単語の頻度情報を応用した文の評価

7G－6

内田友幸　田中英彦　東京大学　工学部
tomo@mtl.t.u-tokyo.ac.jp tanaka@mtl.t.u-tokyo.ac.jp

1　はじめに

近年、社会構造の高度情報化が進み、電子化された自然言語文書が大量に流通するようになって、内容の要約などの自然言語文書をより有効に扱える手法の確立が望まれるようになってきた。しかし、自然言語文書の内容の理解が必要な処理は、電子計算機の飛躍的発展にも関わらず、いまだ困難である。そこで、本研究ではニューラルネットを利用して、従来とは異なった角度から自然言語処理について考察を加えている [1]。本稿ではその応用の一つとして、記事全体におけるセンテンスの重要度の評価を行い、心理実験の結果と比較して有意な結果が得られたことを報告する。

2　システムの構成

2.1　システムの概略

記事中のどのセンテンスがより主旨を表現しているかということは、その文章をとらえる角度によっても変わってしまうので、厳密に解析する事は困難である。そこで本研究ではキーワードを利用した単語レベルの処理からこれらの問題に着目した。

似たような内容の文書同士をクラスタリングし、その文書間の共通した特徴をキーワードの組として抽出すれば、その特徴にそれらの文書の主旨が集まりやすいのではないかと考えた。そして、その特徴に一番近いセンテンスがより主旨を表現していると仮定し、解析を行なった。

具体的には、本システムではまず、大量の自然言語文書を単語に分解し、その単語の頻度情報を元に適応共鳴理論 ART (Adaptive ResonanceTheory)[2] を利用して文書のクラスタリングを行なう。この際、その文書の特徴が重みつきキーワードの組として抽出されることはすでに示した [3]。そこで、本稿ではこのキーワードの組をそのカテゴリの特徴と捉え、この特徴を利用してそのカテゴリ内の記事に含まれるセンテンスの重要性を評価した。

2.2　クラスタリング方法

今回入力したデータは、約2週間分にあたる内政、事件、経済、外国情勢、外国経済などの広範なジャンルの1440 個の新聞記事である。各記事をそれぞれセンテンス毎に分け、単語に分解する。そして、単語の全種類に番号を割り当て、一つの文書内の単語の頻度情報を頻度ベクトルに置き換える。M 種類の単語が存在した場合は、$I = (I_1, I_2, \cdots, I_M)$ によって表される M 次元ベクトルにする。ただし、各要素の内容は以下のようにする。

$$I_j = n \tag{1}$$

(ここで n は文書中に No.jの単語が存在する個数)

これを長さが 1 になるように正規化して頻度ベクトルとし、ART への入力データとする。

ART ではこれらのベクトルを自己組織的にクラスタリングし、最終的にカテゴリの組とその重みつきキーワードの組を得る。

2.3　評価手法

得られた重みつきキーワードの組を利用して記事内のセンテンスの評価を行なう。評価は 2 種類の方法で行なった。一つは、キーワードの組は頻度ベクトルになっているので、各センテンスも頻度ベクトルにして、この 2 つのベクトルの距離を測る方法である。二つ目は、キーワードの重み値を単純に合計する方法である。

キーワードの頻度ベクトル$\boldsymbol{X}$と対象とするセンテンスの頻度ベクトル b との距離 d は以下のように定める

$$d = \|\boldsymbol{X} - b\|^2 \tag{2}$$

また、キーワードの重み値の合計量 s は以下のように定めた。

$$s = \sum_{k=1}^{M} s_k x_k \tag{3}$$

(s_kは No.kの単語が存在する個数、x_kは$\boldsymbol{X}$の第 k成分)

3　実験結果

3.1　タイトルの評価

まず、基本的な能力を把握するため基礎実験を行なった。新聞記事の主旨はその記事のタイトルに書かれているので、本システムでこのタイトルの評価と内容文の評価を行なうことでこのシステムの能力を調べた。全記事のタイトル、センテンスをそれぞれ評価した結果を以下に示す。

記事のタイトルと内容文の評価値

	タイトル	内容文
距離 (d)	1.47±0.02	1.77±0.01
キーワード量 (s)	0.107	0.171

距離 (d) の方はタイトルの方が小さくなっていて、内容文よりタイトルの方がカテゴリの頻度ベクトルに近いということがわかる。キーワード量 (s) の方はタイト

A Study of Sentence Estimation
by Word Frequency Analysis.
Tomoyuki Uchida Hidehiko Tanaka Department of Electrical Engineering, Faculty of Engineering, University of Tokyo, 7-3-1 Hongo, Bunkyo-Ku, Tokyo, 113, Japan

ルの方が小さくなってしまっているが、タイトルより内容文の方が1センテンスの平均文字数は2.94倍大きいことを考慮すれば、逆に1.83倍タイトルの方が大きくなる。

距離(d)の結果を信頼すれば、タイトルの方が内容文よりもカテゴリの特徴に近い。すなわちカテゴリの特徴とその記事の主旨に相関があることがわかる。

3.2 心理実験との比較

次に心理実験を行ない、本システムの評価との比較を行なった。

どのセンテンスが一番大切であるかということは主観的なものであり、決定することは困難である。そこで心理実験を行ない、多数を占めた回答を正解と設定することで評価の指標とした。

