新聞記事からの特徴単語抽出方式とその評価*

1Q－2

木村　誠†　絹川 博之†
東京電機大学大学院　工学研究科

1　はじめに

近年のコンピュータの発展・普及とともに、電子化された文書数は急激な増加の一途を辿ってきた。これに伴い、膨大なテキストデータから必要な文書を見つけ出す際、多くの文書に目を通さなければならなくなっており、性能の良い検索システム・自動要約システムへのニーズが高まっている。

本研究では、複数文書の自動要約システムの開発を目的としている。タスクのサブセットとして、1文書からその主意を表す単語を抽出を行う。抽出手法として、特徴単語抽出方式を適用することとし、最適な方式を選択するために各種の抽出方式の有効性を評価実験にて比較する。

2　特徴単語抽出方式

本研究では、以下の5つの特徴単語抽出方式に関し、実験評価により精度と再現率を算出し、最も性能の良いものを選択する。本タスクでは、これらの方式により抽出された単語のうち上位の数個を要約文生成のための代表単語として抽出するため、精度を重視した選択を行う。

(1)tf：もっとも単純な方式であり、1文書内における各単語の出現頻度のみを利用する。

(2)tf-idf：Salton らにより提案された方法。より少ない文書に偏って出現する単語に高スコアを与える。

(3)クラスタ内 tf-idf (以下 Ctf-idf)：文書が属するクラスタ自体を全文書集合とした tf-idf である。

(4)tf/TF：クラスタ内における語の出現確率 tf と全文書における出現確率 TF を比較したものである。

(5)HGS：久光らによって提案された方法[1]で、超幾何分布を応用した確率計算に基づく方式であり、高頻度語や低頻度語に偏らない公正な重み付けが高速に行えるとされる。キーワード w を含む文書集合を D(w)、全文書の単語数を N、単語 v の全文書中での頻度を K、D(w)の単語数を n、v の D(w)中での頻度を k としたとき、(式 1)で定義され、D(w)を特徴付ける単語に高スコアを与えるように働く。

$$\left.\begin{aligned} HGS(N,K,n,k) &= -\log\left(\sum_{l\geq k} hg(N,K,n,k)\right) \\ hg(N,K,n,k) &= \frac{C(K,l)C(N-K,n-l)}{C(N,n)} \end{aligned}\right\} \cdots(\text{式 1})$$

3　評価実験

3.1　実験方法

各特徴単語抽出方式の性能を比較調査するために、評価実験を行った。実験では、日経新聞 98 年度版電子記事(有効記事数 200879)を全文書集合とし、特定のテーマを表現すると思われるキーワードにより検索を行い、12 種類のクラスタ(述べ記事数 2260)を生成した。各クラスタに含まれる全文書について、その本文に含まれる全単語に各特徴単語抽出方式を適用した。実験では、

(1)見出しに含まれる単語（本文に含まれないものを除く）を正解リストとし、各文書ごとに精度と再現率の相関データを生成した。

(2)各方式ごとに全クラスタの平均精度を算出した。

平均精度は(A)特徴度最高値の1語 (B)再現率 0～0.1 (C)再現率 0～1.0 の3領域について算出する。

なお、本実験は方式間の順位を決定付けることが目的であるため、見出し及び本文に含まれる単語数による精度の正規化は行っておらず、絶対値的な精度を得るものではない。

形態素解析には茶筌(ChaSen) version 2.1 for Windows[2]を用い、品詞区分として名詞及び未知語以外の語はストップワードとした。

---

*Evaluation of Feauture words Selection Methods

†Makoto KIMURA, Hiroshi KINUKAWA
Graduate School of Engineering, Tokyo Denki University

3.2 実験結果

(1)精度と再現率

2.の特徴単語抽出方式のうち最も性能の良かったHGS及びtf-idfに関し、対比的な2クラスタの再現率-精度図を示す。図1は「スケート」、図2は「スポーツ」をキーワードとして生成したクラスタにおける結果である。図中の曲線は、最小二乗法による近似曲線である。

