新聞記事の要約のための一手法

渡辺日出雄
日本アイ・ビー・エム株式会社 東京基礎研究所
watanabe@trl.ibm.co.jp

1 はじめに

一般に人間が行なっている要約とは、「元文章の大意を伝えることができる簡略化した幾つかの文を生成する」ことと捉えることができる。しかし、これには深い意味処理と膨大な知識が必要であり、文章の種類により必要なストラテジーも異なってくる。要約結果は、個人の背景知識や興味の対象の違いなどによっても異なり、同じ人間であっても時により結果が異なることが良く見られる。この様な処理を計算機で忠実に行なおうとすると、現在の計算言語学で扱える範囲をはるかに逸脱しているのは明白である。しかしながら、要約処理は必ずしも、深い意味処理が十分発揮できない状態でもある程度可能である場合もある。例えば、自分の知らない分野の文章であってもなんとか要約をすることは可能であり、これは、表層の様々な手がかりを用いて重要そうな部分を判断しているものと思われる。

このような観察から、計算機を用いた要約処理としては、表層の種々の手がかりを用いて重要な部分を認定するというのが現実的である。そこで、従来の要約システムでは、修辞構造や文の表層の手がかりに着目して、それらを元にヒューリスティックスを用いて要約として適切な文を選ぶという方式を取っていた。[1, 2, 3, 5] しかしながら、これらの手法ではどの特徴素をどの程度重要視すべきかなどを決定するのが開発者に依存してしまうので、文章の種類などが変わった場合などにその手法が適当であるかどうかが問題となる。

本論文では、新聞記事を対象として、文を選択するヒューリスティックスの特徴素の重みを実験データから重回帰分析の手法を用いて推定することにより、人間による要約処理に近付ける手法について報告する。

2 文の特徴素

前節で述べたように本論文の手法は、文の表層上の様々な特徴を抽出し、それらを基にして要約を行なおうというものである。実際には以下の特徴素を用いている。

- 重要なキーワードの出現回数: ここでは、重要なキーワードとは、キーワードとして認定されるもののうち、他の文においても出現したものとする。ポイントは、出現個数とする。
- 時制: 過去・現在の区別。これは、事実を述べた文であることも表している。ポイントは、過去であれば0、現在であれば1とする。
- 文のタイプ: 筆者の主張・推測であるか、事実を述べたものであるかを表す。ポイントは、事実であれば0、推測であれば1、主張であれば2とする。
- 接続関係: 理由 (理由を示す)・例示 (具体的な例をあげる)・逆説 (逆のことがらを述べる)・並列 (並列関係を示す)・対比 (比較をする)・接続 (何らかの接続関係がある) に分類する。ポイントは、理由であれば1、例示であれば2, それら以外は0とする。
- 文章の先頭からの位置: 文章中の位置が重要な要因となる場合が多い。ポイントとして、先頭の段落の文に5、以下続く段落毎に1づつ点数を減らして与える。
- 文章の後方からの位置: 前項と同様で文章の後方からの位置にポイントをつける。

時制は、文末に「た」が使われているかどうかで判断した。[1]

文のタイプは、文末の特定の表現を調べることにより行なった。例えば、「すべきだ」「なければならない」等がある場合は主張と判断し、「だろう」などの場合は推測であるとし、それ以外は事実であるとした。

接続関係は、文末と文頭の特定の表現を調べることにより行なった。例えば、文末に「したからだ」があると、前方文に対する理由と見なし、文末に「、など」があると例示とみなし、文頭に「しかし」があると逆説とみなした。[2]

[1] ここでの「過去」は「現在」でないという程度の意味である。厳密には、文末に「た」があるから現在時制でないとは言えないが、ごく少数の例しかないので無視することにした。

[2] これらの文のタイプや接続関係の判定は [6] を基にしている。

3 要約の手順

従来から行なわれている要約の基本的手法は、入力文を幾つかの特徴素で分析し、これらの特徴素を用いたヒューリスティックスにより要約として必要な文を選ぶという様に捉えることが出来る。

