄与していることがわかる．



(1) なし （漢字＋全品詞）(2) 漢字 （全品詞）(3) 一般名詞 (4) サ変接続名詞
(5) 固有名詞 (6) 形容動詞 (7) 副詞 (8) 動詞 (9) 形容詞

図 5.4 品詞別テキスト分類実験結果

5.3 特徴辞書の削減に関する実験

この実験では，特徴辞書における特徴の出現数のしきい値 m を変化させ実験する．出現数がしきい値 m 以下の特徴を削除することによって特徴ベクトルの次元数を削減する．しきい値 m と特徴ベクトルの次元数の関係を図 5.5 に示す．しきい値 m と F 値の関係を図 5.6 に示す．全品詞を用いるとしきい値 m が大きいほど，特徴ベクトルの次元数が減少し，テキスト分類率が低下している．また漢字はしきい値 m が大きいほど，特徴ベクトルの次元数が減少するが，テキスト分類率は変化はない．漢字で出現数の小さい特徴はテキスト分類に影響を与えない．

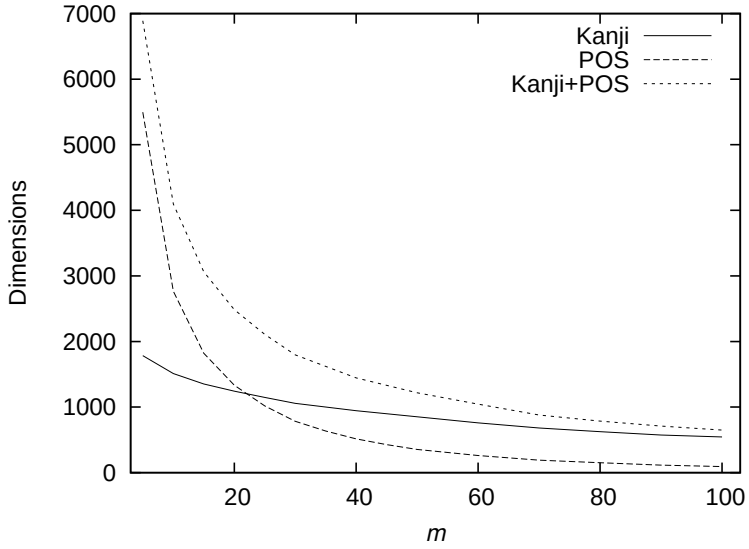
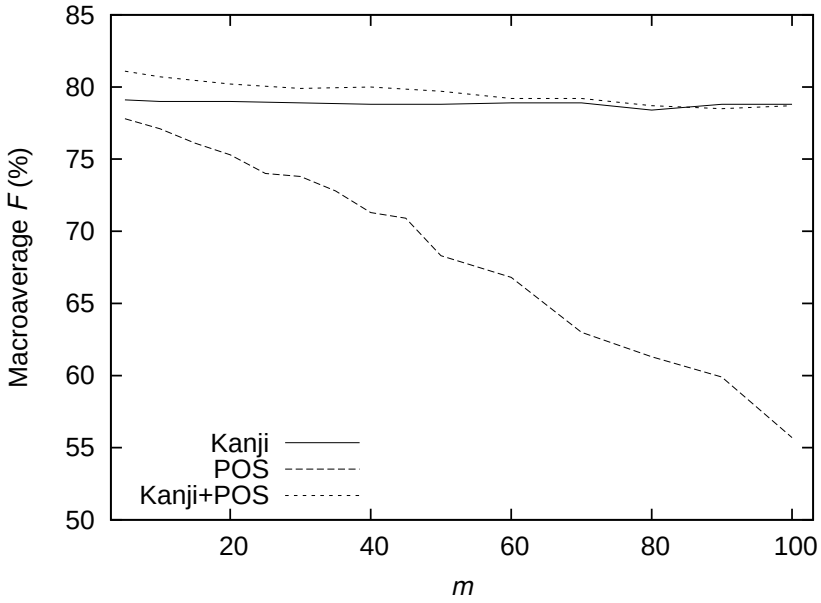
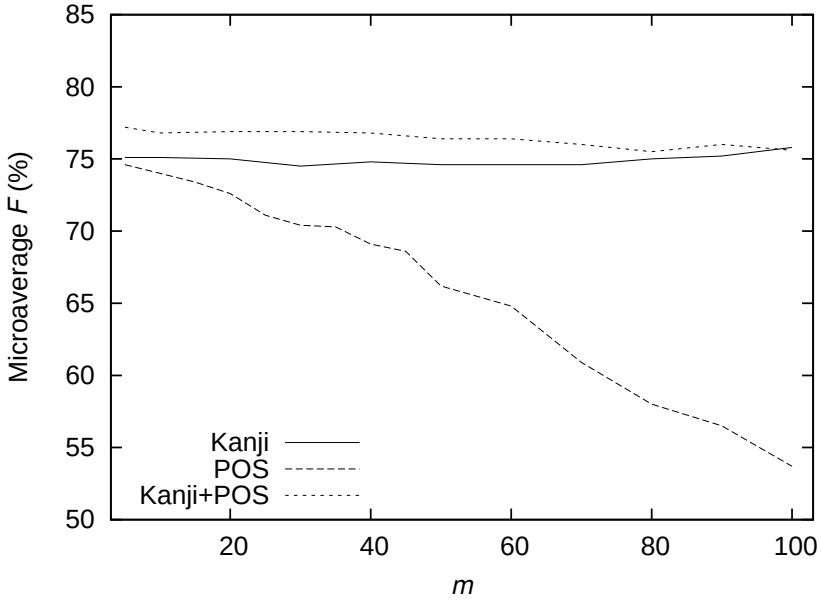


