表層表現を手がかりにした続報記事の抽出

阪元 慶隆　渡辺 靖彦　岡田 至弘
龍谷大学 理工学部 電子情報学科
ysakamot@mail433.elec.ryukoku.ac.jp, watanabe@rins.ryukoku.ac.jp

1　はじめに

新聞の関連記事を検索する研究では、従来、記事間で共起する単語を手がかりにしてそれらの記事の関連度を評価するものが多かった[新谷 96][奥 96][大竹 97]。しかし、記事間にどのような関連があるのかについては、これまであまり検討されていなかった。例えば、以下に示す記事も関連する記事である。

- 同一人物・組織がかかわる事件についての記事
- 事態の新しい推移をあつかう記事(続報記事)
- 同一ではないが、類似する事件をあつかう記事

これらの記事を区別して検索するには、例えば続報記事だけを選んで取り出すことは、記事間で共起する単語の情報だけを用いる方法ではむずかしい。そこでわれわれは共起する単語の情報だけでなく、記事の書き出し文(リード)における典型的な表現を利用して続報記事を取り出す。

本報告の構成は、次の通りである。最初に、関連記事の種類について述べる。次に、新聞記事の書き出し文(リード)中の表層表現を手がかりにして、続報記事を取り出す方法について説明する。最後に、取り出した続報記事の情報を利用するマルチメディアデータベースシステムについて述べる。

2　関連記事の種類

インターネットで配布されている朝日新聞の総合・社会面の記事を対象に、関連記事の種類について調査を行なった。その結果、以下の7種類の関連記事があった。

1. 類似する事件をあつかう記事

   異なる事件ではあるが、その内容が類似している事件をあつかう記事。例えば、以下の2つの記事は、いずれも船舶事故をあつかった関連記事である。

   「マニラ湾で遊覧船転覆、36人救助」 15日午後4時半(日本時間同5時半)ごろ、フィリピンのマニラ湾で乗員・乗客46人を乗せた小型遊覧船が転覆、沈没した。

   「東シナ海で貨物船沈没」 鹿児島県坊津町の西方約460キロの東シナ海で19日午後、パナマ船籍の貨物船アナトリ1(1,388トン、19人乗り組み)が沈没した。

2. 同一人物・組織がかかわる事件をあつかう記事

   以下の2つの記事はそれぞれ異なる事件をあつかっているが、いずれも橋本首相が関わっているという点で関連記事である。

   「CO_2削減京都会議、首相自ら調整へ」 橋本龍太郎首相は10日、日本が議長国として12月に京都で開く気候変動枠組み条約第3回締約国会議(COP3)の成功に向け、自ら調整に乗り出す方針を固めた。

   「北方領土、「相互利益」軸に」 橋本龍太郎首相は24日、都内で開かれた経済同友会の会合で講演し、対ロシア外交の新たな方針を打ち出した。

3. 発生場所が同じ事件をあつかう記事

   以下の2つの記事もそれぞれ異なる事件をあつかっているが、いずれもロシアのモスクワで発生した事件であるという点で関連記事である。

   「差別招く「民族籍」国内旅券から削除－ロシア」 【モスクワ21日＝大野正美】16歳以上の市民すべてに所持を義務づけてきたロシア独特の国内旅券制度が、旧ソ連・ブレジネフ時代の1974年以来23年ぶりに改正される。

   「アルコール密輸団、国境で御用－ロシア」 【モスクワ27日＝大野正美】ロシアとグルジアの国境にあるカフカス山脈の峠で26日、グルジア側からタンクローリー19台に満載した純粋アルコールを密輸しようとした一団が、ロシアの国境警備隊に捕まった。

4. 発生日時が同じ事件をあつかう記事

   発生日時が同じ事件をあつかう記事も関連記事である。ただし、時差などのため、同じ日の新聞に掲載されていても発生日時が異なる記事がある。例えば、次の2つの記事は同じ日の新聞に掲載されたものであるが、それぞれの事件が発生した日時は異なる。

   「武力行使に反対し、ホワイトハウスに3000人がデモ」 米政府による対イラク武力行使方針に抗議して約3000人が21日、ワシントン市内でホワイトハウスまでの約1.6キロをデモ行進した。

   「国連事務総長とイラク側の3回目の協議終了」 国連の兵器査察をめぐる危機打開のためイラクを訪れているアナン国連事務総長は22日未明(日本時間同日朝)、当地の外務省でアジズ副首相らイラク側との3回目の協議を終了した。

5. 執筆した記者が同じ記事

海外特派員の書いた記事には、以下に示すように、記者の名前が示されていることが多い。

『カンボジア市街戦激化、死傷者多数か』 【プノンペン6日＝平井正夫】カンボジアの首都プノンペンで5日から始まったラナリット第1首相派とフン・セン第2首相派の戦闘は6日、夜明けと共に再開され、重火器を交えた激しい市街戦が続いた。