本論文では、これをもう少し形式的にして、それぞれの文の特徴素毎に前節で述べたようにポイント P_i[3]を付け、そのポイントとそれぞれの特徴素の重み W_i とを掛け合わせたものの和（と定数項 a の和）をその文の重要度 S とし、与えられた文章から適当な分量だけ高い重要度を持つ文を要約として取り出すということをする。

$$S = a + \sum_{i=1}^{n} W_i * P_i$$

要約処理は、実際には以下の手順で行なわれる。

1. 文毎の重要度を計算する。
2. まだ選択されていない文の中から最大の重要度を持つものを要約文としてマークする。
3. 今選ばれた文が先行文脈中に何らかの接続関係を持つマークされていない文がある場合はそれにもマークをつける。
4. 指定された分量の文がマークされたかを調べ、足りなければ2へ。

基本的に重要度により選択するが、修辞構造 (rhetorical structure) を全く無視してしまうと、全体として読んだ時におかしな文章となるので、重要度とは別のレベルでそれを反映することにしている。

4 重みの推定手法

従来の手法では特徴素の重みを開発者のヒューリスティックスにより決めていたと考えることもできる。本論文では、この重みを実験データから重回帰分析を行なうことにより求める。

実験では、複数の被験者に対し要約として残すのが適切だと思う文に○を付けてもらい、それぞれの文のポイントを (○の付いた数/被験者数) とした。この数を S とすると、

$$S = a + \sum_{i=1}^{n} W_i * P_i$$

という式 (ここで a は定数項、P_i は特徴素 i のポイント) が複数得られるので、重回帰分析により重み W_i の値が得られる。

[3]ただし値は0から1の間になるように正規化しておく

5 実験例

実験は以下のように行なった。被験者として１０名からなるグループＡに日経新聞から選んだ社説記事５件と一般記事３件を見せ、重要と思う文（全体で１／３くらいの分量）に○をつけてもらった。また、その重みの良さを調べるために、グループＡとは別の１０名からなるグループＢに先ほどとは別の社説記事３件と一般記事３件を見せ、同様に重要だと思う文に○を付けてもらった。

グループＢの実験に用いた一般記事と社説記事の一つづつをそれぞれ図２ (a) と図３ (a) に示す。それぞれの文の先頭の数字は文番号であり、次の数字はグループＢで○をつけた人数（支持者数）を示している。

図１は筆者が適当に与えた特徴素の重み（重み１）と、グループＡの実験から重回帰分析で得られた重み（重み２）を示す。重み１は、文章の前方と後方にあるものは一般に重要そうであり、接続関係の値とは重要さが反対の関係にありそうであり、また、社説記事の場合はより主張文の重要さが大きいのではないかという観察から付けたものである。[4]一般記事の重み２を見てみると、後方にあるより前方にある文の方が重要であり、筆者の主張文などはかえって重要でないこと等が分かる。一方、社説の重み２を見てみると、前と後ろにある文及び現在の事実又は筆者の主張を述べている文が重要そうであること等が分かる。[5]

図２ (a) の記事に対して一般記事の重み１と２で文の重要度を計算し、高いポイントを持つものから３文（３割程度）を選んでみるとそれぞれ図２ (b)(c) となる。一方正解として、支持者数が多い文を３つ選ぶと、0,2,3 となる。図３の社説記事に関しても同様に３割程度の８文を選ぶと図３ (b)(c) となり、正解として支持者数が多い文を８文選ぶと 0,2,3,12,15,20,21,22 となる。どちらの重みで計算した方が人間の結果 (正解) と良く合うかを調べるために以下のような乖離度をいう評価式を用いて調べてみた。

[4]この重み１は従来の開発者のヒューリスティックスに頼る手法に相当する。

[5]ただし、実際にはここで得られた重みは開発者が重みを決定するための基礎資料程度にとどめておくのが無難である。統計的に有意な大量の被験者を対象に実験をするのは困難であり、数十人程度だとかなりのばらつきが見られるからである。しかし、だいたいの傾向は分かるはずであり、これを参考に重みを決定すれば良い。