実験方法はまず、1440個の新聞記事の中から、4つ以上の記事が集まっているカテゴリにあって、短か過ぎず、各センテンスの長さが極端に偏っていないもの14個を、記事内容とは無関係に抽出した。次に、成人の被験者18名に対し「その記事の内容を良く表していて、内容を把握する上で大切な文を大切な順に選んで下さい」という設問とともに配り、各センテンスに振った番号で回答してもらった。

14個の記事のそれぞれのセンテンス数は平均6.1個。これらから一つずつ選んでもらう訳であるが、それぞれの最大数を得た選択肢は平均して12人が選択していた。

この最大数を得た選択肢を正解として、本システムの評価と比較を行なった結果を以下に示す。「3位まで」とは本システムで評価の高い順、3位までのセンテンスのうちに、正答があった記事の数を示している。

システムの正答数

	1位	2位まで	3位まで
距離(d)	6	11	13
キーワード量(s)	7	12	13

この結果から、システムは人間が選んだ答をそれぞれ43 % (6/14)、50 % (7/14) 一致させたことがわかる。また、3位まで許せば93 % (13/14) 一致させることができた。

次に、被験者の回答した選択肢の分布を、システムによる評価値を横軸としたヒストグラムにして示す。

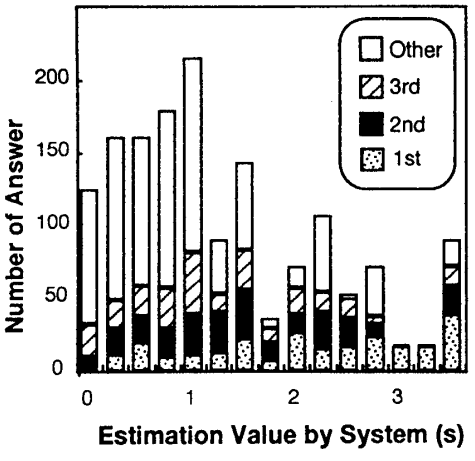

図 1: 被験者からの回答の分布

システムの評価値はキーワードの重み値の方を使っているため、右へ行くほどシステムの評価の高いセンテンスになっている。また、縦軸は回答数であり、下から1番目に選んだ回答、2番目、3番目として選ばれた回答数になっている。そして、棒の高さ全体は選択肢全体の分布になっている。

このグラフから、右へ行くほど回答として選ばれる割合が増加している様子がわかり、かつ、その傾向は回答者が選んだ順位が3、2、1と上がるほど強くなっている。これを平均値で表すと、以下のような表になる。

回答された選択肢のシステムによる評価値の平均

	選択肢全体	1位	2位	3位
距離(d)	1.80±0.04	1.72	1.79	1.80
キーワード量(s)	1.42±0.21	2.26	1.67	1.41

回答の平均値の95 %信頼区間は距離(d)で0.02〜0.03、キーワード量(s)で0.12〜0.13となっており、距離(d)での3位のデータ以外は重複していない。この結果から回答順位が2位より1位の方がシステムの評価が上がり、なおかつ、1、2位の回答は選択肢全体の評価に比べて全体的に高くなっていることがわかる。

4 考察

文書をカテゴリ分類し、その共通の特徴を基準として文書内のセンテンスを評価することで、タイトルの評価が内容文より高くなり、また心理実験の結果に近い結果を得ることができた。

これは大量の文書の中から似たような文書をクラスタリングすることで、それらの文書の特徴がクローズアップされ、その特徴により近いセンテンスがその文書の中で、文書の意図する内容を表現している確率が高いためではないかと考えられる。

このことから、単語の頻度情報などの自然言語の表層的な情報を統計的に捉える試みは、自然言語文書を人間が有効に利用しやすくするという技術を開発する上では、ある程度の効果が期待できるといえる。

5 おわりに

本稿では、ARTによる文書のクラスタリングを応用することで、自然言語文書中「内容を良く表しているセンテンス」を評価できることを示した。これを応用すれば、論文データベースや新聞データベースなどのより効率的な閲覧などができると考えられる。

しかし、カテゴリ分類が適切に行なわれなかった文書などは今回取り扱わなかったし、単語の頻度情報しか利用していないため、短い文書や複雑な概念を持った文書への応用は難しいことが予想される。今後は単語の頻度だけでなく構文情報などを取り扱ったり、辞書を利用することでより可能性を広げていく予定である。

なお、心理実験に御協力いただいたNIFTY-Serve、FGALRAYの会員の皆様に感謝いたします。

参考文献

[1] 内田 友幸 田中 英彦:“ARTを利用した多義語の分類とその評価” 情報処理学会, 研究会報告, 自然言語処理, 101-15, 1994.

[2] Stephen Grossberg et al. : “The ART of Adaptive Pattern Recognition by a Self-Organizing Neural Network,” IEEE Computer, Vol.21, No.3, (March 1988), pp.77-88

[3] 内田 友幸 田中 英彦:“ARTを用いた自然言語中の単語の頻度の情報処理” 情処第47回全国大会, 2-215, 1993.