図 5.5 しきい値 m と特徴ベクトルの次元数の関係



(1) Macro average F



(2) Micro average F

図 5.6　しきい値 m と F 値の関係

5.4　次元削減に関する実験

この実験では PCA，CDA，PCA 後に CDA（PCA+CDA）の 3 通りの次元削減方法について検討する．変数変換は RFPT，SVM のカーネルには RBF を採用する．

特徴として漢字，全品詞，漢字＋全品詞の 3 通りについて検討した結果を図 5.7 に示す．次元削減では PCA を用いた場合に最も良い分類率が得られた．CDA の場合では，分類率の向上が見られなかった．理由としてカテゴリ数が 10 と小さいため，CDA 後の特徴ベクトルの次元数が 9 次元となり小さ過ぎるためであると考えられる．

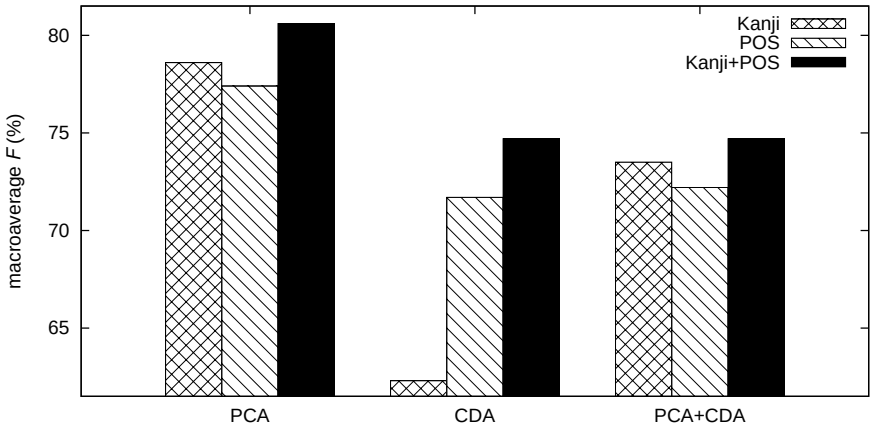


図 5.7　変数変換実験結果

5.5　OCR によるテキスト分類実験

OCR によるテキスト分類実験では以下のように実験を行う．

評価用記事は，紙媒体資料をスキャナーの解像度を 100dpi，125dpi，150dpi，200dpi，300dpi の 5 通りでスキャンし，OCR で読み込ませて文字コードに変換し用いる．

評価用記事の OCR の文字認識率（F 値）を図 5.8 に表す．解像度が 200dpi，300dpi では 98%を超える認識率であったが，解像度が 150dpi を境に認識率が大幅に低下した．

