情 報 学 基 礎　68－12
自 然 言 語 処 理 151－12
（２００２．９．１８）

日本語テキストの自動分類のための特徴素抽出手法の比較

石田栄美　辻慶太
国立情報学研究所
〒101-8430　東京都千代田区一ツ橋 2-1-2
{ emi, keita }@nii.ac.jp

日本語テキストを対象に、自動分類において、分類の手がかりとなる特徴素の抽出手法の分類性能を比較した。本実験では、テキストから特徴素を抽出する手法として、形態素解析を用いて抽出する単語ベースの方法とN-gramによって抽出する文字列ベースの方法をもとにした6つの手法を比較した。その結果、単語ベースの方法に比べ、bigram、trigramなど文字列ベースによる特徴素抽出手法を用いた場合の分類性能が高かった。さらに、分類性能に影響した可能性がある特徴素を定義し、それらの特徴素を文字種ごとに分類した結果、漢字のみからなる特徴素、漢字と助詞の組み合わせ、漢字と記号の組み合わせからなる特徴素の割合が高いことがわかった。

A Comparison of Feature Extraction for Japanese Text Categorization

Emi Ishida, Keita Tsuji
NII (National Institute of Informatics)
2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo, 101-8430, Japan
{ emi, keita }@nii.ac.jp

In the present paper we examine the influence of features on the performance of text categorization. There are two types of approaches in the methods for extracting features form Japanese texts. One is to segment texts using morphological analyzer and extract ‘words’. The other is a simple one, to use N-gram. In this experiment, five methods based on these two approaches are examined; (1) word (which is segmented by morphological analyzer), (2) β-unit (which is segmented manually based on some Japanese segmentation rule), (3)unigram, (4) bigram, (5) trigram and (6) 4gram. The categorization result based on (4) was better than those based on (1) and (2). We examined the reason and found the combination of kanji-character and hiragana particle or kanji-character and symbol has some influence.

1. はじめに

テキストの自動分類(text categorization)は、あらかじめ決められたカテゴリにテキストを自動的に分類することである。これは、既存のカテゴリに分類された学習用データセットを用いて各カテゴリの特徴を表現し、これを用いて分類対象テキストを分類することである。

自動分類研究では、統計的手法や機械学習手法を用いた様々な自動分類手法が提案されている。Naive Bayes(Lewis(1992))、Decision Tree(Apte ら(1994))、Support Vector Machine (SVM)

(Joachims(1998))などを用いた分類手法が提案されている。

テキストの自動分類に関するほとんどの論文の分類対象は英語テキストである。英語テキストは空白によって単語が区切られているため、これらを自動分類のための手がかりとして用いることは比較的容易である。ほとんどの研究では、単語をテキストとカテゴリの関係を表す分類の手がかりとして用いている。

一方、日本語テキストは連続した文字列であるため、何らかの特徴素抽出手法を用いて、手がかりとなる特徴素を抽出しなければならない。

しかしながら、日本語テキストを対象とした自動分類実験では、十分な考察も無いまま、形態素解析システムを用いて文を分割し、名詞を用いている例が多く、どのようなタイプの特徴素が分類にとって有効であるかは明らかになっていない。

本研究の目的は、日本語テキストにとって有用な特徴素のタイプ、抽出手法を見つけ出すことである。日本語テキストの特徴素抽出手法においては、一般的に 2 つのアプローチが考えられる。1 つは、形態素解析システムを用いて、テキストから単語を抽出する手法(以下、単語ベース手法と呼ぶ)であり、もう 1 つは、N-gram を用いる方法である(以下、文字列ベース手法と呼ぶ)。これら 2 つの手法は、特徴素を意味のあるものであるか、意味のないものであるかという違いがある。

本研究では、日本語テキストの自動分類において、これら 2 つのアプローチを用い、有用な特徴素抽出手法は何か、また、なぜその特徴セットが有効であったのかを調べた。

2. 関連研究

テキストの自動分類における特徴素の抽出手法を比較した論文は少ない。Wong ら(2000)は、中国語のテキストを対象に、bigram、単語、それらの組み合わせの比較を行っている。しかしながら、日本語テキストを対象に比較した実験は無い。

