文の結合度に基づく内容抽出手法

福本 淳一
沖電気工業(株) 研究開発本部 関西総合研究所
fukumoto@kansai.oki.co.jp

1 はじめに

情報検索において，文書中の重要部分など，検索文書の一部や内容を取り出す情報抽出に関する研究が行なわれており，これは大量の検索結果のブラウジングの支援としても有効である．情報抽出の方式としては，あらかじめ抽出パターンとして文書中のどのような内容を抽出するのかを登録しておき，そのパターンに当てはまる情報のみを抽出する方法がある[1][2]．しかし，この方法では，抽出パターンとして登録した情報のみしか取り出すことができず汎用性に欠ける．また，文書中の単語頻度などの統計量を用いて抽出する方法などもあるが[3]，抽出された文間の関連性について考慮されていないため抽出文を抄録として取り扱うには問題がある．

我々は，文書中の各文間の関連度情報を用い，関連の高い文を順に抽出することにより，抽出された文同士にも関連性を保持した抽出方法を提案する．文間の関連性を扱ったものとしてMannらはRSTとして言語的な制約に基づいたいくつかの文間の関係を設定した[4]．そして，このような文間関係を基にして得られた文章構造から抄録を作成する手法を提案されている[5][6]．また，統計的な手法によって文間の関連度としていくつかの類似度の計算方法が提案されている[7]．

一方，情報検索においてユーザの要求はさまざまであり，どのような目的で情報抽出を行なうのかによって文書に対する注目点が異なる．上で述べた抽出パターンに基づく内容抽出ではユーザの要求に合致した抽出パターンをあらかじめ登録しておかなければならず，柔軟なユーザの要求に対して対応が難しい．また，注目内容としてキーワードにマッチする文を抽出することにより，それに関する文を抽出することが可能であるが，独立に抽出された文間には互いに関連が少なく，抽出される文数も一定数にはならない．

本稿では，文書中の各文間の関連度を用いることで，ユーザによって指定された注目語句と関連する文を各文間で関連の高いもの性質を保持しながら文書中から文を抽出する方法について述べる．

2 文の結合度

各文間の結合度を計算するため，まず，日本語形態素解析システムChaSen「茶筅」[8]を用いて文書中の各文の形態素解析を行ない各文から名詞を抽出した．抽出する名詞の品詞細分類としては「普通名詞」「サ変名詞」「固有名詞」「地名」「人名」を用いた．次に，この名詞情報を用いて文書中の各文間の関連度を求める．関連度の値は，文の長さの影響を考慮し，文書中の各文の単語ベクトルの内積の値を文の単語数で正規化することにより求めた．ここで，ベクトルの各要素は文中での名詞の頻度とする．

文xの文yに対する関連度$cv(x,y)$は，それぞれの単語ベクトルv_x, v_yの内積値を文xの単語数$w(v_x)$(総数n)で正規化することによって得る．

$$cv(x,y) = \sum_{i=1}^{n} v_x(i) * v_y(i)/w(v_x) \quad (1)$$

3 重要部分の抽出

重要部分の抽出方法として指定された注目点である語句を含む文を中心にし，その文と関連性の高い文を順に選択することで文の抽出を行なう．また，選択された文とも関連度の高い文も抽出候補として選択される．注目点である語句を含む文が文書中に複数存在した場合，関連度の高さにより順に選択される．そして，あらかじめ指定され

た抽出率に相当する文数になった時点で，または，抽出済みの文と関連性のある文が存在しないと判断された時点で抽出処理を終了する．

抽出アルゴリズムを以下に示す．

1. 文書中の各文間の関連度 $cv(i,j)$ を式(1)に基づき計算する．但し，同一文の関連度 $cv(i,i)$ は0とする．
2. 指定された注目語句を含む文書中のすべての文 X をマークする．
3. マークされたすべての文 X について，$cv(X,Y)$ が最大となる文 Y のペア (X,Y) をペアリストとする．
4. ペアリストのなかで $cv(x,y)$ が最大となるペア (x,y) を1つ選び，文 x, y をこの順に抽出文とする．
5. 4. で選ばれた文 x, y について，$cv(x,y)$, $cv(y,x)$ を0にする．
6. 文 y について $cv(y,z)$ が最大となるペア (y,z) を選び，ペアリストに追加する．
7. 抽出文数が抽出率を満足する文数に達した，または，ペアリスト中の $cv(i,j)$ の最大値が0でなければ，4. に戻る．

4 評価

4.1 評価用データ

本抽出方式の評価のため，日本経済新聞記事 [9] 1994年6月1日版より会社人事関連の記事，箇条書スタイルの記事を除いたものから文数が10文以上のものを任意に50文書 (平均文数12.9文，最大17文，最小10文) 選び，抽出実験を行なった．

抽出用の注目語句としては，各記事見出し文の形態素解析を行ない，文中に含まれる名詞語句[1]を注目点を示す語句として選択した (全注目語句数299，平均注目語句数6/文書)．但し，評価者により，「期」「品」などのように注目語句としてふさわしくないと判断されたものは削除した．