特徴素	一般記事		社説	
	重み1	重み2	重み1	重み2
定数項	0.0	0.183	0.0	0.039
キーワード	1.0	0.216	1.0	0.151
時制	0.3	-0.180	0.3	0.046
文のタイプ	0.3	-0.331	1.0	0.089
接続関係	-1.0	0.127	-1.0	-0.279
位置(前方)	1.0	0.437	1.0	0.242
位置(後方)	1.0	-0.015	1.0	0.214

図1: 特徴素の重み（一般記事と社説）

乖離度=∑(正解で選ばれたがシステムで選ばれなかった文の支持者数)-
∑(システムで選ばれたが正解で選ばれなかった文の支持者数)

図2の記事に関しては、重み1での結果の乖離度は14、重み2での結果は6となり、図3の記事に関しては、重み1での結果は13、重み2での結果は12となり、重回帰分析でパラメーター推定をした結果の方がより被験者に近い要約を生成できることがわかる。グループBに提示した他の一般記事2つについても同様に重み2の方が良い結果が得られた。図3を含めてその他の社説記事2つについては3割程度の分量（だいたい7文程度）で比較するとほとんど同じ結果となったが、これは社説記事に関しては重み1がそれほど悪い値ではなかったからと考えられる。

6 おわりに

本論文では、文の幾つかの特徴素とそれらの重みにより計算された重要度を元に文を選ぶことにより要約をする手法について述べた。特徴素の重みを実験結果を元に重回帰分析で求めることにより、特徴素を新たに増やした場合や、別の分野・スタイルの文章の場合でも簡単に対応が可能であり、開発者の恣意的な値に頼ることなく、より人間の処理に近い要約システムを構築することが可能となる。

現在この要約システムはマルチビューを使った検索支援システム[4]の中で記事内容の概略を見るためのツールとして使われている。

参考文献

[1] 喜多壮太郎：“説明文を要約するシステム”, 情報処理学会自然言語処理研究会報告 63-6, 1987

[2] 住田一男、酒井哲也、他：“自動抄録機能を持つ対話的文書検索システム - システムの機能と構成 -”, 情報処理学会第48回全国大会, Vol. 3, 275–276, 1994

[3] 田村俊哉、田村直良：“文章の表現形式に基づいた要約文章の生成について”, 情報処理学会自然言語処理研究会報告 92-1, 1992

[4] 野美山浩、丸山宏、他：“電子図書館 IV - ナビゲーションシステムプロトタイプ”, 情報処理学会第49回全国大会, Vol. 4, 215–216, 1994

[5] 山本和英、増山潔、内藤昭三：“文章内構造を複合的に利用した論説文要約システムＧＲＥＥＮ”, 情報処理学会自然言語処理研究会報告 99-3, 1994

[6] 渡辺日出雄、辻井潤一、長尾真：“文の表層上の手掛かりを用いたテキスト構造の解析”, 情報処理学会第３２回全国大会, pp.1633–1634, 1986.

0 (10) 【ニューヨーク１０日＝松本元裕】米ＩＢＭは十日、最新鋭マイクロプロセッサー（ＭＰＵ、超小型演算処理装置）「パワーＰＣ」を搭載したパソコンを来年夏に発売することを明らかにした。
1 (3) まずノート型を発売、続いてデスクトップ型二機種を投入する。
2 (5) 低価格で処理速度の高いパワーＰＣは、ＩＢＭ再生のカギを握ると言われるコンピューターの心臓部品。
3 (9) 同社がパワーＰＣ搭載パソコンの商品計画を明らかにしたことで、他の日米欧のパソコン各社も対抗策を迫られることになりそうだ。

4 (3) 発売する三機種はＣＤ―ＲＯＭ、マイク、ステレオオーディオ、音声認識機能を標準装備して、マルチメディア機能を高める計画。
5 (0) ＯＳ（基本ソフト）はＩＢＭの「ＯＳ／２」のほか、米マイクロソフトの「ウィンドウズＮＴ」、サン・マイクロシステムズの「ソラリス」などにも対応できるようにする。