日本語テキストの自動分類研究においては、藤井ら(1997)，河合 (1992)のように、形態素解析システムを用いて抽出した特徴素を用いることが一般的である。

森本ら(1996)は、新聞記事における特徴素がどのような振る舞いをするかを実験しているが、森本ら(1996)は、名詞と動詞をあらかじめ定義されたキーワードとして用いており、特徴素抽出手法は比較していない。

渡辺ら(1995)は、特徴素として、漢字 1 文字だけを用いた実験を行っている。渡辺ら(1995)は、形態素解析システムは用いてはいないが、辞書やいくつかのルールを用いて、特徴素を抽出しており、その上でどの特徴素セットがよいかということを調べている。

以上のように、特徴素セットの中から分類に有効な特徴素を選択する方法を検討する文献はあるが、単語ベースと文字列ベースのどちらの性能がよいかというような文から特徴素を切り出す特徴素抽出手法に関する考察はない。本研究では、単語ベースと文字列ベースの特徴素抽出手法を比較する。

3. 実験

本章では、実際の分類実験について述べる。3.1 では、実験で比較した 6 つの特徴素抽出手法に関して説明する。3.2 では、テストコレクションについて述べる。3.3 では、この実験で用いた自動分類システムの概要を述べる。3.4 では、結果を示す。

3.1 特徴素抽出手法

どの特徴素抽出手法がよいかを調べるために、以下の 6 つ手法に関して比較した。

(1) 単語: 形態素解析システム茶筌(茶筌(1999))を用いて特徴素を抽出した。テキストは、タグつきの単語に分けられる。この分割された単語すべてを用いた。

(2) β単位: 手作業でβ単位(国語研究所(1962))を基準に分割した。β単位とは、現代語で意味を表す最小の言語単位であり、ほぼ形態素に相当する。特徴素の長さは単語ベースに比べて短くなる。

(3) unigram: N-gram をもとにテキストを分割した。この場合、N=1 である。

(4) bigram: N-gramをもとにテキストを分割した。この場合、N=2 である。

(5) trigram: N-gram をもとにテキストを分割した。この場合、N=3 である。

(6) 4gram: N-gramをもとにテキストを分割した。この場合、N=4 である。

3.2 テストコレクション

日本語の新聞記事を用いて、テストコレクションを作成した。テキストには、「毎日新聞 CD-ROM データ

集」の 1994 年の新聞記事を用い、分類カテゴリは、毎日新聞縮刷版の記事索引で用いられているものをそのまま用いた。

1994 年の 6 月分のうち、3,006 件の見出しを実験に用いた。実験では、テストコレクションを学習用データと評価用データに分割した。学習用データはカテゴリを表現するためのものであり、評価用データは分類システムの性能を評価するためのものである。2,004 件を学習用に、1,002 件を評価用に用いた。

このデータセットは大規模でなく、また見出しは長い文を含んでいない。本研究では、WWW 上の電子メールアーカイブや Q&A のような比較的小規模なテキスト集合の自動分類が将来的に重要であると考え、それらの分類を想定しているため、見出しだけを用い、学習用に用いたデータも 2,000 件程度にした。

分類カテゴリは、第 1 階層から第 3 階層まであり、第 1 階層のカテゴリは、政治、外交、経済、労働、社会など 10 カテゴリであり、第 3 階層までを含めた総カテゴリ数は 309 である。分類には、第 3 階層まで用いた。

第 1 階層のカテゴリに割り当てられた記事数を表 1 に示す。テストコレクション全体の 4 分の 1 が「政治」に属していることがわかる。また、「政治」に次ぎ、「社会」、「国際」「スポーツ」の順で件数が多い。カテゴリごとの件数はばらつきが大きいが、学習用データと評価用データのばらつきは同様である。

表1　各カテゴリに属する記事数

カテゴリ名	学習用データ	評価用データ	合計
政治	503	257	760
外交	130	54	184
経済	256	138	394
労働	18	10	28
社会	401	210	611
国民生活	150	71	221
地方	0	0	0
文化	105	50	155
スポーツ	256	176	432
国際	349	179	528
合計	2,168	1,145	3,313