『ASEAN、対カンボジア調停継続』 【クアラルンプール25日＝平井正夫】クアラルンプールで開かれた東南アジア諸国連合(ASEAN)の第30回定例外相会議は25日、カンボジア問題について、内政不干渉の原則を維持しつつもASEANとして調停作業を続けることを明記した共同声明を採択し、閉幕した。

この2つの記事はどちらも同じ記者が作成した記事である。同じ海外特派員の書いた記事には、地域・時期対象となる事件などに継続性があることが多い。

6. 続報記事

事態の新しい推移を先の記事に続けて報道する記事(続報記事)は関連記事である。例えば、以下に示す『失跡の社長、最低賃金法違反も?』は『障害者を雇用の会社社長、預金引き出し姿消す』の続報記事である。

『障害者を雇用の会社社長、預金引き出し姿消す』 従業員の半数以上が障害者である埼玉県比企郡の漆器製造会社の社長(56)が、障害者の預金約140万円を無断で引き出したうえ、障害者の保護者らからも約2000万円を借りたまま、6月初めから姿を消していることが3日、関係者の話で分かった。

『失跡の社長、最低賃金法違反も?』 従業員の半数以上が障害者である埼玉県比企郡の漆器製造会社が6月初めに事実上倒産し、社長(56)が障害者の保護者らから約2000万円を借りたまま姿を消した問題で、知的障害者全員の賃金が最低賃金を大きく下回っていたことが4日、関係者の話で分かった。

7. 解説記事

事態の新しい推移をあつかう続報記事に対し、これまでの経過をまとめた記事(解説記事)もまた関連記事である。例えば、以下に示す『苦しい－負担する側』は『医療保険改正、きょう衆院通過』であつかわれているできごと、すなわち、医療保険制度改正案についてのこれまでの経過をまとめて解説を行なう記事である。

『医療保険改正、きょう衆院通過』 患者負担増を柱とする医療保険制度改正案を審議していた衆院厚生委員会は7日、薬剤費の新たな負担を薬の種類に応じて段階的に重くすることなどを盛り込んだ与党3党の修正案を、自民、社民両党などの賛成多数で可決した。

『苦しい－負担する側』 診療報酬制度の見直しや薬価差益の解消など、医療費のむだをなくすための抜本的な改革は先送りされたまま、患者負担だけが引き上げられる。8日の衆院本会議で可決された医療保険制度改正案は、サラリーマンやお年寄りの自己負担を平均で2倍以上に引き上げる内容だ。

◆貯金を取り崩して

昨春亡くなった父と、79歳になる母が10年以上入退院を繰り返してきたという札幌市白石区の無職豊村滋子さん(56)。いま母は月額12万円の年金だけでは生活できず、貯金を取り崩しており、年金暮らしの知人の中には、お金がかかるから入院しないという人もいるという。

本研究では、これらの関連記事の中から続報記事を対象に検索を行なう方法について検討する。

3 続報記事の抽出

先の記事と続報記事は同一の事件をあつかっているので、それらの見出しとリードには共通して用いられる単語が多い。さらに、続報記事のリードにはその記事が続報記事であることを示す以下の表現が含まれることが多い。

- ～事件 / 事故 / 問題で

  (例文1) 旧ユーゴ各国を視察中のブラウン長官ら乗員・乗客計33人が乗った米空軍機が墜落した事故で

- ～について

  (例文2) 患者負担増を柱とする医療保険制度改正案について

そこで、われわれは

- 続報記事のリードにあらわれる、その記事が続報であることを示す表現
- 見出しおよびリードに含まれる、共通して用いられる単語

の2つを手がかりに、続報記事を取り出す。以下にその手順を示す。

手順1 新聞記事の見出しおよびリードを形態素解析する。
形態素解析にはJUMANを用いた[黒橋 97]。

手順2 リードに以下の表現が含まれる記事を続報記事として取り出す。

- 「事件 / 事故 / 問題」＋「で、」
- 「に」＋「ついて、」

表 1: 続報記事の抽出実験

(a) 学習サンプルを対象に行なった実験結果

閾値	適合率	再現率
40	78.9	74.8
45	84.9	72.3

(b) テストサンプルを対象に行なった実験結果

閾値	適合率	再現率
40	84.5	80.0
45	88.1	73.2

手順 3 手順 2 で取り出した記事と、その前後に報道された記事との意味的な関連度を計算する。関連度の計算方法は後述する。関連度の値がしきい値よりも大きければ、それらの記事の間には続報関係があるとして記事を関連づける。

手順 4 同じ記事と続報関係にある記事の間には続報関係があるとして記事を関連づける。例えば、記事 A と B がそれぞれ記事 C と続報関係にある場合、記事 A と B もまた続報関係にある記事として関連づける。