OCR 出力テキストの分類実験結果を図 5.9 に示す．特徴抽出方法の有意差をマクネマー検定した結果を表 5.10，表 5.11 に示す．漢字と漢字＋全品詞の場合は 100dpi，125dpi の場合を除いて有意水準 5%で後者が優れている．全品詞と漢字＋全品詞の場合は 100dpi，125dpi，150dpi の場合は有意水準 1%で，200dpi，300dpi の場合は有意水準 10%で後者が優れているといえる．全品詞は OCR 誤認識の影響を受けやすいが，漢字や漢字＋全品詞は相対的にその影響を受けにくいことがわかる．

文字認識率の低下とともにテキスト分類率も低下していることがわかる．文字認識率とテキスト分類率の関係を図 5.12，全品詞認識率とテキスト分類率の関係を図 5.13 に示す．全品詞は漢字より解像度が低いほど，テキスト分類率の低下が著しい．

記事をあらかじめ文字コード化されている場合，漢字＋全品詞のテキスト分類率は 80.6%である．紙媒体の記事を OCR で文字コード化した場合，漢字＋全品詞，スキャナの解像度 300dpi のテキスト分類率は 80.9%である．解像度が 300dpi 程度であれば，OCR の認識率低下によるテキスト自動分類への悪影響は認められなかった．

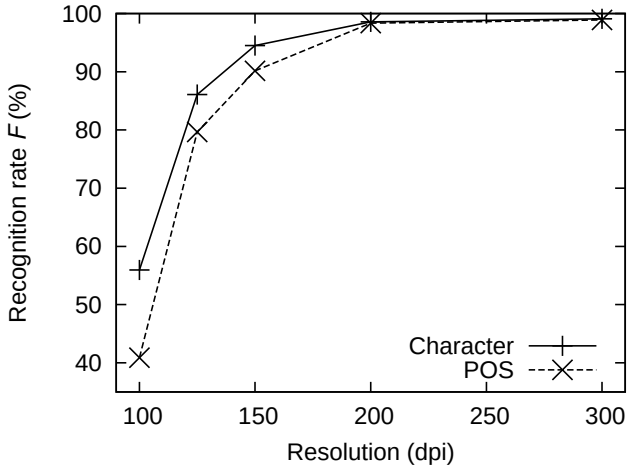
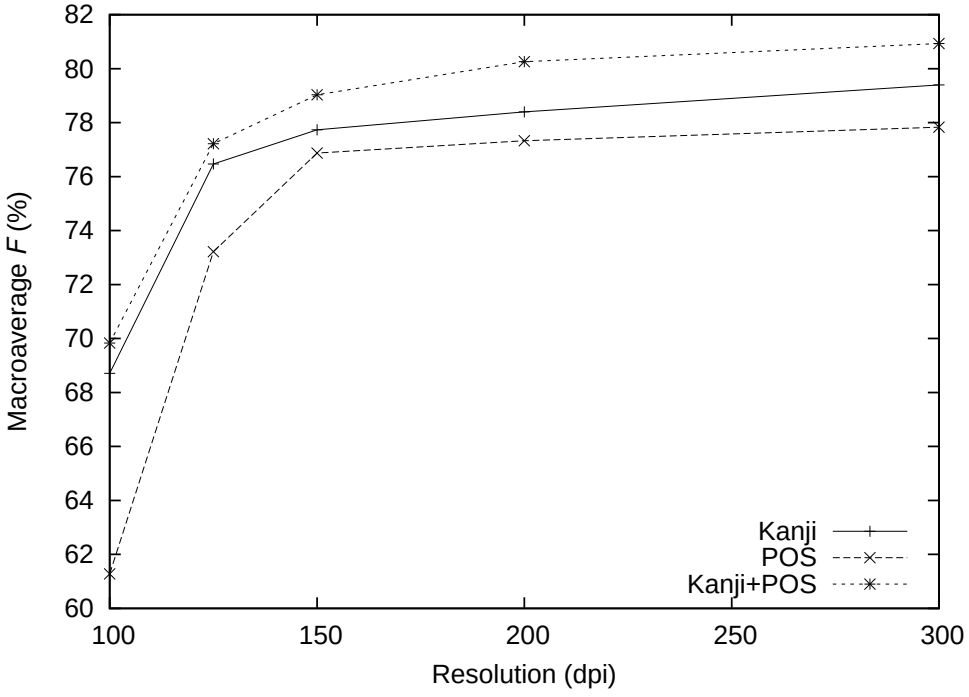
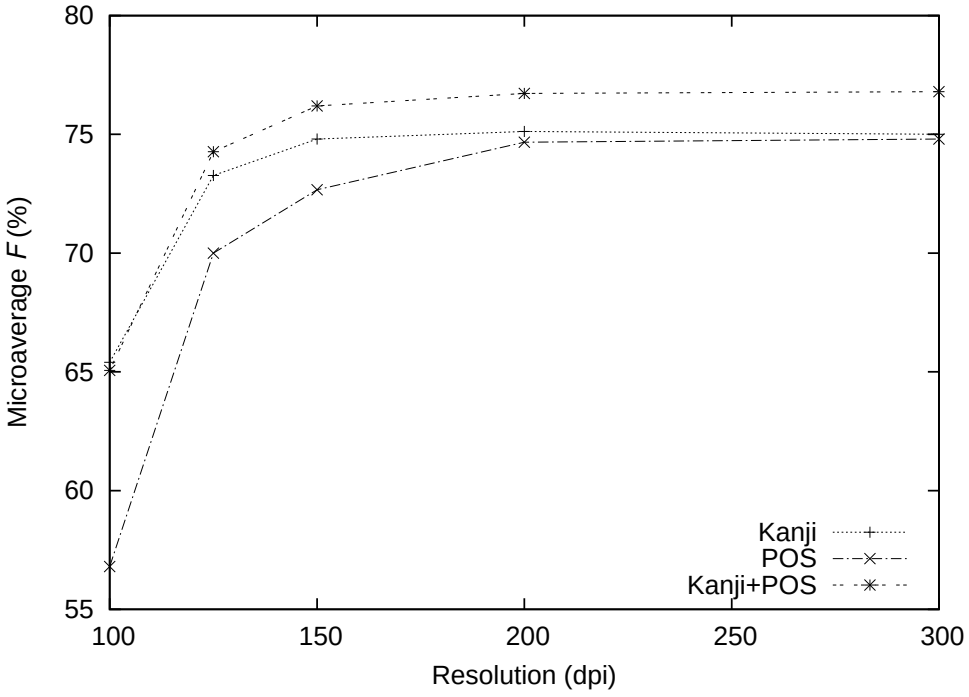


