概念ベースを用いた新聞記事からの主テーマ抽出手法

高橋良和　渡部広一　河岡司

同志社大学工学部

１．はじめに

近年, パソコンや携帯電話などのコンピュータの発達により，人間が取得できる情報は，非常に多量なものとなってきている. それらの情報にすべて目を通すだけでも非常に多大な時間が必要になっている. その情報を短時間で理解するためには, コンピュータが情報の最も重要な部分を選んで，人間にその情報を与える必要がある．そこで，人間が長文をコンピュータに与えて, コンピュータが最も重要な語を出力することができれば, 短時間で長文の内容を理解することができる.

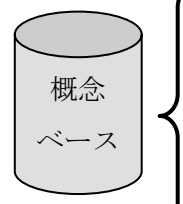
本研究では，新聞記事を対象にして，概念ベース[1]と単語の種類（品詞など）による重要度の設定を用いて新聞記事から最も重要な語（テーマ）を抽出する手法を提案する.

２．「概念ベース」と関連度計算

概念ベースとは，電子辞書などから自動構築された知識ベースである. ある一つの概念 A を属性 a_i と重み w_i によって，次のように定義する.

$$A=\{(a_1,w_1),\cdots,(a_N,w_N)\} \qquad (1)$$

つまり，概念 A は概念 A の意味特徴を表す属性とその属性の重要度（重み）の対の集合で表される．概念数は，約 9 万語で一つの概念につき平均 30 個の属性が存在する. (図 1).



概念：約 9 万語

概念	属性／重み
医者	医師／0.3, …, 治療／0.1
太陽	日光／0.2, …, 月／0.05
美術	芸術／0.4, …, 鑑賞／0.07
：	：

図 1　概念ベース

関連度とは，二つの概念 A と B の関連の強さを定量化した相対的な値である．関連度は，0 から 1 までの連続値をとり，関連の強い概念同士では高い値となり，関連の弱い概念同士では低い値となる．例えば，概念「医者」と「病院」の関連度は 0.72，概念「医者」と「太陽」の関連度は 0.04 となる. このように概念同士の関連の強さを定量化すれば, 曖昧である概念間の関連性の強弱を数値の大小関係を比較することでコンピュータに判断させることができるようになる.

語を表記的記号としてではなく, 意味属性を持つ概念としてとらえることで, コンピュータが言葉の意味を理解することができる. 人間とコンピュータとの双方向の会話の実現には欠かせない技術である.

本研究では重み付き関連度計算[2]を利用している.

２．１ 重み付き一致度

２つの概念A，Bをその一次属性をa_i，b_j，重みをu_i，v_jとし，概念Aと概念Bを

$$\begin{aligned} A&=\{(a_i,u_i)\mid i=1\sim L\} \\ B&=\{(b_j,v_j)\mid j=1\sim M\} \end{aligned} \qquad (2)$$

と表現することにする．概念 A，B の重み付き一致度は以下のようになる.

$$W(A,B)=(s_A/n_A+s_B/n_B)/2 \qquad (3)$$

$$\begin{aligned} s_A&=\sum_{a_i=b_j}u_i \quad & n_A&=\sum_{i=1}^{L}u_i \\ s_B&=\sum_{a_i=b_j}v_j \quad & n_B&=\sum_{j=1}^{M}v_j \end{aligned} \qquad (4)$$

重み付き一致度 *MatchW* は概念 A から見たとき，概念 B の属性と一致した属性の重みの割合と，概念 B から見たときの概念 A の属性と一致した属性の重みの割合の平均を表している.

２．２ 重み付き関連度

概念 A，B のうち属性数の少ない概念を $A(L\leqq M)$とし，概念 A の一次属性の並びを固定する.

$A=\{(a_1,u_1),\ (a_2,u_2),\ \cdots,\ (a_L,u_L)\}$

概念 *B* の各一次属性を対応する概念 *A* の各一次属性との一致度(*MatchW*)の合計が最大になるように並べ替える．

$$B_x=\{(b_{x1},v_{x1}),\ (b_{x2},v_{x2})\ ,\cdots,\ (b_{xL},v_{xL})\}$$

概念 *A* と概念 *B* との関連度 *ChainW*(*A*,*B*)は，

$$ChainW(A,B)=\left(s_A/n_A+s_B/n_B\right)/2 \qquad (5)$$

$$s_A=\sum_{i=1}^{L}u_i MatchW(a_i,b_{xi}) \qquad n_A=\sum_{i=1}^{L}u_i$$
$$s_B=\sum_{i=1}^{L}v_i MatchW(a_i,b_{xi}) \qquad n_B=\sum_{j=1}^{M}v_j \qquad (6)$$

である．すなわち，重み付き関連度は，概念 *A* から見たときの一致している属性の重みの割合 s_A/n_A と概念 *B* から見たときの一致している属性重みの割合 s_B/n_B の平均になる．

