ニュース記事一覧作成のための自動要約

中野貴之, 黒橋禎夫, 中村順一
京都大学大学院工学研究科電子通信工学専攻
{t-nakano, kuro, nakamura}@pine.kuee.kyoto-u.ac.jp

1 はじめに

インターネットにおける電子掲示板としてスタートしたネットニュースは, コンピュータネットワークの成長とともにその規模を拡大し, 現在ではインターネット上の一大マスコミュニケーション・メディアとして機能するまでに至っている。

しかし, 日々雑多な情報を含む大量の記事が配送され, それらの記事は内容の質に大幅な差がある。このため, 大量の記事を読んで必要な部分を読者が選別しなければならない。つまり, 得られる情報量のわりに手間と時間がかかるという問題が生じている。

この問題を解決するため, ネットニュースを効率的に利用することを目ざした研究として次のようなものがある。

佐藤らは, アナウンス型ニュースグループで記事のダイジェスト作成 [1] する手法や, 質問回答型ニュースグループで質問文や回答文を抽出する手法 [2] を提案・開発した。これらは, 特定のニュースグループにおいて記事文の形から核となる情報を抽出する手法である。

井佐原らは, 対話型ニュースグループにおいて, リファレンスツリー (参照関係のある記事群) の中で話題が転換した記事を推定する手法 [3] や, 関連ある記事群を抜き出す手法 [4] を提案・開発した。

以上の手法はネットニュースを効率的に利用する手法であるが, 読みたい記事を簡単に探すという用途には不十分である。本研究では, 対話型ニュースグループにおいて記事の取捨選択を容易にする手法として, 記事一覧への自動要約付与を提案する。

2 記事一覧作成のための自動要約

ネットニュースを利用するためのクライアントソフト (一般にニュースリーダという) では, 投稿された記事の一覧として記事番号・投稿日時・差出人・表題などの情報が図1のように表示される。記事の参照関係は, インデントなどを利用して示される。

```
286 11/14 mari-kun@m9.people | バス料金は先？あと？
289 11/14 nao_wada@na.rim.or || バス料金は先？あと？
290 11/15 ka-co@at1.peanet.n ||| バス料金は先？あと？
295 11/16 tar-gz@fan.gr.jp   |||| バス料金は先？あと？
298 11/16 mari-kun@m9.people |||| バス料金は先？あと？
291 11/15 Annie--k@ma2.seiky || バス料金は先？あと？
292 11/15 tishida@quartz.ocn ||| バス料金は先？あと？
294 11/15 portsys@po.aianet. |||| バス料金は先？あと？
296 11/16 Annie--k@ma2.seiky |||| バス料金は先？あと？
293 11/15 take5@mars.dti.ne. || バス料金は先？あと？
299 11/16 seki63@he.mirai.ne || バス料金は先？あと？
```

図1: ニュースリーダによる記事一覧 (japan.town.kyoto)

ここで取り上げられたのは, 日本各地のバス料金の支払い方法について情報を交換しているニュース記事群である。しかし, この記事一覧を見ただけではどのように情報交換が進んでいるのかということが全く分からず, これらの記事が興味あるものであるかどうかの判断材料としては全く不十分である。

そこで, 記事一覧にその内容を端的に示す要約を付与すれば話題とされている内容がはっきりする。各記事の要約と参照関係の情報をあらかじめ読者に提示できれば, 記事の一覧性が高まり, 必要な記事の取捨選択が容易になる。本研究では, このような要約情報を自動的に生成することを考える。

本研究は, テキストの自動要約研究の観点からは次のような位置づけにある。現在のテキスト自動要約技術では文の意味を理解して要約文を生成することはできないが, 文の表層的特徴によって重要文を抽出することはある程度可能となってきている。

しかし, このようなレベルでは, 実際上有効であることは少ない。新聞記事の場合には第1文目を抽出するという方法で十分であり, 科学技術論文などの場合には重要文が適当な精度で抜き出されるだけではほとんど意味をなさない。

ところが, ネットニュース記事のような雑文の場合には, 第1文が重要であるという統一性はなく, またそれほど深い議論がされているわけではないので, 1,2文の重要文を取り出せばそれで大体の内容を把握することができる。さらに, 記事間の参照・引用関係は自動的に

検出できるため，これに重要文の抽出を組み合わせれば，記事群の要約として有効となる可能性がある。

すなわち，現在のレベルの要約技術の非常によい応用例と考えられる。

3 要約作成の手法

3.1 概要

本研究では，表層の手がかりを利用してそれぞれのネットニュース記事から重要文を判断・抽出し，それを要約文として用いることにする。重要文抽出に利用する表層の特徴に対してそれぞれをどの程度重要視するべきであるかは，人手でサンプルを用意し C4.5[5] による決定木学習で自動的に学習する。