図 5.8　文字認識率（F 値）



(1) Macro average F



(2) Micro average F

図 5.9　OCR 出力テキストの分類実験

表 5.10 マクネマー検定結果 (漢字と漢字＋全品詞の比較)

Resolution	100dpi	125dpi	150dpi	200dpi	300dpi
$\alpha(\%)$	39.4	12.3	4.27	2.28	1.46
$\hat{b}$	60	32	26	24	25
$\hat{c}$	63	42	40	40	43

表 5.11 マクネマー検定結果 (品詞と漢字＋全品詞の比較)

Resolution	100dpi	125dpi	150dpi	200dpi	300dpi
$\alpha(\%)$	0.0301	0.111	0.379	5.82	5.71
$\hat{b}$	114	78	68	71	70
$\hat{c}$	197	121	103	91	90

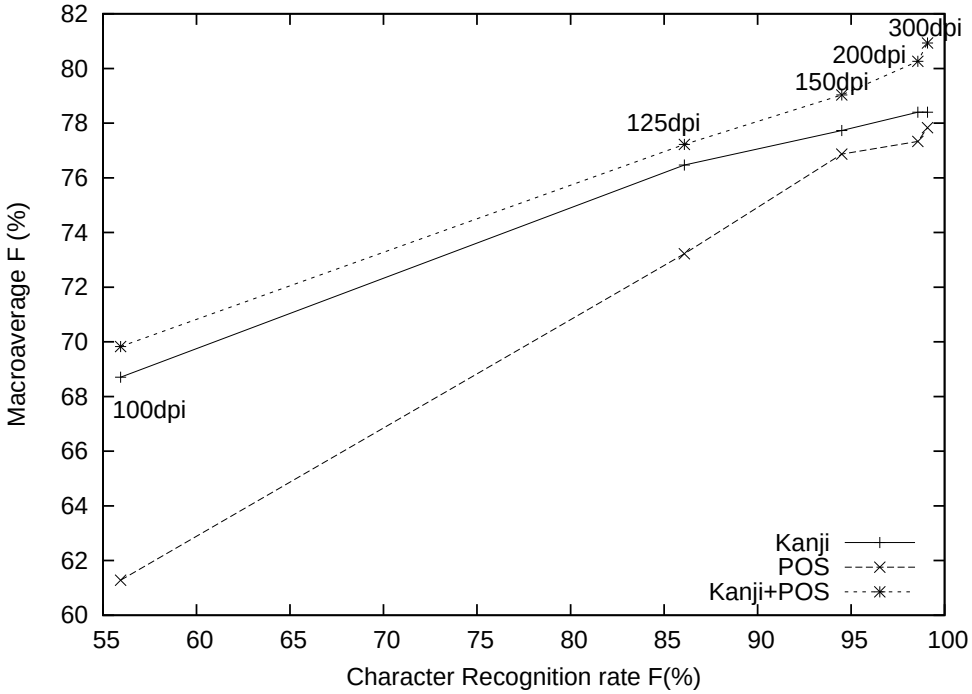


図 5.12　文字認識率とテキスト分類率の関係

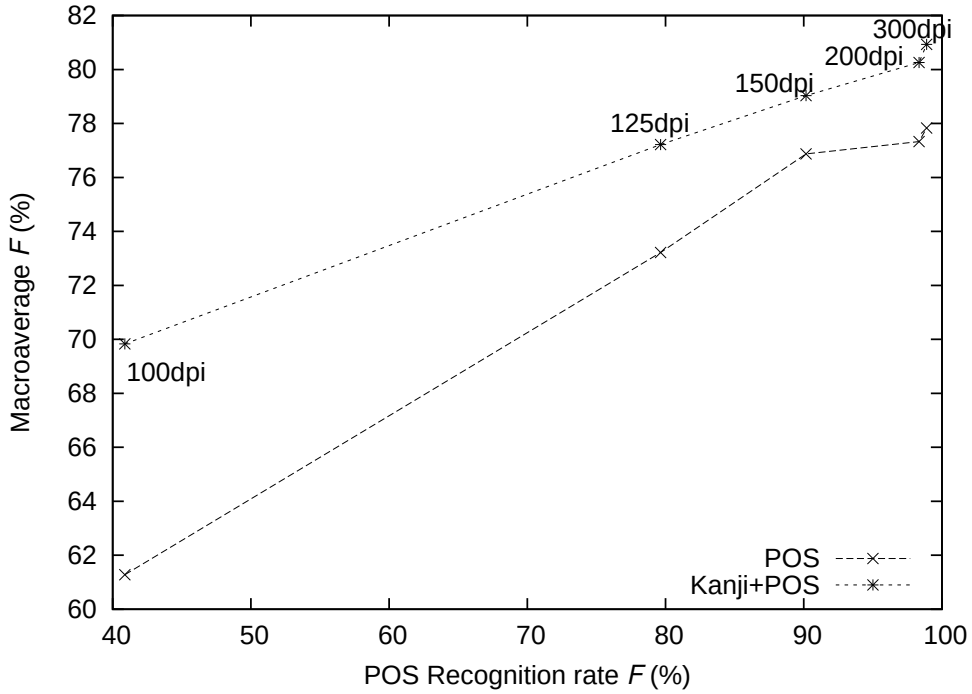


図 5.13　全品詞認識率とテキスト分類率の関係

第 6 章

結論

6.1　まとめ

本研究では，日本語テキスト自動分類における特徴抽出方法や変数変換，特徴変換を提案し，テキスト分類精度の向上と比較を行った．特徴抽出方法は漢字＋全品詞を用いた場合に最も良い分類率 80.6%が得られた．漢字と品詞の特徴を組み合わせることでテキスト分類の精度が向上することがわかった．単独の特徴としては漢字がテキスト分類に最も寄与している特徴であり，品詞の中では一般名詞がテキスト分類に最も寄与している特徴であることがわかった．

紙媒体からのテキスト分類実験では，記事があらかじめ文字コード化されている場合のテキスト分類率が 80.6%，紙媒体の記事を解像度 300dpi のスキャナで画像化して OCR で文字コード化した場合のテキスト分類率が 80.9%となり OCR の認識率低下によるテキスト自動分類への悪影響は認められなかった．また OCR の認識エラーによる形態素解析への悪影響も認められなかった．OCR による日本語テキスト画像の自動分類の実用性があることがわかった．