また，評価用データとして，一人の評価者により注目語句と関連する内容の文を文書より任意の数取り出したものを正解例とした (全抽出文数1253文，注目語句あたりの平均文数4.2文)．

[1]文の関連度の計算で抽出した名詞の品詞細分類と同様のものを用いた．

4.2 実験方法

抽出方式の評価として以下の4つの抽出結果について正解例との比較を行なった．

(a1) 本方式で各抽出率 20%, 30%, 40%にあたる文数を抽出

(a2) 注目語句をキーワードとし，それを含む文を各抽出率 20%, 30%, 40%で抽出 (注目語句を含む文数が各抽出率に達しない場合は注目語句を含む文から順に選択)

(b1) (a1) のうち，注目語句を含む文が文書中に1文しかないものを抽出

(b2) (a2) のうち，注目語句を含む文が文書中に1文しかないものを抽出

4.3 実験結果

本方式での抽出実験結果を示す．図1に例文を，表1に例文中の各文間の関連度の計算結果を示す．

「住宅」「持ち家」「分譲」「プラス」を抽出のための注目語句とした場合の抽出結果を次に示す．抽出文の番号は，抽出された順に並べられている．語句の右の括弧内の文番号はその語句が含まれている文を示したものである．

- 注目語句 = “住宅” (1, 2, 3, 5, 11)

  11 - 1 - 5 - 3 - 2 - 8 - 13 - 6 - 7 - 12 - 4 - 9
- 注目語句 = “持ち家” (2, 3, 5, 6, 8, 11)

  11 - 1 - 5 - 8 - 2 - 3 - 13 - 6 - 7 - 12 - 4 - 9
- 注目語句 = “分譲” (2, 6, 8, 13)

  13 - 2 - 6 - 11 - 1 - 8 - 5 - 3 - 9 - 12 - 4
- 注目語句 = “プラス” (1, 11, 13)

  13 - 2 - 11 - 1 - 8 - 5 - 3 - 6 - 9 - 12 - 4

また，抽出率を 20% 30% 40% とした時の本方式の抽出結果 (文番号) を表2に示す．

各抽出実験 a1, a2, b1, b2 の抽出率 20%, 30%, 40%における抽出結果の適合率，再現率を表3に

表 1: 文 i, j の関連度

$i\backslash j$	1	2	3	4	5	6	7	8	9	10	11	12	13
1	-	.23	.17	0	.29	.5	.23	.23	.11	0	.35	0	.11
2	.28	-	.14	0	.14	.14	.21	.35	0	0	.14	0	.28
3	.33	.22	-	0	.44	.11	0	.33	.11	0	.44	0	0
4	0	0	0	-	.25	0	0	0	0	0	0	.25	0
5	.50	.20	.40	.10	-	.10	.20	.20	.20	0	.50	.20	0
6	.16	.33	.16	0	.16	-	0	.33	0	0	.50	0	.16
7	.50	.37	0	0	.25	0	-	.12	0	0	.12	0	.12
8	.36	.45	.27	0	.18	.18	.9	-	.9	0	.18	0	.27
9	.33	0	.16	0	.33	0	0	.16	-	0	.16	.16	0
10	0	0	0	0	0	0	0	0	0	-	0	0	0
11	.60	.20	.40	0	.50	.30	.10	.20	.10	0	-	.10	.20
12	0	0	0	.33	.66	0	0	0	.33	0	.33	-	.33
13	.33	.66	0	0	0	.16	.16	.50	0	0	.33	.16	-

1: 建設省が三十一日発表した建築着工統計によると、四月の新設住宅着工戸数は前年同月比一一・六%増の十三万五千八百十戸で、二カ月ぶりにプラスに転じた。

2: 持ち家（自分の土地に建てる家）が二一・〇%増の五万五千七百七十二戸、分譲住宅（マンションと建て売りの合計）が六二・二%増の三万千六百七十戸といずれも大幅に伸びた。

3: 持ち家については三月までに住宅金融公庫が受け付けた大量の融資案件が着工段階に入っている。

4: 一方、貸家は一三・二%減の四万五千七百十八戸と不振が続いている。

5: 住宅着工戸数は持ち家の伸び鈍化、貸家の不振などで三月は二十二カ月ぶりに前年同月比マイナスに落ち込んでいた。

6: これに対し、四月は持ち家・分譲の伸びが拡大、再び増加基調に乗った。

7: 季節調整済みの年率換算戸数は前月比四・七%増の約百五十八万九千戸と高水準。

8: 建設省は「公庫融資に裏打ちされた持ち家の着工増や、マンションを中心とした分譲の大幅な伸びに支えられて、夏場までは好調が続く」とみている。

9: ただ九四年度全体としては、前年度の前倒し着工の反動でマイナスになるとの見方が多い。

10: 秋口以降、転機が訪れる可能性が高い。

11: 四月の住宅着工の内訳をみると、持ち家は十一カ月連続の前年同月比プラスで、伸び率は三月の二・九%に比べて拡大。

12: 貸家は三カ月連続のマイナス。

13: 分譲は十二カ月連続のプラスで、特にマンションは九八・二%増とほぼ倍増した。

図 1: 例文 (日本経済新聞記事 94 年 6 月 1 日版)

