用語抽出技術を利用したテキスト自動分類

[1]太田 晋　[2]美馬 秀樹
[1]コグニティブリサーチラボ株式会社
[2]東京大学大学院工学系研究科

本論文では用語抽出技術を利用したテキスト分類について述べる。テキスト分類では、低頻度語が分類において有用な手がかりとなることが知られている。しかし、低頻度語を考慮して学習を行うためには属性数を増やす必要があり、計算コストの増大を招いてしまう。そこで本研究では、属性数を減らすために用語抽出の技術を利用し、(専門)用語を属性として利用するテキスト分類手法を提案する。新聞データを利用した実験の結果、おおむね専門性が高いと思われるカテゴリに関しては、テキスト分類が専門用語を利用した特徴付けにより精度よく行なうことができることがわかった。

1 はじめに

近年、インターネットの急速な成長とともに、膨大かつ多様な情報に容易にアクセスすることが可能となった。しかし、その反面、自分の探し求めている情報を見つけだすことが非常に困難となっている。情報を容易に取得するための技術の 1 つとしてテキスト情報を自動的に分類する技術が求められている[1]。これまでテキスト自動分類では、人手で分類規則を書く方法が用いられてきたが、1990 年代に入ると、大量のテキストデータが利用可能になったことや、コンピュータの性能が大幅に向上したことから、機械学習による分類手法が用いられているようになった。

このような背景から、Naive-Bayes、決定木、k-最近接法、Support Vector Machine、ブースティング法、確率的決定リストなどさまざまな機械学習アルゴリズムが適用されている。これらの学習手法を適用する際の問題として属性選択がある。テキスト分類問題に対して学習手法を適用する場合、出現頻度の小さい単語まで考慮して学習を行なわなければ分類精度が落ちることが知られている。しかし、低頻度語を考慮して学習を行うためには属性数を増やす必要があり、計算コストの増大を招いてしまう[2]。そこで、分類に必要な単語属性だけをあらかじめ選ぶ属性選択が用いられるが、学習手法によって最適な単語属性数が異なるため、学習手法ごとに適切な属性選択基準を選ぶ必要がある。

近年、Support Vector Machine(SVM) によるテキスト分類が研究されており、従来の学習手法に比べて高い分類精度が得られることが報告されている[3]。SVM によるテキスト分類における属性選択に関する研究として、相互情報量を基準とした選択手法と、品詞を基準とした選択手法とを比較する研究がおこなわれており、品詞選択のよ

うな簡単なフィルタリング手法が有効であるということ、属性の数を増やすほど分類精度が向上することが報告されている[4]。

先に述べたように低頻度語を考慮するためには属性数を上げる必要があるが、一方で計算コストの増大を招いてしまう。そこで本研究では、テキスト分類の属性選択に用語抽出の技術を利用し、(専門)用語を属性としてテキスト分類を行うことを提案する。用語はそのテキストを特徴付ける量と考えられるため、用語を属性とすることによって属性の質を高め、より少ない属性数で分類精度を向上させることを目的として実験を行った。

2 用語認識

今回の実験では、C/NC-value[5]用語認識技術を基にした用語の自動認識を利用した。C-value とは、用語構成に関する基本語彙の組み合わせパターンと用語の対象ドメインにおける出現頻度、さらには、用語のネスティングに関する性質に注目し、スコア付けを行うことで用語の高精度な自動認識を行う。また、NC-value では、候補となる用語の実際の文書上でのコンテキスト中にある語彙とのコロケーションの情報を用いて、用語としての確からしさ(termhood)の指標を求め、求まった指標を基に候補となる用語の再順序付けを行う。我々の行った実験では、本手法により、英語、および日本語に関しても、ドメインによらず、上位の候補では 90%以上の正解率を得られることが示されている[5][6]。

3 実験

本章では日本語テキスト分類における属性選択手法として用語フィルタリングを用いた場合の実験結果について述べる。また、比較のため、形態素解析した単語のうち名詞のみをフィルタリングして同様の実験を行った。

3.1 実験設定

分類対象として毎日新聞(1994 年発行分)の 30207 記事からなる RWCP コーパス[7]を使用した。このコーパスには各記事が属するカテゴリを表す複数の UDC コード[8]が付与されている。この中からスポーツ、刑法、政府、教育、交通、軍事、国際関係、言語活動、演劇、農業の 10 カテゴリについて、訓練記事とテスト記事をそれぞれ 1000 記事ずつ選んだ。各カテゴリに属する記事数はそれぞれ 100 個ずつである。

これらの記事を C/NC-value 用語認識技術によって、カテゴリ毎に用語としての確からしさ(termhood)のスコアを算出し、カテゴリ毎にスコアの高い上位 100 個をそれぞれ選び(計 1000 個)、それらから重複を取り除き、計 870 個の異なり語を得た。

また、品詞フィルタリングについては、日本語形態素解析システム Chasen 2.2.9 [9]を用い、普通名詞、固有名詞、サ変名詞のみを選び、計 22630 個の異なり語を得た。