新聞記事の意味的な関連度は次のように計算する。続報記事として取り出された記事 x と任意の新聞記事 y の見出しおよびリードに共通する普通名詞、地名、人名の数をそれぞれ $N_c(x,y)$、$N_{pl}(x,y)$、$N_p(x,y)$ とすると、その関連度 $SCORE(x,y)$ は以下の式で求める。

$$SCORE(x,y) = w_c N_c(x,y) + w_{pl} N_{pl}(x,y) + w_p N_p(x,y)$$

ただし、w_c、w_{pl}、w_p はそれぞれ共通する名詞の種類 (普通名詞、地名、人名) の重要度を表す重みである。関連記事の種類の調査に用いた 524 個の記事 (学習サンプル) を対象に、この重みをさまざまに変化させて続報記事の抽出実験を行なった。その結果、$w_c = 3$、$w_{pl} = 5$、$w_p = 5$ とした時が比較的精度よく続報記事を抽出できた (表 1(a))。この重みの値を用いて、朝日新聞の総合および社会面の記事約 1 カ月分 481 記事 (テストサンプル) から新たに続報記事を取り出した結果を表 1(b) に示す。なお、この実験では 1 カ月間と短期間の新聞記事を対象にしたので、続報記事を検索する期間は限定しなかった。

関連度を計算する時に用いる名詞が以下のような場合は、関連のない記事とも関連づける失敗が多く、適合率を低下させた。

- 「容疑者」「起訴」「逮捕」といったどの記事にもよくあらわれる名詞
- 「橋本龍太郎首相」「東京」など、続報関係にない記事にも頻繁にあらわれる人名・地名

一方、以下のような記事では、続報関係にある記事を抽出することができないことが多く、再現率を低下させた。

- 書き出し文 (リード) が短く、記事間に共通する名詞の数が少ない記事
- 事件に関わる組織の名称や事件の呼び名が略称に変わってしまい、元の記事で用いられている単語と一致しない記事

続報記事の関連づけの処理の例を 1 つ示す。実験に用いた新聞記事の中で、「医療保険」という単語が用いられている記事が 12 個あった。このうち、国会で審議された医療保険改革関連法に関連するのは 8 記事である。こられ 8 記事のうち、『9 月から患者負担増を実施』(平成 9 年 5 月 7 日)、『医療保険改正、きょう衆院通過』(平成 9 年 5 月 8 日)、『医療保険改正案、今国会成立で』(平成 9 年 5 月 20 日)、『医療保険改正案、参院委可決』(平成 9 年 6 月 13 日)、『医療保険改正法が成立』(平成 9 年 6 月 17 日) の 5 つは続報関係にある記事、すなわち続報記事である。残りの 3 つの記事、『苦しい——負担する側』(平成 9 年 5 月 9 日)、『ひと息——利害絡む団体は』(平成 9 年 5 月 9 日)、『医療保険改正で慢性病ほど重い負担』(平成 9 年 6 月 14 日) はそれまでの経過の要約・解説を行なう記事で、事態の新しい推移を報告する続報記事ではなかった。5 つの続報記事のうち、手順 2 によって次の 2 つの記事が続報記事として取り出された。

- 『9 月から患者負担増を実施』
- 『医療保険改正案、参院委可決』

この 2 つの記事は手順 3 によってそれぞれ以下の記事と関連づけられた。

- 『9 月から患者負担増を実施』は『医療保険改正、きょう衆院通過』と『医療保険改正案、参院委可決』
- 『医療保険改正案、参院委可決』は『9 月から患者負担増を実施』『医療保険改正、きょう衆院通過』および『医療保険改正法が成立』

さらに手順 4 によって『医療保険改正案、今国会成立で』をのぞく 4 つの記事が続報記事として関連づけられた。『医療保険改正案、今国会成立で』だけが続報記事として取り出せなかったのは、そのリードで述べられている内容が他の 4 つの記事のリードとかなり異なるからである。以下に『医療保険改正案、今国会成立で』および『9 月から患者負担増を実施』のリードを示す。

『医療保険改正案、今国会成立で』 参院自民党の村上正邦幹事長、社民党の村沢牧議員会長、さきがけの堂本暁子議員団座長らは 19 日、国会内で医療保険制度改正案の扱いを協議し、今国会での成立を目指すことで一致した。

『9 月から患者負担増を実施』 患者負担増を柱とする医療保険制度改正案について、自民、社民、さきがけの与党 3 党は 6 日、薬剤費の新たな負担は薬の種類に応じて段階的に重くする、などの見直しで合意したことを受けて、衆院厚生委員会理事会に修正案を示した。