３．「テーマ候補語」の抽出手法

「テーマ候補語」を抽出するときに，まず新聞記事を形態素解析する．形態素解析には形態素解析ソフト茶筌[3]を用いる．テーマは名詞であることから，これを用いて品詞が名詞の単語のみを抽出する．しかし，名詞の中にもテーマ候補語として不適切なもの，例えば単位（一人，二人などの「人」）が存在する．さらに，「一」もテーマとしてふさわしくない単語と考えられる．このような単語（単位，数詞，地域名，助数詞，非自立（こと，ものなど），形容詞語幹，形容動詞語幹）や位置語はシソーラス[4]などを用いて削除する．

４．テーマ候補語の重要度の設定方法

この章では，テーマ候補語の重要度の設定手法について説明していく．以下の節で説明する手法で得た重要度を元にして，その重要度が最も大きくなった単語を「仮テーマ」とする．

４．１ TF と出現順序による重要度の設定

TF[5]とは，索引語頻度を意味し，索引語がどれだけ多く，文書中に出現するかを示している．何度も繰り返し使われる語は，重要であると考えられる．

この考えを利用する．まず一つの記事内のテーマ候補語の TF を計算する．これを重要度とする．

さらに，これは新聞記事独特の性質であるが，新聞記事は冒頭から順に重要な事実を書き記しておいて，編集の段階で文字数の調整が行われる場合に，どこで文書を切断してもいいような文章構成となっている[6]．そこで，見出しに存在する単語（以降「理想のテーマ」と呼ぶ）がどこに出現しているかを Yahoo!ニュース[7]１００件を対象に目視で調べた（図２）．これを利用して，出だしの一文目に出現したテーマ候補語の重要度を５倍する．この値は実験的に得た値である．

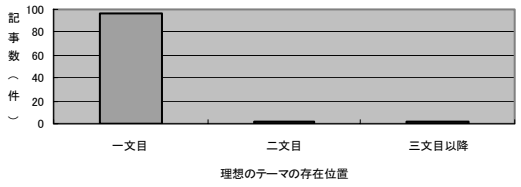


図２　記事１００件の理想のテーマの存在位置

４．２ 文法による重要度の設定

先ほどの理想のテーマを見て，その品詞や用法を調べたところ，理想のテーマには「サ変接続」(「～する」に接続して動詞の役割をする名詞)，「目的語」，「主語」が多いことがわかった．このことから，テーマ候補語がこれらの三つに該当する場合，重要度を大きくすると精度が上がると考えられる（図３）．テーマ候補語がサ変接続の場合，重要度を４倍，主語の場合は４倍，目的語の場合は２５倍する．この値は実験的に得た値である．

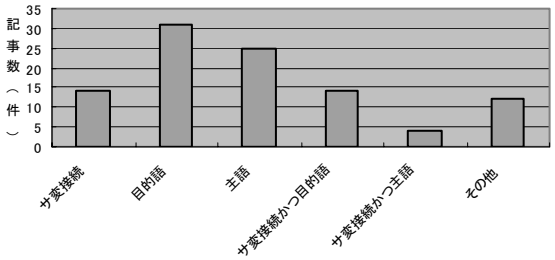


図３　「テーマ」となっている単語の文法

４．３ 概念ベースによる重要度の設定

テーマ候補語の中でも意味的に重要な語と重要でない

語が存在する．この重要度を概念ベースを用いて生成する．テーマ候補語から概念ベースを用いて連想される単語がテーマ候補語の中に存在した場合，重要度に概念ベースにより得られる属性 a_i の重み w_i を３０倍したものを足していく．

４．４ 関連度の使用

４．２節までの手法で表記的にテーマとなる語を抽出することは可能であるが，４．３節の手法では，テーマ候補語の意味があまり考慮されていない．そこで，テーマ候補語とテーマ候補語すべてを関連度計算し，その総和を求める．そして，関連度の総和を先ほど求めた重要度に掛けていく．この処理を行うことで表記的，意味的に重要な語を仮テーマとすることができる．

４．５ 未定義語の処理

概念ベースには固有名詞が登録されていないためそれらは「未定義語」として扱い，重要度の計算をしていなかった．しかし実際には未定義語が理想のテーマとなる場合は多く存在する．これより，未定義語も表記だけで得た重要度で他のテーマ候補語の重要度と比較できるようにした．この処理を行って得た結果を図４に示す．

５．テーマの出力

４章での重要度の設定手法で得られた重要度により仮テーマを得るが，理想のテーマのほとんどが複合語であるため，最後に仮テーマと新聞記事を照らし合わせ，仮テーマとなる語の前後が名詞だった場合，それらを組み合わせてテーマとして出力することにする．

６．評価

６．１ 評価用データベース

評価には，Yahoo!ニュースから無作為に抽出した記事１００件をデータベースに登録し，それを用いた．評価用記事データベースに登録した記事の平均単語数は 519.84 単語，平均テーマ候補語数は 42.8 語である．

６．２ 評価方法

評価用記事データベースを用いて，そのデータベースに保存されている記事を入力し，テーマを出力させる．記事は１００件あり，抽出したテーマが理想のテーマと一致した場合は○，明らかに違うものは×，理想のテーマとは一致しないが意味が近いものは△とする．（１００件の中の○と△の数）/１００を精度とする．