*1記事に対して複数のカテゴリが付与されている場合を含む

3.3 実験で用いた自動分類システム

本実験における分類システムは、学習フェーズと分類フェーズの 2 つのフェーズにわけることができる。学習フェーズでは、学習用データを用いて、各カテゴリの特徴をもとに分類の基準となるものを作成する。本研究では、これをカテゴリ表現とよぶ。

本実験で用いた分類システムの概要を図 1 に示す。図に示したように、分類には様々な処理が必要であるが、本研究では、テキストから特徴素を抽出する手法に焦点を当てる。

以下では、学習フェーズと評価フェーズで用いた各手法について説明する。

3.3.1 学習フェーズ

学習フェーズでは、学習用データを用いて分類の基準となるカテゴリ表現を作成するが、これを作成するためには、テキストからの特徴素の抽出、カテゴリ表現に用いる特徴素の選択、カテゴリ表現を作成するという手順が必要である。特徴素の抽出には先に述べた 6 つの手法を用いた。また、特徴素の選択は行わず、抽出した特徴素全てを用いた。

カテゴリ表現では、様々な手法が提案されている。本実験では、相対出現率による重み付け手法を用いたカテゴリ表現手法を用いた。この手法は、書名を用いて図書を日本十進分類法(NDC)カテゴリに分類した実験(石田(1998))で最も精度が高かった手法である。

カテゴリ表現は、特徴素の重みベクトルで表現される。カテゴリ C_i（i=1,2,3,・・・・,N）における特徴素の重み w_{ij} （j= 1,2,3,・・・・,M）は、以下のように表される。

$$w_{ij} = \frac{T_{ij}}{\sum_{i=1}^{N} T_{ij}}$$

ここで、 T_{ij} は、カテゴリ C_i における特徴素 t_j の出現回数である。

3.3.2 分類フェーズ

分類フェーズでは、分類対象テキストをカテゴリに分類する。このフェーズでは、テキストからベクトル表現を用いたテキスト表現を作成することが必要である。

分類対象テキスト q_i は、以下のように表される。

$$q_i = \{w_{q1}, w_{q2}, \cdots w_{qk}, \cdots, w_{qM}\}$$

ここで、 w_{qk} は、分類対象テキスト中の出現回数である。

カテゴリ C_i と分類対象テキスト q_i の類似度計算には、以下の式を用いた。

$$similarity = \sum_{j=1}^{M} w_{ij} w_{qj}$$

ここで、w_{ij} はカテゴリ C_i における重みであり、w_{qj} は、分類対象テキスト q_i 中での出現回数である。

この類似度計算により、分類対象テキストに類似しているカテゴリが降順にランキングできる。本実験では、最上位のカテゴリにテキストを分類した。

3.3.3 評価尺度

再現率、精度、F_1 値の 3 つの評価尺度を用いた。再現率(r)と精度(p)(Yang (1999))には以下の式を用いた。

$$r = a / (a + c)$$
$$p = a / (a + b)$$

ここで、a は正解と同じカテゴリに分類された件数であり、b は正解と異なるカテゴリに分類された件数であり、c は正解カテゴリに分類された異なるカテゴリに属する記事件数である。

F_1 値は、以下の式で計算する。

$$F_1 = 2rp / (r + p)$$

この場合、精度と再現率は、同じ重みを持っている。

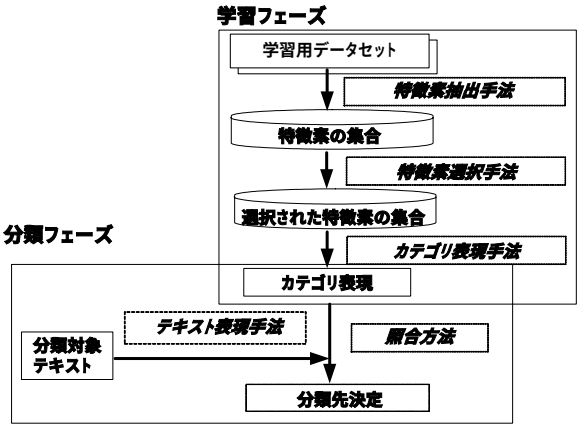

図1　実験に用いた自動分類の構成

3.4　分類結果

6 つの特徴素抽出手法を用いた分類結果を表 2 に示す。最も F_1 値が高かったのは、bigram を用いた結果である。次いで、trigram となっている。一方、unigram が最も F_1 値が低かった。全体的にみて、bigram、trigram が単語ベースよりも F_1 値が高かった。