6.2　今後の課題

今後の課題として，実験で用いる記事のカテゴリ数やカテゴリあたりの記事数を増加させテキスト分類実験を行うことがあげられる．カテゴリあたりの記事数を増加させることでテキスト分類の精度向上が期待できる．また記事のカテゴリ数の増加により，CDA による分類率の向上も期待できる．

また日本語や英語以外の他の言語に対しテキスト分類を行い，結果の比較・検討が残されている．

付録 A

追加実験

追加実験として学習用記事数を 1000 件（100 件/カテゴリ），1250 件（250 件/カテゴリ），1400 件（140 件/カテゴリ）の 3 通りで実験を行った．学習用記事 1250 件の場合は，まず評価用記事 500 件をランダムでそれぞれ 250 件（25 件/カテゴリ）のグループ 1 とグループ 2 の 2 グループに分け，グループ 1 を学習用記事 1000 件に加え，グループ 2 を評価用記事で実験する．次にグループ 2 を学習用記事 1000 件に加え，グループ 1 を評価用記事で実験する．評価方法はそれぞれの実験結果 F 値の平均とする．学習用記事 1400 件の場合も，同様に評価用記事 500 件をランダムでそれぞれ 100 件（10 件/カテゴリ）の 5 グループに分けて実験を行う．

実験結果を図 A.1 に示す．結果より学習用記事数を増加することによりテキスト分類率が増加していることがわかる．

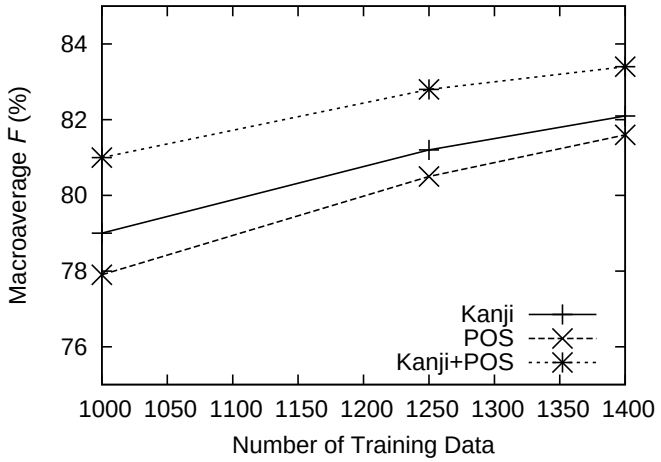


図 A.1　学習用記事数変化の実験結果（F 値）

付録 B

近年のテキスト自動分類の研究

鈴木ら [17] は提案手法に相互情報量，TFIDF，ナイーブベイズ法でそれぞれ値を算出し，それらの和で分類を行ったと報告されている．文章中から単語，単語の N-gram，文字の N-gram を抽出し特徴としている．実験結果は単語 5-gram の 89.6%である．実験で用いるテキストは新聞記事で，記事数が学習用に 7000 記事，評価用に 3500 記事と多く，カテゴリ数は 7 と少ないため，本研究より良い結果が得られていると考えられる．テキスト分類ではカテゴリが少なく，記事数が多い実験が良い分類率が得られている．

山田ら [18] は提案手法に kNN（k 近傍法）の前処理として HEOM，HVDM などの距離関数を提案し実験を行った．複数の距離関数を行うことでテキスト分類の精度が向上し，また距離関数を組み合わせることでテキスト分類の精度が向上する．実験で使用するテキスト，カテゴリ数が違うため，本研究と比較することができない．

付録 C

謝辞

本論文では様々な方の協力のもと執筆することができました．研究を進める中，終始適切な提案や助言を下さった木村文隆教授，確信のある指摘をして下さった若林哲史准教授，研究の他に教養やコンピュータの基礎知識を教え面倒を見て下さった大山航助教，様々な視点でご指摘して下さった三宅康二名誉教授に深く感謝致します．また研究で使用する機器や書物，書類の管理をして下さった田中みゆき事務官にも感謝致します．

多忙ながら私の些細な質問にも丁寧に指導して下さった Busagala さん，研究室全体を楽しい雰囲気に盛り上げた陳さん，英語の添削を丁寧にして下さった駱さん，人に対する感謝の気持ちや思いやりの大切さを教えて下さった研究室の皆さん，にも感謝致します．