6 (0) パワーＰＣはＩＢＭ、アップルコンピュータ、モトローラの三社が共同開発したＲＩＳＣ（縮小命令セットコンピューター）型ＭＰＵ。
7 (2) パソコン用ＭＰＵ市場で事実上の標準機種になっているインテル製ＭＰＵに対抗するための商品で、低価格・高速処理が特徴だ。
8 (1) パソコン業界二位のアップルがパワーＰＣ搭載パソコンを来年発売する計画を発表。
9 (0) 同一位のＩＢＭはパワーＰＣ内蔵のワークステーションをすでに発売しているが、パソコンについては製品計画を明らかにしていなかった。

10 (1) ＩＢＭはパワーＰＣを外販するだけでなく、搭載パソコンの技術仕様を外部に有償で公開、パワーＰＣ搭載パソコンのファミリー作りを進める計画。
11 (3) アップルなどと合わせたパワーＰＣ搭載パソコン全体で、世界の市場に占めるシェアを最低二〇％程度までもっていきたい考え。

(a) 元記事

2 (5) 低価格で処理速度の高いパワーＰＣは、ＩＢＭ再生のカギを握ると言われるコンピューターの心臓部品。
10 (1) ＩＢＭはパワーＰＣを外販するだけでなく、搭載パソコンの技術仕様を外部に有償で公開、パワーＰＣ搭載パソコンのファミリー作りを進める計画。
11 (3) アップルなどと合わせたパワーＰＣ搭載パソコン全体で、世界の市場に占めるシェアを最低二〇％程度までもっていきたい考え。

(b) 重み１による要約結果

0 (10) 【ニューヨーク１０日＝松本元裕】米ＩＢＭは十日、最新鋭マイクロプロセッサー（ＭＰＵ、超小型演算処理装置）「パワーＰＣ」を搭載したパソコンを来年夏に発売することを明らかにした。
1 (3) まずノート型を発売、続いてデスクトップ型二機種を投入する。
2 (5) 低価格で処理速度の高いパワーＰＣは、ＩＢＭ再生のカギを握ると言われるコンピューターの心臓部品。

(c) 重み２による要約結果

図２:一般記事の要約例

.....

0 (10) ドイツの議会が基本法（憲法）を改正して、貧困など経済的理由による外国人の移住を認めないことにした。
1 (3) 受け入れは政治亡命に限り、いわゆる経済難民を締め出そうというわけだ。
2 (8) 理想主義の後退は残念だが、重い財政負担、ドイツ社会の現状などから見て、やむを得ない措置といえよう。

3 (6) 七月から実施する規制措置は、「迫害のない国」（ルーマニア、ブルガリア、ハンガリーなど）からの亡命は例外を除いて受け入れず、政治亡命を認めている「安全な第三国」（西欧とポーランド、チェコの計十八カ国）を経由してきた難民は経由国に送り返す、というものだ。
4 (1) 東欧革命などで民主主義体制に転換した国々から政治亡命者が出るはずがないという論理である。

5 (2) 第二次大戦後の一九四九年に制定された基本法の第一六条二項は「政治的に迫害された者は庇護（ひご）を受ける権利がある」と規定し、亡命者に寛大だった。
6 (0) これは、ナチ時代の排外主義が他民族に危害を加え、ドイツからも多くの移民を生み出したことに対する反省に基づいている。
7 (0) 社会主義国家の存在を強く意識して、この理想主義的な条項をつくったともいう。

8 (2) だが、ドイツが統一され、東西冷戦が終わると、状況は一変した。
9 (1) 昨年一年間で西欧主要国に流入した難民は申請ベースで七十万人に達し、規制が緩いドイツは四十四万人と全体の六割を超えた。
10 (0) この四月には四万三千人がドイツに亡命を申請し、四万一千人が審査を受けたが、亡命を認められたのは七百人だけだった。
11 (0) 申請者の七三％はルーマニア、旧ユーゴスラビアなど旧ソ連・東欧からの人々で占められる。一方で、不法入国も後を絶たない。