従来の日本語テキストを対象とした自動分類研究においては、詳細な考察もないまま、ほとんどの論文で、形態素解析システムを用いた単語ベースの手法が用いられていた。しかしながら、この結果は文字列ベースによる特徴素も分類にとって有効である可能性を示唆している。

表2　各特徴素抽出手法を用いた分類結果

	特徴素抽出手法					
	単語	β単位	unigram	bigram	trigram	4gram
再現率	62.3	62.1	43.5	65.7	60.6	54.7
精度	54.5	54.4	38.1	57.9	57.5	53.8
F_1	58.2	58.0	40.6	61.6	59.0	54.2

4.　考察

先にも述べたように、単語ベースの方法は単語の意味を考慮しているのに対し、文字列ベースの方法は単語の意味を考慮していない。結果として、その特徴素セットの特徴も大きく変わってくる。

ここで、特徴素の数、特徴素の特徴という 2 点から、各アプローチの特徴素セットの違い、分類性能への影響などを考察する。

4.1　特徴素の数

各手法で抽出した特徴素数を表 3 に示す。この表から、4gram で抽出した特徴素数が最も大きく、β単位で抽出した特徴素数が最も少ないことがわかる。全体的にみて、文字列ベースによる特徴素数は多く、単語ベースの特徴素数は少ない。

さらに、表 2 と表 3 を比較すると、特徴素数と分類性能の順位は対応していないことがわかる。このことから、特徴素数が、分類性能に直接的な影響を与えている可能性は低いといえる。

表3　特徴素抽出手法による特徴素数

	特徴素抽出手法					
	単語	β単位	unigram	bigram	trigram	4gram
異なり数	5,959	5,943	1,802	19,952	32,864	37,667

4.2　特徴素が分類に与える影響

4.2.1 影響のある特徴素

次に、特徴素セットの性質が分類性能にどのような影響を与えるかを調べるために、分類性能に影響を与える可能性がある特徴素を「影響のある特徴素」と定義し、どのような性質があるかを調べた。

影響のある特徴素の特徴は、以下の手順によって、決定する。

(1) 分類対象テキスト中に出現する特徴素のカテゴリの重みを求める。そのカテゴリは、分類実験で付与されたカテゴリである。
(2) 特徴素の該当するカテゴリの総重みを求める。
(3) テキスト中に出現する特徴素の重みを高い順から加算していく。

(4) 総重みの x%以上の重みになるところでやめ、該当するものを分類に影響を与えた特徴素(影響のある特徴素)と考える。今回の分析では、x=30 とした。

ここでは、単語、β単位、bigram、trigram の特徴素セットに関して分析した。

表 4 に、各手法における影響のある特徴素の延べ数と異なり数を示す。分類対象テキストは、1 つ、もしくは 2 つの影響のある特徴素で分類されていた。また、文字列ベースの手法では、単語ベースに比べて、影響のある特徴素の数が多い。

次に、影響のある特徴素を、特徴素の長さと、漢字、カタカナ、ひらがな、記号などの文字種によって分類した。

表4　影響のある特徴素数

	特徴素抽出手法			
	単語	β単位	bigram	trigram
異なり数	652	685	1,671	2,448
延べ数	1,111	1,147	2,686	3,996

4.2.2　単語ベースにおける影響のある特徴素

単語ベースから抽出した特徴素のうち影響のある特徴素を分類した結果を表 5 に示す。この表において、「重みの平均」は、それぞれの組み合わせにおける重みの平均を示している。最大の重みは 1.0 である。「割合(異なり数)」は、それぞれの組み合わせの異なり数をもとに求めた全体の割合である。「割合(延べ数)」は、それぞれの組み合わせの出現頻度である。「組み合わせ」は、文字種の種類を示している。「記号」は記号だけで構成されている特徴素であり、「漢字＋ひらがな」は漢字とひらがなで構成されている特徴素である。さらに、漢字を含む特徴素が多かったため、漢字を含む特徴素を詳細に分類した。「漢字(1 文字)」は、特徴素が漢字 1 文字だけで構成されていることを示しており、「漢字(2 文字)」は、漢字 2 文字で構成されていることを示している。