最後に長きにわたり日常の生活を支えて両親や友人，私に関わった全ての人々に感謝し，本論文の結びと致します．

参考文献

[1] 祖国威，大山航，若林哲史，木村 文隆 :“ 統計的分類手法による英文新聞記事のテキスト自動分類 ”，電気学会論文誌 C Vol.124 No.3，2004，pp.852-860

[2] 三宅康二 :“ 実用になる OCR 実現のための基礎的研究ー総合報告ー “ ，情報科学リサーチジャーナル，Vol.14，2007，pp.31-55

[3] 三宅康二 :“ 実用性の高い OCR の実現を求める基的研究“ ，電子情報通信学会，Vol.3，2007，pp.55-60

[4] 渡辺靖治，竹内雅人，村田真樹，長尾真 :“ x^2 法を用いた重要漢字の自動抽出と文献の自動分類 ” 情報学基礎 39-4，1995，pp.25-32

[5] 呉勇，山田祥，岸本陽次郎 :“ 名詞頻度を使った分類用辞書の構築と評価 ”，電子情報通信学会論文誌 D-1 Vol. J84-D-I No.2，2001，pp.213-221

[6] 平博順，春野雅彦 :“ Support Vector Machine によるテキスト自動における属性選択 ”，情報処理学会論文誌，Vol.41，No4，2000，pp1113-1123

[7] Mayo MURATA, Lazaro S.P BUSAGALA, Wataru OHYAMA, Tetsushi WAKABAYASHI, Fumitaka KIMURA: “ The Impact of OCR Accuracy and Feature Transformation on Automatic Text Classification ” : Document Analysis Systems VII by Bunke, H. and Spdz, A.(Eds.) DAS 2006, pp.506-517

[8] Gudila Paul Moshi, Lazaro S.P. Busagala, Wataru Ohyama, Tetsushi Wakabayashi and Fumitaka Kimura: “ An impact of linguistic features on automated classification of OCR texts ” : Proceedings of the 9th IAPR International Workshop on Document Analysis Systems, 2010，pp.287-292

[9] 松本裕治 :”形態素解析システム「茶筌」”，情報処理，Vol.41，No.11，pp. 1208-1214，2000.

[10] 津田宏治 :“ サポートベクタマシンとは何か ”，信学論誌，Vol.83，No.6，2000，pp.623-633

[11] Lazaro S.P. Busagala, ”Improving Automatic Text Classification by Integrated Feature Analysis”，IEICE TRANSACTIONS on Information and Systems Vol.E91-D No.4

pp.1101-1109，2008.

[12] 村田真代：”OCR 文字認識率テキスト自動分類におよぼす影響に関する研究”，三重大学大学院工学研究科，平成 17 年度修士論文.

[13] 野村愛：”英文 OCR の高精度化”，三重大学工学部情報工学科，平成 19 年度卒業論文.

[14] 堤智英：”日本語新聞記事のテキスト自動分類に関する研究”，三重大学工学部情報工学科，平成 19 年度卒業論文.

[15] 馬場こづえ，藤井 敦，石川徹也：”小説テキストを対象としたジャンル推定と人物抽出”，言語処理学会，第 11 回年次大会発表論文集，pp. 157160，2005

[16] 和泉潔，松井宏樹，松尾豊：”人工市場とテキストマイニングの融合による市場分析”，人工知能学会論文誌，Vol.22 No.4 pp.397-404,2007

[17] 鈴木誠，平澤茂一，”単語と $N-gram$ の各カテゴリにおける出現頻度の比の和を用いたテキスト自動分類手法”：電子学会論文誌 C Vol.129，No.1，2009，pp.118-124

[18] 山田貴大，石井直宏，中島豊四郎：”重みを用いた距離関数の結合によるテキスト分類”，電子学会論文誌 C Vol.127，No.12，2007，pp.2077-2085

[19] 高須淳宏，相原 健郎：”テキスト分類における訓練データと性能の実験的考察”，電子文書処理，NII Journal，No.6，2003.3

[20] 相澤彰子：”低頻度語の利用によるテキスト分類性能の改善と評価”，情報処理学会論文誌，Vol.44，No.7，pp1720-1730

[21] 橋本泰一，村田浩司，乾孝司，内海和夫，石川正道：”文書クラスタリングによるトピック抽出および課題発見” 社会技術研究論文集，Vol.5，216-226，2008，pp216-226