表 2: 抽出結果

注目語句	20%	30%	40%
住宅	1, 5, 11	1, 3, 5, 11	1, 2, 3, 5, 11
持ち家	1, 5, 11	1, 5, 8, 11	1, 2, 5, 8, 11
分譲	2, 6, 13	2, 6, 11, 13	1, 2, 6, 11, 13
プラス	2, 11, 13	1, 2, 11, 13	1, 2, 8, 11, 13

示す．注目語句を含んだ文書として，実験 a1, a2 については 50 文書が，実験 b1, b2 については 46 文書が評価対象の文書であった．

$$\text{再現率} = \frac{\text{抽出文中の評価者との一致文数}}{\text{評価者による抽出文数}}$$

$$\text{適合率} = \frac{\text{抽出文中の評価者との一致文数}}{\text{本方式による抽出文数}}$$

4.4 考察

本手法 (a1) とキーワード抽出 (a2) による結果を比べた場合は，再現率，適合率ともにキーワード抽出による方法が本方式を上回ったが，注目語句を含む文が 1 文しか存在しない場合，本方式の方が良い結果を得た．このことから，文書中に注目語句に相当する単語が少ない場合，関連度を用いることで最初の文と関連のある文をうまく抽出できていることがわかる．しかし，関連するキーワードが文書中に多く存在する場合，実際にはそ

表 3: 評価結果

抽出率	再現率 (%)		適合率 (%)	
a1. 20%	44.0	(551/1253)	75.2	(551/733)
30%	54.7	(686/1253)	61.5	(686/1115)
40%	62.7	(786/1253)	54.7	(786/1438)
a2. 20%	46.6	(584/1253)	78.9	(584/740)
30%	61.4	(769/1253)	67.6	(769/1137)
40%	70.4	(882/1253)	59.0	(882/1496)
b1. 20%	15.5	(177/1145)	63.0	(177/281)
30%	18.7	(214/1145)	50.2	(214/426)
40%	21.8	(250/1145)	46.2	(250/541)
b2. 20%	13.8	(158/1145)	55.2	(158/286)
30%	18.9	(216/1145)	49.0	(216/441)
40%	22.5	(258/1145)	44.8	(258/576)

れらの文を抽出しなければならないにもかかわらず，関連度の値により別の文の抽出を行なったため良い結果が得られなかった．

本抽出方式は，文中の各語の表層と品詞といった簡素な情報のみを用いて文の関連度を求めている．しかし，文書中には言い替えなどの同一の概念を違った言葉で表わされる場合などがあり，単語間の意味的な類似性を取り入れるなど，関連度の計算方法の改良がある．また，文書中には接続語句などのように明示的に文間の関係を示しているものがあり，これも関連度の情報についても抽出の際に考慮する必要がある．

抽出結果の文を抄録としてとらえた場合，文頭の接続詞などが抽出された文の前後と正しいつながりでなくなってしまうといった問題があり，抽出結果を抄録として利用する際にはこのような語句の削除などの処理も必要である．

5 おわりに

本稿では，文章中の各文間の単語情報を用いて各文間の関連を計算し，それを基に文章の指定された注目語句と関連する部分を抽出する方法について述べた．

今後は，考察でも述べたように単語の類似度情報や接続語句などのように文間の関係を求めるための鍵となる表層表現も採り入れるなどの改良がある．

参考文献

[1] Hobbs, J. R., Appelt, D. E., Bear, J. S., Israel, D. J., and Tyson, W. M. : Fastus: A system for extracting information from natural-language text, Technical Note 519, SRI International, (1992).

[2] 松尾, 木本 : 抽出パターンの階層的照合に基づく日本語テキストからの内容抽出法, 情報処理学会論文誌, Vol. 36, No. 8, pp.1838–1844, (1995).

[3] Zechner, K. : Fast generation of abstracts from general domain text corpora by extracting relevant sentences, In *COLING'96*, pp.986–989, (1996).

[4] Mann, W. C. and Thompson, S. A. : Rhetorical structure theory: A theory of text organization, USC/Information Science Institute RR-87-190, (1987).

[5] Miike, S., Itoh, E., Ono, K., and Sumita, K. : A full-text retrieval system with a dynamic abstract generation function, In *SIGIR'94*, pp.152–161, (1994).

[6] Watanabe, H. : A method for abstracting newspaper articles by using surface clues, In *COLING'96*, pp.974–979, (1996).

[7] Salton, G. : *Automatic Text Processing*, Addison-Wesley, (1989).

[8] 日本語形態素解析システム ChaSen「茶筅」 http://cactus.aist-nara.ac.jp/lab/nlt/chasen.html (1996).

[9] 日本経済新聞社 (編) : 日本経済新聞社 CD-ROM 版 1994 年版, 日本経済新聞社, (1995).