相対頻度を利用した不要語除去によるテキスト自動分類

Issei Yoshida
IBM Japan, Yamato Software Laboratory

本論文では、テキスト自動分類システムにおいて、カテゴリーの特徴付けに役立たない語（不要語）の新しい定義を提案する。一般的に、特徴づけに役立たない語は除去するかまたはその重みを低くすることによって分類性能を上げることができるが、除去する場合はstop listを参照したり、TFの上限・下限を設定する方法をとることが多い。本論文では、除去に際してカテゴリー間の出現頻度の違いに注目した手法を述べる。また、その手法を用いたベクトル空間モデルによるシステムを用いて、この手法の有用性を確認した。

1. はじめに

テキスト自動分類システムは、学習機能と分類機能の２つの要素から構成される。これらの機能を実現するために、決定木、Neural network、ベクトル空間モデルなど様々なモデルが提案されている（[1][2]）。いずれの方法においても、各カテゴリーや文書を特徴付ける語を文書から抽出することが重要である。しかし、文書から単語を頻度順に取り出すと、カテゴリーを一意に決定するために有用でない語（不要語）が上位を占めてしまう。この不要語を学習・分類前に除去しておくことが、分類性能の大きな改善をもたらす。

不要語には大きく分けて、機能語と一般語の2種類が存在する。機能語は語と語の関係を表す助詞・助動詞などを指す。機能語はカテゴリーに依存しないものが多いので、語の品詞を調べたり、あらかじめ不要語リストを作成しておくことにより除去出来る。一般語は、機能語以外の一般的に頻繁に用いられる語を指す。一般語は機能語と異なり、語の頻度によって決定されることが多く、与えられた文書集合中の出現頻度が或る上限または下限を超えた語を不要語とする手法が一般的である（[1]）。この上限・下限を決める手法としてZipfの法則などが知られているが、上限・下限を超えない語の中にも不要語が含まれていたり、逆に、超えた語の中にカテゴリーを特徴付ける重要な語が含まれている場合がある。

そこで本論文では語の偏在性の指標として他の全てのカテゴリーでの出現頻度が閾値以下であることを条件に用いる方法を提案する。また、山本らが提案している手法[3]と本手法の不要語除去効果をコーパスを用いて比較したところ、閾値を適切に設定することにより、recall を83.4% から86.0%に、precision を83.6% から86.7%にそれぞれ向上させることが出来た。

2 .相対不要語の定義と不要語除去のアルゴリズム

2-1. 準備

本論文で述べるテキスト自動分類の前提となる諸条件についての定義を行なう。

定義2-1

Cat = {C(1),C(2),...,C(N)}：与えられたカテゴリー集合
W(C) (C∈Cat)：カテゴリーC内の、不要語除去前の語集合
Weight(w,C) (C∈Cat、w∈W(C))：
　Cに属する学習用文書のうち語wを含む文書の個数
Tw(C):= Σ Weight(w,C) (C ∈Cat)：
　和はすべての語w∈W(C)　に対してとる

Weight(w,C)の定義はこの他にも、常に非負の値をとるという条件の下で、値の大小が語wのC内の出現頻度を反映しているような別の指標も採用出来る。

2-2. 学習時における不要語除去

本論文で提案する不要語の定義は以下で与えられる。

定義2-2（相対不要語）

閾値 R (0≦R≦1) に対し、
w ∈ W(C(i)) がC(i)の相対不要語である
⇔ w ∈ W(C(k)) かつ Weight(w,C(k))＞R×Tw(C(k))
を充たすk (k≠i)が存在する

各カテゴリーC(i)に対し、相対不要語をW(C(i))からすべて取り除いた残りの語集合を、改めてカテゴリーC(i)に属する語集合として定義する。この定義から、Rが0に近づくほど不要語の数が増える。R=1のときは、W(C(i))内に不要語は存在しない。また、R=0のときは他のカテゴリーに存在する語は、その重みに依らずすべて不要語となる。