表 5 から、以下のことがいえる。

(1) 漢字 2 文字からなる特徴素が全体の半分を占めている。

(2) カタカナだけで構成されている特徴素も比較的大きな割合を占めている。

(3)記号で構成されている特徴素の重みの平均は、とても低い。

4.2.3　文字列ベースにおける影響のある特徴素

文字列ベースから抽出した特徴素セットのうち効果的な特徴素を分類した結果を表 6 に示す。表 6 からは、以下のことがいえる。

(4) 意味を表さない漢字で構成されている特徴素も大きな割合を占める。

(5) 漢字と助詞で構成されている特徴素は、全体の 10%を占める。

(6) 漢字と記号で構成されている特徴素は、全体の 20%を占める。

4.2.4 影響のある特徴素の性質

以上のことから、特徴素の数は分類性能に影響を与えないこと、分類に影響のある特徴素は、漢字を含むもの、漢字と助詞、漢字と記号で構成された特徴素であることがわかった。

伝統的なテキストの自動分類研究においては、ほとんどの論文では名詞や漢字のようなテキストにおいて、ある役割を持つ特徴素が用いられてきた。これは、カテゴリや分類対象においてある意味をもつ単語が、効果的な特徴素として考えられているからである。しかしながら、この結果は、意味を表す漢字だけなく、意味を表さない漢字や助詞、記号など単体では主題を表現できない特徴素も分類性能に影響を与える可能性がある。つまり、テキスト分類において考慮すべき役割を果たしている。

5. 結論

本報告では、テキストの自動分類における特徴素抽出手法の比較を行い、特徴素が分類性能にどのような影響を与えているかを調べた。実験の結果は、文字列ベースの性能がよいことがわかった。さらに、それぞれの手法による影響のある特徴素を分析した結果、漢字と助詞、漢字と記号の組み合わせなども、分類性能に影響を与える可能性があった。

今回は単純な比較と分類性能に影響する可能性のある特徴素の性質を調べたが、これらが本当にどのような影響を与えているのかを実証するためのさらなる分析が必要である。

また、本実験において、カテゴリ表現において確率モデルを用いた。確率モデルは、機械学習モデルとは異なるものである。そのため、これらの結果が SVM のような表現手法でも同様であるかを確かめる必要がある。

テストコレクションはこれらの結果に大きな影響を与えると考えられるので、他のテストコレクションを用いた実験も必要である。

表5　単語ベースにおいて影響のある特徴素を分類した結果

組み合わせ	単語			β単位		
	重みの平均	割合(異なり数)(%)	割合(延べ数)(%)	重みの平均	割合(異なり数)(%)	割合(延べ数)(%)
記号	0.21	0.9%	2.8%	0.21	1.2%	2.5%
アルファベット	0.93	2.3%	1.9%	0.93	2.2%	1.8%
ひらがな	0.88	4.6%	4.2%	0.88	8.5%	7.0%
カタカナ	0.95	16.9%	19.5%	0.95	16.2%	18.8%
漢字(1文字)	0.81	9.2%	8.2%	0.75	9.3%	10.5%
漢字(2文字)	0.88	51.1%	49.8%	0.85	53.4%	52.6%
漢字(3文字)	0.83	5.2%	4.7%	0.87	2.3%	1.5%
漢字(4文字)	0.94	1.5%	1.3%	1.00	0.1%	0.2%
漢字+カタカナ	0.97	1.4%	1.6%	0.97	5.3%	4.2%
漢字+ひらがな	0.98	6.3%	5.1%	1.00	0.9%	0.1%
数字+漢字	1.00	0.2%	0.1%	0.83	0.3%	0.3%
アルファベット+数字	0.83	0.3%	0.4%	1.00	0.1%	0.1%
アルファベット+カタカナ	0.92	0.2%	0.5%	0.91	0.1%	0.4%
合計	0.86	100%	100%	0.86	100%	100%