構築された決定木を分類器として用いて，記事のおのおのの文の重要度を求める。各文の重要度が求まれば，重要度の高いものを取り出すことで，記事の重要文を抽出することができる。抽出した重要文をリファレンスツリーに各記事の要約として付加すれば，記事一覧における要約が完成する。

以下，決定木学習を行う際に利用した，記事の文に対する表層的な特徴の作成について説明する。

3.2 記事の前処理

まず，特徴を算出するために，記事に対して引用符号の分離と文単位への分割という前処理を行なった。

3.2.1 引用符号の分離

対話型ニュースグループでは，他の記事に対する応答やコメントが多く投稿される。このような応答記事やコメント記事では，元となる記事の一部分を引用して，その後に自分の意見を述べることがよく行なわれる。引用部分は，行頭に引用符号を付けて引用であることを明示するのが習わしである。

そこで，他の記事を参照して引用していることを特徴として利用するため，引用符号の分離を行なった。これは，特定の記号群 (>,), |, :, 〉, 》, ≫) が行頭に表れたかどうか，を検出して分離した。

3.2.2 文単位への分割

重要度の割り当ては，ニュース記事において意味のあるまとまりの単位である文ごとに行う必要がある。しかし，ニュース記事は行の区切りと文の区切りが一致しておらず，行で記事を区切ると文が途中で分断されてしまう。このため，記事を文単位に分割する必要がある。

本研究では，表層的な手がかりを元に記事を文単位に分割した。分割の手がかりには，句点／ピリオド／感嘆符／疑問符／箇条書きなど特定の文字列パターンを利用した。例えば，

来春４月から入居可能の２ベッドルームアパートまたは
マンションを探しています．‖
１件屋を何人かでシェアする下宿スタイルも可．‖
１）ＪＲ京都駅，または京阪３条駅への交通の便がよく‖
２）月７万前後‖
３）１年契約可能な物件‖
こちらは，日本人大学生１人と留学生１人のよていです．‖
（女性）‖
学生用マンション welcome．‖
full furnished（家具付き）だと申し分ないです．‖

のように分割した。句点「．」や箇条書き (「1）ＪＲ京都駅…」など) で文が区切られている。ここで，「‖」は分割時に挿入した分割記号である。

3.3 特徴の算出

要約の自動生成に利用する特徴には，文の形式的特徴と文末表現に着目した言語的特徴を利用した。

3.3.1 形式的特徴

形式的特徴は，引用の深さ・他の記事から引用を受けた参照数・文の長さ・文頭からの距離・最後の引用文からの距離・日本文か英文か，の 6 つを用いた。

引用の深さは，他の記事を引用しているという情報である。引用部分以外は「0」であり，通常の引用部分 (「| 」など) は「1」である。引用が入れ子になっている場合は 2 以上の値をとる。例えば，「| >> 」という引用符号の場合，「>> 」で示された引用部分が投稿者によって「| 」という引用符号で多重引用されたことを示している。この入れ子になっている段階数を算出して，入れ子が 1 つの場合は引用の深さは「2」，2 つの場合は「3」，…とした。

他の記事から引用を受けた参照数は，1 つの記事だけからは直接得ることはできない。参照関係を検出する記事群のおのおのの記事に対し，他の記事を参照している部分を調べ参照先の該当部分に「参照された」という情報を付与するという，以下のような間接的な処理を行う必要がある。

- 元記事と参照記事で，引用符号を除いた部分が共通している行を調べる

- 共通している行があれば，参照記事側に被参照数を加算する

文の長さは，文のバイト数である。重要文が比較的短い事実に着目したものである。

文頭からの距離は，その文が先頭から何文目かを示す。つまり，先頭文から順に 1,2,3,... という値になる。この特徴は，記事の先頭部分に重要文が多いという事実に着目したものである。

最後の引用文からの距離は，その文が最後の引用文から何文目かを示す。これは，引用文の直後に，重要な応答やコメントが記述されることが多いという事実に着目したものである。

日本文か英文かとは，その文の言語を示す。実験対象のニュースグループは日本語の記事がほとんどであり，英文字だけで構成される文は，投稿者のメールアドレスや Web ページアドレスであることが多く，重要であることは少ないという事実に着目したものである。