2-3. 分類時における不要語除去

2-2節の相対不要語除去を含む学習を経たシステムが、或る文書DをC(1),...,C(N)のいずれかに分類させる（または、どこにも分類させない）際に、Dの不要語が分類のためのデータとして採用されないようにしたい。このためには、Dに含まれる語のうち、W(1),...,W(N)のいずれかに含まれるもののみを評価の対象とすればよい。各W(i)からは既に相対不要語が取り除かれているので、Dから相対不要語を取り除く効果が得られる。

3.不要語除去効果の評価

3-1. 実装システムおよび評価作業の概要

本論文で紹介した手法を実装したプログラムを用いて、(1)閾値Rの変化に伴う不要語除去効果の変化の観察(2)閾値Rの変化に伴う分類性能の変化の測定および山本らの手法[3]との性能比較を、Recall/Precision ベースで行なった。評価方法およびプログラムの概要を表1に示す。

表1. 評価方法およびプログラム仕様の概要

カテゴリー集合の要素	政治、経済、スポーツ、社会、国際の5個
学習用文書集合	各カテゴリーに対し100文書、計500文書
テスト用文書集合	各カテゴリーに対し100文書、計500文書
語の定義	形態素解析にはIBMのJMAを使用し、JMAによって単語に分割された中から、品詞コードが動詞の一部と名詞に対応するもののみを語と定義した。
分類システムのモデル	ベクトル空間モデル
語wがW(C)の要素となる必要十分条件	wがCの学習用100文書に2個以上出現する（同一文書内の複数出現もカウントする）
関数Weight(w, C)	Cの学習用100文書のうちwが存在する文書の個数

カテゴリーベクトルの定義	すべてのW(C)の合併を語空間とし、不要語の重みを0、それ以外の語wの重みを、wのC内での延べ出現回数とする
ベクトルの距離	２つのベクトルのなす角の余弦(cosine)
分類基準	カテゴリーベクトルと文書ベクトルの近さが最大となるCに分類。また、その値が0以下ならばどこにも分類しない

3-2. 閾値Rの変化に伴う不要語除去効果の変化

表2は、不要語除去操作後の、各カテゴリーの頻度順上位20語に占める不要語の個数を示す。不要語かどうかは目視により判断した。政治の例を表3に示す。太字の語が、目視により不要語と判断された語である。

表2：頻度順の上位20語に占める不要語の個数

R	政治	経済	スポーツ	社会	国際
1(除去なし)	16	18	11	20	15
0.09	6	7	4	7	6
0.01	0	3	0	3	3
0	0	0	0	1	1

表3:カテゴリー「政治」の不要語除去状況（品詞コードを省略）

R	カテゴリー「政治」内の、頻度順の上位20語リスト
1	**い、し**、小泉、首相、**ついて、れ、な、あ、述べ**、自民党、**考え、示、行、い、日本**、政府、**られ、問題、対、する**
0.09	小泉、自民党、**考え、国民**、幹事長、**対応**、民主党、外務省、公明党、橋本、改革、幹部、内閣、機密費、**必要、事件**、外相、参院選、自民党総裁選、**日夜**
0.01	自民党、幹事長、公明党、橋本、内閣、機密費、参院選、自民党総裁選、党首、憲法、首相公選制、憲法改正、山崎、記者団、橋本、衆院、福田、保守党、公明、官房長官
0	公明党、参院選、憲法、首相公選制、山崎、橋本、衆院、福田、保守党、官房長官、両党、派閥、総裁選、神崎、共産党、行政改革、国外退去、参院、亀井、政調会長

或るRの値に対して除去されている語は、より小さなRの値に対しても、相対不要語の定義によって除去されている。R=1のときの不要語上位16個は、R=0.01の段階ですべて除去されている。一方で、R=0.01またはR=0の場合に、政治を特徴付ける語（重要語）が除去されずに残っていることも表3からわかる。