学習手法として Support Vector Machine (SVM) を用いた。使用したソフトウェアは TinySVM 0.09 [10]である。

3.2 実験結果

図 1・表 1 にカテゴリ毎の適合率、**図 2・表 2** に再現率を示す。

図 1 において、適合率は一部のカテゴリを除いて用語フィルタリングが名詞フィルタリングとほぼ同等の精度を出していることがわかる。一方、**図 2** において、再現率では用語フィルタリングは多くのカテゴリで名詞フィルタリングより劣っているが、軍事と国際関係のカテゴリに関しては用語フィルタリングの適合率、再現率が高い結果となっている。このカテゴリの用語を観察すると専門性の高いと思われる用語が多数含まれているため、用語抽出した属性の効果が現われていると考えられる。一方、スポーツのカテゴリでは、試合結果等を伝える記事が多く、記事中に専門性の高い用語があまり含まれていないため、適合率、再現率ともに低い結果となっていることがわかる。

4 結論

用語抽出技術を利用し、用語を属性としてテキスト分類問題に適用して実験を行った。その結果、用語を属性とした場合、名詞を属性とした場合に比べて極めて少ない属性数でほぼ同等の適合率を得ることができた。再現率の悪かったカテゴリについて今後さらに検討する必要はあるが、おおむね専門性が高いと思われるカテゴリに関しては、テキスト分類が専門用語を利用した特徴付けにより精度よく行なうことができることがわかった。

謝辞　毎日新聞 94 年版の使用に関して、記事データの研究許可をいただいた毎日新聞社に感謝いたします。

5 参考文献

[1] 永田昌明, 平　博順 “テキスト分類―学習理論の「見本市」―”, 情報処理学会誌, Vol. 42, No. 1, pp. 32-37 (2000).

[2] 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, pp. 1720-1730 (2003).

[3] 平　博順, 向内隆文, 春野雅彦 “Support Vector Machine によるテキスト分類”, 情報処理学会研究報告, 98-NL-128-24, pp. 173-180 (1998).

[4] 平　博順, 春野雅彦 “Support Vector Machine によるテキスト分類における属性選択”, 情報処理学会論文誌, Vol. 41, No. 4, pp. 1113-1122 (2000).

[5] Frantzi, K. T., Ananiadou, S. and Mima, H. “Automatic Recognition of Multi-Word Terms: the C-value/NC-value method”, International Journal on Digital Libraries, Vol. 3, No. 2, pp. 115―130 (2000).

[6] Mima, H. and Ananiadou, S., “An Application and Evaluation of the C/NC-value Approach for the Automatic term Recognition of Multi-Word units in Japanese”, International Journal on Terminology, Vol. 6(2), pp. 175-194 (2001).

[7] 豊浦　潤, 徳永健伸, 井佐原均, 岡　隆一 “RWC における分類コード付きテキストデータベースの開発”, 電子情報通信学会研究報告, NLC96-13, pp. 27-32 (1996).

[8] Universal Decimal Classification Consortium http://www.udcc.org/

[9] 松本裕治, 北内　啓, 山下達雄, 平野善隆, 松田寛, 高岡一馬, 浅原正幸 “形態素解析システム『茶筌』version 2.2.9 使用説明書”(2002).

[10] http://chasen.org/~taku/software/TinySVM/

表 1　カテゴリ毎の適合率

	名詞	用語
スポーツ	0.860	0.821
刑法	0.743	0.636
政府	0.611	0.490
教育	0.864	0.868
交通	0.739	0.700
軍事	0.608	0.652
国際関係	0.608	0.622
言語活動	0.764	0.596
演劇	0.976	0.928
農業	0.909	0.908
平均	0.768	0.722

図 1　カテゴリ毎の適合率グラフ

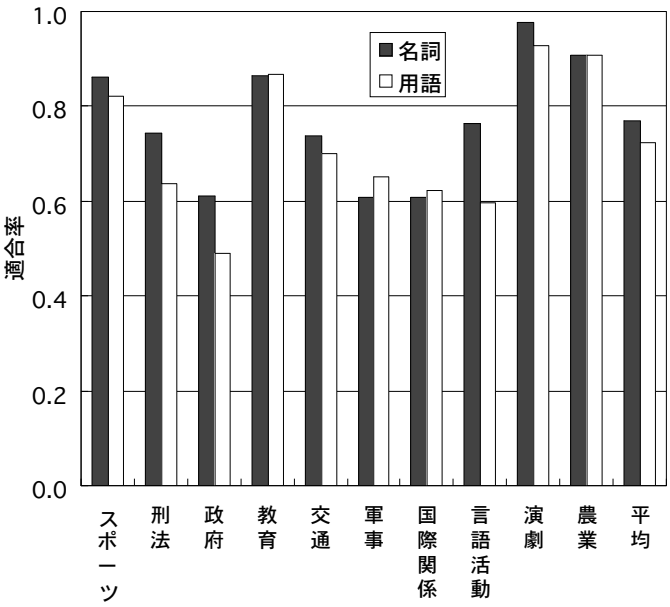

表 2　カテゴリ毎の再現率

	名詞	用語
スポーツ	0.740	0.550
刑法	0.550	0.350
政府	0.330	0.240
教育	0.570	0.460
交通	0.510	0.420
軍事	0.450	0.450
国際関係	0.480	0.460
言語活動	0.420	0.310
演劇	0.800	0.640
農業	0.700	0.590
平均	0.555	0.447

図 2　カテゴリ毎の再現率グラフ