3.3.2 言語的特徴

言語的特徴は，4 種の記号 (句読点，感嘆符，疑問符，継続「…」) や 22 種の表現 (「ですかね」，「かなあ」等) について，これらの記号や表現が文末に表われるかどうかを示すものである。特徴として含めた文末記号や文末表現は，実験対象の記事に多く含まれる表現を人手で選んだ。

3.4 重要度の付与

決定木学習では，重要度を与えるための手がかりのほかに，重要度そのものの正解となるデータが必要である。この正解データにより学習を行い，どの手がかりを重視もしくは軽視するのかを決定する。

重要度を自動で割り当てる実験を行うため，学習事例記事の各文に対して，正解となる重要度を人手で付与した。

重要度は，(0) 重要でない，(1) やや重要，(2) 重要であるの 3 段階である。

重要度の付与は，その記事の内容を端的に表すと思われる 1,2 文に重要度 2 を与えた。記事の内容を特徴づける部分が長い場合などは，先頭の文に 2 を与えて，残りは 1 を与えた場合もある。

例えば，重要度をつけたテキストは次のようになる。行頭の数字が重要度である。

```
2 : >  あたしが今住んでいる福岡の西鉄バスは「中乗り・
     >  後払い」です。‖
0 :
2 : 福岡のバスは、名古屋とかなり違いますね。‖
1 : まず、福岡・北九州とも「にしてつ」の文字が目立ち、
    福岡で、ＪＲ九州、昭和バスがすこし路線を持つくらい
    なのですが、名古屋では、市バスが大半の路線をにぎっ
    ていて、名鉄、ＪＲ東海、三重交通は名古屋駅から郊外
    への路線が中心です。‖
0 : 福岡市交通局も地下鉄しか経営していないし、北九州市
    営バスも若松区が中心で、市役所には申し訳程度に乗り
    入れているだけです。‖
0 : また、福岡は繁華街天神に郊外や本州・九州各所からば
    んばん乗り入れてくるのに、名古屋の繁華街栄は、殆ど
    市バスの牙城で、名鉄、三重交通の数路線しか乗り入れ
    てないのが特徴ですね。‖
```

「あたしが今住んでいる…」は，引用文中で重要であり，この部分に対する応答が続く。次の文「福岡のバスは…」は投稿者の書いた文では最も重要な文であり，その次の部分「まず，福岡・北九州とも…」は前の文について福岡と名古屋との違いを詳しく説明している文であり，重要度としては一段落ちる。それに続く文はさらに細かい内容であるため重要度は 0 とした。

4 要約作成の実験

4.1 対象とするニュースグループ

実験対象のニュースグループは，対話型ニュースグループ 3 つである。それぞれ約 100 通の記事を取得して，前処理を行なったのち人手で重要度を付与した。

japan.town.kyoto 「京都に関する地域ネタを語ろう」というニュースグループ

fj.sys.mac 「Apple 社の Macintosh および Lisa に関する話題」を扱うニュースグループ

japan.lang.japanese 「日本語に関する話題」を扱うニュースグループ

4.2 学習事例とテスト事例の分割

実際に記事の要約をする状態を考えると，お互い参照関係を持つ記事群に対して要約を行なうことが多いと考えられる。そこで，テスト事例ではお互いの記事が参照関係を持つことが望ましい。

そこで，記事の References フィールドを利用して実験対象記事でリファレンスツリーを構築した。そして，記事を，大きなリファレンスツリーを 1,2 個含むグループ，小さなリファレンスツリーを 4 ～ 8 個含むグループ，どのリファレンスツリーにも属さない独立記事のみを含むグループに分割した。各グループに含まれる記

表 1: 再現率と適合率 (形式的特徴のみ)

正解レベル	寛容		厳格	
ニュースグループ	再現率	適合率	再現率	適合率
japan.town.kyoto	78.2	84.8	66.0	66.4
fj.sys.mac	55.1	79.2	51.0	78.9
japan.lang.japanese	46.3	69.8	30.7	56.8

表 2: 再現率と適合率 (言語的特徴も使用)

正解レベル	寛容		厳格	
ニュースグループ	再現率	適合率	再現率	適合率
japan.town.kyoto	80.1	85.9	68.9	71.2
fj.sys.mac	63.8	78.5	53.1	74.4
japan.lang.japanese	51.6	69.4	31.9	50.1

事数は 20 程度になるようにして 5 つのグループに分割した。

こうして作成された記事グループに対して交叉検定 (cross validation) により実験を行なった。

4.3 実験結果

実験は, 形式的特徴のみを使用して決定木を構築した場合と, 形式的特徴と言語的特徴をすべて使用して決定木を構築した場合の 2 通りを行なった。これにより, 言語的特徴が結果に与える影響を調べた。