3-3. 閾値Rの変化に伴う分類性能の変化

図1は、Rの値を0から0.09まで0.01刻みで変化させたときの、分類結果のRecall、Precisionの値の変化を示している。（参考として、R=0.1, R=0.2のときの値も載せてある）

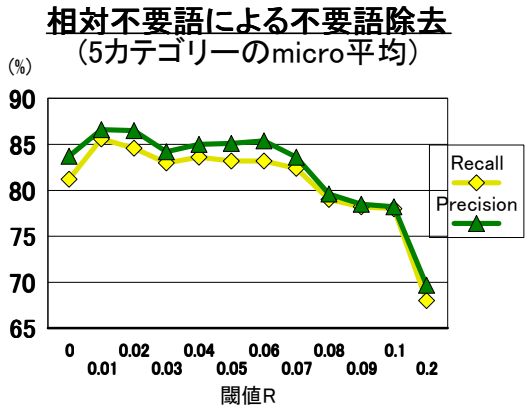

図1:相対不要語による分類性能

R=0.01以上については、Rの値が小さくなるほどRecall、Precision共に向上する傾向が見られる。一方、R=0のときの値が低いことは、語を除去し過ぎていることを示唆している。特に、学習用文書集合がより多くの文書を含む場合、或るカテゴリーが偶然他のカテゴリーの特徴語を含む可能性が高くなり、R=0のときの分類性能はさらに低下すると考えられる。

山本らの手法[3]で用いられている手法(Process2-1)は、「或るカテゴリーの語wを除去するための必要十分条件を、wが他のすべてのカテゴリーに出現する場合」としている。この手法を、他の条件を3-1の表1と同じにして評価した結果は、Recall = 78.2, Precision = 79.0 であった。相対不要語を利用すると、R が 0.01 から 0.07 の範囲で、Recall, Precisionともにこの数値を上回っており、相対不要語の効果が確認された。 なお、山本らの手法によって除去される語は、R < 1.0 の相対不要語による除去によって必ず取り除かれることに注意されたい。

また、山本らの手法は、不要語除去が重み付けの定義と結びつけて適用されており、上の結果は必ずしも公平とは言えない。そこで、3-1の表1の設定のうち、ベクトルの成分の定義を山本らの手法で用いられている重み付け(Process2)に変更し、再度測定を行なった。重み付けの定義は、

「w ∈ C に対し、w の C 内での延べ出現回数を e, w を含む カテゴリー(Cを含む)の個数を m として、e (N - m) 」

である。

上と同様にRを変化させて分類性能を測定したところ、重み付けを変えたこの測定では図1ほどはっきりと閾値と性能の相関関係は見られず、R = 0.05 で最高値 Recall = 86.0, Precision = 86.7 を記録した。一方、山本らの手法(Process2)による結果は、Recall = 83.4, Precision = 83.6 であった※。相対不要語による除去の場合、R が 0.01 から 0.1 までの範囲で、Recall, Precisionともにこの数値を上回っている。

（※ベクトルの距離の測定にcosineを用いているため、論文[3]のProcess3（ベクトルの正規化）も同時に実行されている）

4.結論

本論文の提案する不要語除去の仕組みが、従来の手法と比較して有効である可能性が示された。今後の課題は、(1) 実用的なRの値の範囲が、適用される分野集合によってどの程度変化するかを調べること (2) 今回のコーパスでは分野間の距離が比較的大きいと考えられるため、分野間の距離が小さいと考えられるようなカテゴリー集合（例えば「野球」「ソフトボール」「サッカー」・・・）に対しても有用であるかを調べることである。

参考文献

[1] 徳永健伸, 言語と計算5 情報検索と言語処理, 東京大学出版会
[2] Y.Yang, An Evaluation of Statistical Approach to Text Categorization, *Tech.Rep. No. CMU-CS-97-127*
[3] K.Yamamoto,S.Masuyama,S.Naito, Automatic Text Classification Method with Simple Class-Weighting Approach, *Natural Language Processing Pacific Rim Symposium '95*