表6　文字列ベースにおいて影響のある特徴素を分類した結果

組み合わせ	bigram			trigram		
	重みの平均	割合(異なり数)(%)	割合(延べ数)(%)	重みの平均	割合(異なり数)(%)	割合(延べ数)(%)
記号	1.00	0.4%	0.3%	0.00	0.0%	0.0%
アルファベット	0.88	1.7%	1.7%	0.95	1.0%	0.9%
ひらがな	0.92	3.4%	3.0%	0.97	2.4%	2.2%
カタカナ	0.96	7.8%	8.7%	0.97	11.3%	13.2%
数字	0.00	0.0%	0.0%	1.00	0.3%	0.2%
漢字(意味ある)	0.91	18.1%	16.8%	0.94	4.9%	4.1%
漢字(意味なし)	0.94	22.4%	23.1%	0.97	21.0%	19.8%
漢字+カタカナ	0.96	3.5%	5.5%	0.96	5.7%	8.1%
漢字+ひらがな(助詞を除く)	0.95	4.2%	3.6%	0.00	0.0%	0.0%
漢字+助詞	0.71	9.4%	7.2%	9.83	11.0%	8.2%
漢字+記号	0.96	18.3%	18.5%	0.98	20.7%	19.9%
漢字+数字	0.97	2.8%	2.3%	0.98	3.9%	4.4%
アルファベット+数字	0.33	0.1%	0.0%	0.87	0.1%	0.1%
アルファベット+記号	0.97	0.7%	0.9%	0.99	0.8%	0.8%
アルファベット+カタカナ	0.92	0.1%	0.3%	1.00	0.0%	0.0%
アルファベット+ひらがな	1.00	0.2%	0.1%	1.00	0.2%	0.2%
アルファベット+漢字	0.96	0.5%	0.4%	0.98	0.8%	0.7%
カタカナ+ひらがな	0.93	0.5%	0.6%	0.95	0.8%	1.1%
カタカナ+数字	0.80	0.5%	0.4%	0.97	0.6%	0.4%
カタカナ+記号	0.93	2.9%	3.6%	0.97	5.3%	6.6%
ひらがな+数字	0.97	1.5%	1.3%	1.00	1.0%	1.4%
数字+記号	0.94	1.2%	1.7%	0.95	1.0%	0.8%
その他	0.00	0.0%	0.0%	0.94	7.1%	6.9%
合計	0.82	100.0%	100.0%	1.27	100.0%	100.0%

引用文献

Apte, C., Damerau, F. and Weiss, S. M. (1994) “Automated Learning of Decision Rules for Text Categorization”. *ACM Transaction of Information Systems*, Vol.12, No.3, pp.223-251

Joachims, T. (1998) “Text Categorization with Support Vector Machines: Learning with Many Relevant Features”. *Proceedings of ECML-98, 10th European Conference on Machine Learning*, pp.137-142

Lewis, D. D. (1992) “An Evaluation of Phrasal and Clustered Representations on a Text Categorization Task”. *Proceedings of SIGIR, 15th ACM International Conference on Research and Development in Information Retrieval*, pp.37-50

Wong, C. K. P., Luk, R. W. P., Wong, K. F. and Kwok, K. L.(2000) “Text Categorization using Hybrid (Mined) Terms”. *Proceedings of the Fifth International Workshop on Information Retrieval with Asian Languages*, no page

石田栄美(1998) “図書をNDCカテゴリに分類する試み”, Library and Information Science. No.39, pp.31-45

河合敦夫(1992) “意味属性の学習効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.1114-1122

茶筌(1999) 形態素解析システム茶筌 2.0. http://chasen.aist-nara.ac.jp/index.html.ja

藤井洋一, 鈴木克志, 他(1997) “共起情報を利用した文書の自動分類”, 情報処理学会自然言語処理, No.118-16, pp.97-104

森本由起子, 間瀬久雄, 辻洋(1996) “記事データからの分類知識獲得に関する実験シミュレーション”, 情報処理学会デジタルドキュメント, Vol.1, No.6, pp.1-8

渡辺靖彦,村田真樹, 他(1995) “χ^2法を用いた重要漢字の自動抽出と文書の自動分類”, 情報処理学会情報学基礎, Vol.4, No.39, pp.25-32

Yang, Y.(1999) “An Evaluation of Statistical Approaches to Text Categorization”. *Journal of Information Retrieval*. Vol.1, No. 1/2, pp.67-88

国立国語研究所(1962), “現代雑誌九十種の用語用字” 国立国語研究所