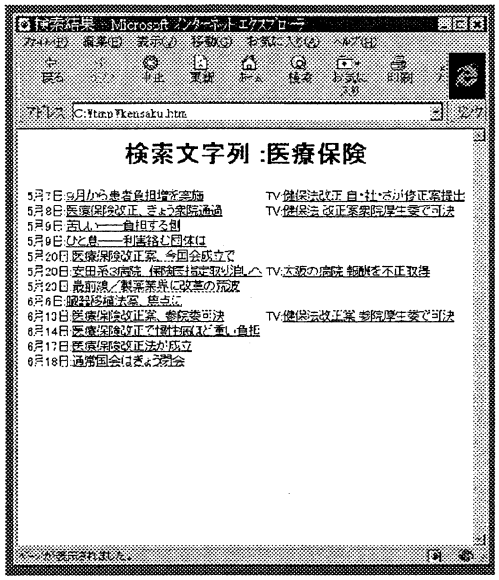

図 1: 全文検索による検索の結果の例

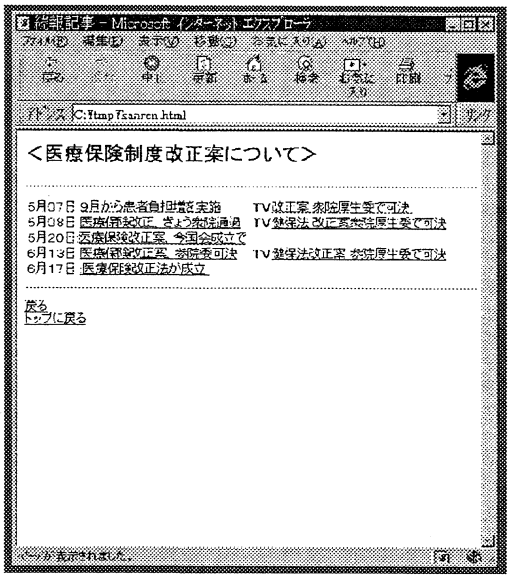

図 2: 続報関係にある記事の一覧

『医療保険改正案、今国会成立で』以外の 4 つの続報記事のリードでは、『医療保険改正案、参院委可決』のように、医療保険改革関連法の内容およびその国会での審議の 2 点が述べられているが、『医療保険改正案、今国会成立で』では法案の内容は触れられていない。このため、『医療保険改正案、今国会成立で』は医療保険改革関連法についての他の続報記事と共通して用いられている名詞が少なく、続報記事として関連づけするのに失敗した。

4 TV ニュースと新聞記事を対象にしたマルチメディアデータベースシステム

われわれは TV ニュースと新聞記事を対象にしたマルチメディアデータベースシステムを作成している [渡辺 98]。このシステムでは内容が対応関係にある TV ニュース映像と新聞記事が対応づけられていて [渡辺 97]、WWW ブラウザを利用して、それらの記事を相互参照できる。このシステムに続報記事の情報を組み込み、内容が続報関係にある TV ニュースおよび新聞記事の検索を実現した。その例を以下に示す。

図 1は、「医療保険」という検索語に対する全文検索結果である。システムは、ユーザが入力した検索語「医療保険」を含む新聞記事を取り出し、そのタイトルと日付を表示する。さらに、取り出した新聞記事と対応関係にある TV ニュースがあれば、そのタイトルも同時に表示する。同じ行にある新聞記事と TV ニュースは内容が対応関係にあるものである。表示されている記事の中には、医療保険制度の改正に関する続報記事がいくつか含まれている。例えば、その中の 1 つである「9 月からの患者負担増を実施」という新聞記事をユーザが選択すると、システムはその記事と続報関係にある記事を表示する。その結果を図 2に示す。この検索結果の中からユーザは、任意の TV ニュース映像あるいは新聞記事をシステムに表示させることができる。

参考文献

[新谷 96] 新谷, 角田, 大石, 長尾: 形態素の共起頻度と出現位置による新聞関連記事の検索手法, 電子情報通信学会技術研究報告, NLC96-1, (1996).

[黒橋 97] 黒橋, 長尾: 日本語形態素解析システム JUMAN 使用説明書 ver.3.4., 京都大学長尾研, (1997).

[奥 96] 奥, 鷲崎, 田中: 関連記事の判定に関する検討, 言語処理学会第 2 回年次大会, (1996).

[大竹 97] 大竹, 増山, 山本: 名詞を中心とした連接に着目した新聞の関連記事検索手法, 情報処理学会研究報告, 97-NL-122, (1997).

[渡辺 97] 渡辺, 岡田, 角田, 長尾: TV ニュースと新聞記事の対応づけ, 人工知能学会誌 Vol.12 No.6, (1997).

[渡辺 98] 渡辺, 岡田, 金地, 阪元: TV ニュースと新聞記事を対象にしたマルチメディアデータベースシステム, 電子情報通信学会技術研究報告, NLC97-, (1998).