○の例…「理想のテーマ」：おれおれ詐欺
　　「抽出テーマ」：おれおれ詐欺

△の例…「理想のテーマ」：エルニーニョ
　　「抽出テーマ」：水温上昇

×の例…「理想のテーマ」：大豆飲料
　　「抽出テーマ」：甘み

６．３ 評価結果

評価は以下のように場合分けをして行った（図４）．

①：TF のみを重要度としたもの
②：①に出だしの文の重要度の設定を行ったもの
③：②に文法的重要度の設定を行ったもの
④：③に「概念ベース」を用いたもの
⑤：④に関連度計算を行ったもの
⑥：⑤に未定義語処理を行ったもの

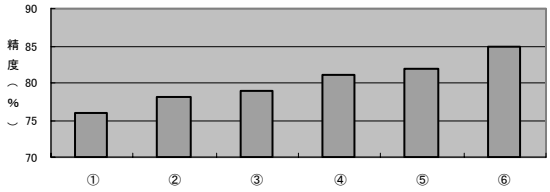


図４　評価結果

⑥の時，精度８５％という最も高い精度が得られた．そこで，⑥の時の詳細結果を示す（図５）．

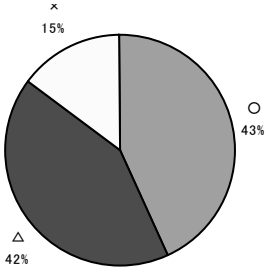


図５　⑥の時の詳細結果

７．考察

評価結果（図４）を見ると，①の状態でも精度７６％を得られた．

②の状態を見ると，精度は７８％と向上したが，①の状態では「○」だったものが「×」に変わってしまった場合が多く存在した．これは，新聞記事の出だしの一文目に出現した単語の重要度の設定を行ったため，理想のテーマが二文目，三文目以降にある新聞記事の場合，理想のテーマの重要度が出だしの一文目の単語より小さくなってしまったためと考えられる．

③の状態を見ると，精度が７９％と向上した．②の状態に比べ精度は１％しか向上しなかったが，これは②の状態で抽出していたテーマが重要度を設定したサ変接続，目的語，主語のどれかに当てはまっていたためと考えられる．

④の状態を見ると，精度は８１％で③の状態に比べ２％向上したが，これは概念ベースを用いてテーマ候補語から単語を連想することによって，その新聞記事の意味的に重要な単語を得ることができたため，精度が向上したと考えられる．しかし，概念ベースには固有名詞が登録されていないため，固有名詞が理想のテーマとなる新聞記事を対象にした場合，テーマの抽出に失敗するようになってしまった．

⑤の状態を見ると，精度は８２％となり，④の状態に比べ１％向上した．わずかしか精度が向上していない．しかし，⑥の状態と組み合わせることによって，精度は８５％となり，④の状態に比べ４％精度が向上した．これは，④の状態では抽出できなかった意味的に重要な語や固有名詞のテーマが抽出できるようになったため，精度が向上したと考えられる．

８．おわりに

提案した手法で精度８５％を得ることができた．①の状態から⑥の状態を見ると全体で９％の精度を向上させることができた．さらに，③と⑥の状態を見ると，概念ベースを用いることによって意味的にも表記的にも重要な語も出力できたと考えられる．

しかし，図５の詳細結果を見ると，「○」が４３％，「△」が４２％，「×」が１５％であり，「○」と「△」がほぼ同じ数存在する．今回の評価では（１００件の中の○と△の数）/１００としているため，「△」を正解としている．「△」を不正解にした場合，精度は４３％となり，とても精度の低いテーマ抽出手法となってしまう．よって，これからの課題として「○」の数を増やしていくことが考えられる．

これからの精度の向上手法として，IDF を考えている．IDF とは，ある索引語がどの程度その文書に特徴的に現れるのかという特定性を表す尺度である． IDF は，どの文書にも出てくる語なら小さく，特定の文書にしか出てこない語なら大きい値をとる．

今回提案した手法は TF が各テーマ候補語の重要度を大きくしているため，IDF を用いることによって精度が向上するのではないかと考えられる．

謝辞

本研究は文部科学省からの補助を受けた同志社大学の学術フロンティア研究プロジェクトにおける研究の一環として行った．

参考文献

[1] 小島一秀，渡部広一，河岡司，“連想システムのための概念ベース構成法－属性信頼度の考え方に基づく属性重みの決定”，自然言語処理，Vol.9，No.5，pp.93-110，2002

[2] 渡部広一，河岡司，“常識的判断のための概念間の関連度評価モデル”，自然言語処理，Vol.8，No.2，pp.39-54，2001

[3] 奈良先端科学技術大学院大学情報科学研究科，http://chasen.naist.jp/hiki/ChaSen/

[4] NTT コミュニケーション科学研究所監修日本語語彙体系，岩波書店，1997

[5] 徳永健伸，情報検索と言語処理，東京大学出版会，1999

[6] 山本和英，要約技術と検索技術，日本語学，Vol.23，No.2，pp60-68，2004

[7] Yahoo!ニュース，http://dailynews.yahoo.co.jp/fc/