結果の評価は, 2 つの正解レベルを設定して行なった。すなわち, 「(1) やや重要」と「(2) 重要である」を同一視して正解とするレベル (寛容) と, 重要度クラスが完全に一致するときのみを正解とするレベル (厳格) の 2 つである。

実験結果の再現率と適合率を表 1, 2に示す。言語的特徴を追加すると再現率は上がるが, 適合率が少々低下するという結果になった。

4.4 自動要約の例

こうして作成した要約の例を図 2に示す。この例では記事の流れを見通しやすくするため, 引用文で重要と判断された文は除外し, 引用文以外で重要度 2 と分類された最初の 1 文のみを, 重要度 2 の文がなければ重要度 1 の最初の 1 文のみを採用している。

図 2の例から分るように, 今回の実験の精度でも, 抽出した重要文をリファレンスツリーとして構成すればそのリファレンスツリーでの話題を掴む目的には十分有効である。したがって, 本研究で提案した手法は, 読みたいリファレンスツリーや記事を選びやすくするという目的をほぼ達成しているといえる。

```
286  <<バス料金は先？あと？>>
|    東日本の政令都市で、横浜、川崎は、公営、民営とも前乗り
|    先払いだったような気がします。
+-289  <<バス料金は先？あと？>>
|  |    札幌は中（後）乗り前降りで後払いです。
|  +-290  <<バス料金は先？あと？>>
|  |      あたしが今住んでいる福岡の西鉄バスは「中乗り・
|  |      後払い」です。
|  |    +-295  <<バス料金は先？あと？>>
|  |    |      京都の場合，運賃均一区間でないところがあるか
|  |    |      ら，今は運賃後払いに統一されているのでしょう。
|  |    +-298  <<バス料金は先？あと？>>
|  |           福岡のバスは、名古屋とかなり違いますね。
+-291  <<バス料金は先？あと？>>
|  |    均一区間のバスと区間外に行くバスとで違っていてやや
|  |    こしかったような・・・。
|  +-292  <<バス料金は先？あと？>>
|  |      いや　後払いだと思います
|  |    +-294  <<バス料金は先？あと？>>
|  |    |      でも、ぼくも、25 年以上前（小学生のとき）、
|  |    |      降れりようとした京都の市バスの後ろの出口には
|  |    |      さまそうになった記憶があるんです。
|  |    +-296  <<バス料金は先？あと？>>
|  |           中学の卒業アルバムを見たら、学校の前のバス停
|  |           の風景を移した写真があって、確かに前から乗っ
|  |           ています。
+-293  <<バス料金は先？あと？>>
|      さて、東京都の場合でも都下（23 区以外）だったりしま
|      すと後乗り運賃後払いだったりします。
+-299  <<バス料金は先？あと？>>
       僕の出身地の仙台では公営、民営とも後ろ乗り後払いで
       す。
```

図 2: 自動要約の例 (japan.town.kyoto)

5 おわりに

本研究で提案した手法は, 個々の重要文の抽出精度はあまり高くないが, 要約全体を眺めれば記事の流れを知ることができるレベルには達している。つまり, 記事の参照関係をもとに重要文による木を構成すれば記事内容の流れを掴むことができる。利用者は, 重要文で構成されたいくつかの木を見れば興味ある記事群を見つけることが可能となる。

今後は, 文の特徴を追加するなどして精度向上をはかり, またネットニュースだけではなく電子メールへの適用も考えていきたい。

参考文献

[1] 佐藤理史, 佐藤円, ネットニュースグループ fj.wanted のダイジェスト自動生成, 言語処理学会論文誌, Vol. 3, No. 2, (1996).

[2] 佐藤円, 佐藤理史, ネットニュース記事群の自動パッケージ化, 情報処理学会論文誌, Vol. 38, No. 6, (1997).

[3] 内元清貴, 小作浩美, 井佐原均, 対話型ネットニュースグループにおける話題転換記事の推定, 言語処理学会第 3 回年次大会発表論文集, (1997), pp. 377–380.

[4] 井佐原均, 小作浩美, 内元清貴, 討論型ニュースグループを対象とする知的ニュースリーダーの開発, 自然言語処理研究会 119-3, (1997), pp. 13–18.

[5] J.Ross Quinlan, *C4.5 : Programs for Machine Langueage*, (Morgan Kaufmann Publishers, 1993), (邦訳:『AI によるデータ解析』 J.R. キンラン, 古川康一監訳, トッパン, 1995 年).