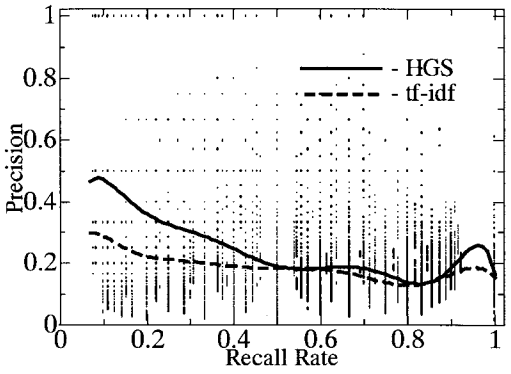

図1　再現率-精度(スケート)

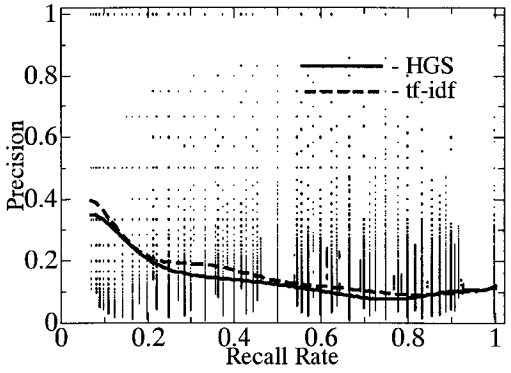

図2　再現率-精度(スポーツ)

(2) 全クラスタにおける平均精度

表1に、全クラスタにおける各方式の平均精度を示す。

表1　全クラスタにおける各方式の平均精度

領域	HGS	tf-idf	Ctf-idf	tf	tf/TF
(A)	0.020	0.013	0.010	0.007	0.001
(B)	0.019	0.012	0.011	0.008	0.005
(C)	0.025	0.024	0.021	0.023	0.010

4　考察

・表1より、今回調査した12クラスタにおいては5方式のうちHGSの性能が最も良好であり、tf-idfが次点となった。他の方式と比べHGSは特にスコア最高語(領域(A))やその近辺の語(領域(B))における精度が高く、本タスクのような、低再現率領域における精度を重視する使い方には適しているといえる。また、大半のクラスタ中では再現率の全域に渡って他の方式よりも優れた性能を示した。

・一部のクラスタについては、HGSとtf-idfとがほぼ同等の性能となった。図1及び図2は、HGSの性能差の状況を示している。図1では、概念体系の下位に位置するような、テーマ特定性の強い語をキーワードとしているのに対し、図2ではそれよりも上位に位置するものをキーワードとしてクラスタを生成している。今回試行した他のキーワード対でもほぼ同様であり、クラスタのテーマ性が強いほどHGSの性能が大きく向上する傾向が見られた。また図2のように、テーマ性の低いクラスタにおいては、HGSはtf-idfと同程度の性能を示した。

・以上より、本タスクにおける特徴単語抽出方式としてはHGSが有効であるといえる。ただし、本実験において、HGS計算パラメータのうちn及びKを動的に算出するために計算機の主記憶を500MB前後使用しており、一般的目的のシステムでの利用に際しては、少量の主記憶で高速に動作可能な計算エンジン[3]を利用するなど、実装上の効率化を図る必要があるといえる。

5　おわりに

文書自動要約のための特徴単語抽出方式として、tf,tf-idf,Ctf-idf,tf/TF,HGSのうち、最適なものを選択するために評価実験を行った。実験の結果、HGSが最も良好な性能を示した。今後、さらに多くの方式間での比較を行い、文書要約システムへ実装する予定である。

参考文献

[1]久光 徹,丹羽 芳樹:「組み合わせ的確率モデルに基づく特徴単語選択方法-超幾何分布の応用-」,情報処理学会・自然言語処理研究会,140-12(2000).
[2]「日本語形態素解析システム ChaSen」,
http://chasen.aist-nara.ac.jp/
[3]高野 明彦,他6,「汎用連想計算エンジンの開発と大規模文書分析への応用」,第19回IPA技術発表会(2000)
[4]徳永 健伸,「情報検索と言語処理」,東京大学出版会(1999)