12 (5) ドイツはいま、戦後最悪とも言える不況下にある。
13 (0) 旧西ドイツ地域の失業率は四月で七・一％だが、経済再建が遅れている旧東ドイツ地域は一四・七％と高い。
14 (1) 極右勢力による外国人襲撃など排外的な風潮が強まっているのは、大量に流入してきた外国人が自分たちの職を奪うのでは、という不安があるからだ。

15 (4) 難民はいったん、収容施設に入り、審査が終わるまでそこで暮らすが、その費用負担が各州や市町村に重くのしかかっている。
16 (2) このため、多くの地方政府は難民規制を強く働きかけていた。
17 (0) 西欧各国並みの難民規制を設けようという基本法の改正には、連立与党各党のほか、野党・社会民主党の議員の多くも賛成した。
18 (3) 戦後処理、冷戦の後始末という意味あいもあり、ドイツにとってひとつの時代が終わったという印象を受ける。

19 (0) フランスなど他の西欧各国も経済難民の規制強化に踏み切っている。
20 (4) 貧しい国の人々が豊かな国へ行こうとするのは自然の成り行きだが、大量の移動は混乱と摩擦を生む。
21 (7) 国際協力で経済難民が発生する素地をなくすことがより重要であり、西欧各国が一定の制限を設けるのもやむを得まい。
22 (5) ただ、これらの例をそのまま、国情や環境が違う日本に当てはめて考えるのは問題があろう。

(a) 元記事

0 (10) ドイツの議会が基本法（憲法）を改正して、貧困など経済的理由による外国人の移住を認めないことにした。
1 (3) 受け入れは政治亡命に限り、いわゆる経済難民を締め出そうというわけだ。
2 (8) 理想主義の後退は残念だが、重い財政負担、ドイツ社会の現状などから見て、やむを得ない措置といえよう。
3 (6) 七月から実施する規制措置は、「迫害のない国」（ルーマニア、ブルガリア、ハンガリーなど）からの亡命は例外を除いて受け入れず、政治亡命を認めている「安全な第三国」（西欧とポーランド、チェコの計十八カ国）を経由してきた難民は経由国に送り返す、というものだ。
17 (0) 西欧各国並みの難民規制を設けようという基本法の改正には、連立与党各党のほか、野党・社会民主党の議員の多くも賛成した。
19 (0) フランスなど他の西欧各国も経済難民の規制強化に踏み切っている。
20 (4) 貧しい国の人々が豊かな国へ行こうとするのは自然の成り行きだが、大量の移動は混乱と摩擦を生む。
21 (7) 国際協力で経済難民が発生する素地をなくすことがより重要であり、西欧各国が一定の制限を設けるのもやむを得まい。

(b) 重み１による要約結果

0 (10) ドイツの議会が基本法（憲法）を改正して、貧困など経済的理由による外国人の移住を認めないことにした。
1 (3) 受け入れは政治亡命に限り、いわゆる経済難民を締め出そうというわけだ。
2 (8) 理想主義の後退は残念だが、重い財政負担、ドイツ社会の現状などから見て、やむを得ない措置といえよう。
3 (6) 七月から実施する規制措置は、「迫害のない国」（ルーマニア、ブルガリア、ハンガリーなど）からの亡命は例外を除いて受け入れず、政治亡命を認めている「安全な第三国」（西欧とポーランド、チェコの計十八カ国）を経由してきた難民は経由国に送り返す、というものだ。
4 (1) 東欧革命などで民主主義体制に転換した国々から政治亡命者が出るはずがないという論理である。
19 (0) フランスなど他の西欧各国も経済難民の規制強化に踏み切っている。
20 (4) 貧しい国の人々が豊かな国へ行こうとするのは自然の成り行きだが、大量の移動は混乱と摩擦を生む。
21 (7) 国際協力で経済難民が発生する素地をなくすことがより重要であり、西欧各国が一定の制限を設けるのもやむを得まい。

(a) 重み２による要約結果

図３:社説記事の要約例