自然言語処理研究：「要約する」とはどういうことか－予備的考察

言語・情報講座　舟阪　晃

0　まえがき

われわれは、現在、情報の時代、情報社会に生きている。膨大な情報に取り囲まれ、日々それを適切に処理することが期待されている。情報の洪水のなかで、いかに重要な情報を、また、関連の深い情報を、効率よく入手するかは、学問研究、実業界、また、一般生活において重要かつ不可欠の関心事であるといえる。本稿では、この観点から、「要約する」ということについて考察を加えてみたい。要約というのは、日本では、もっぱら、実用的な面から話題にされることが多く、学問的、理論的意味合いは必ずしも深くないが、外国の文献を調べて見ると、実践的な研究ばかりでなく、理論的な研究も予想外に多く見られる。本稿では、その研究の初歩的な概観を試みているので、予備的考察とした。

1　要約の諸相

要約には、種々の側面があるが、簡単にいえば、原文の縮約 (reduction) といえる。その縮約に、量的、質的要因が関連する。量的な面でいえば、どの程度縮約するかということが問題になり、たとえば、下で取り上げるソフトでは、原文100に対して何パーセントの縮約を求めるか、また、縮約した文の語数を何字にするかを、指定することができる。一方、質的な方は、極端な両極をいえば、原文のなかの重要な一文、または、複数の文を、そのままで

指摘するものから、原文になかった表現を使って、また、要約文の集合のなかに一貫性を実現しようとする高度なものまで考えられる。

要約というのは、人間の頭脳に、かなりの負担を要求するものであることが推測できる。というのは、たとえば、英語の講読の時間に、学生にパラグラフの要約を求めると、全訳でことをすませようとする学生が多くみられる。全訳は、全体の意味がわからなくとも、単語を遂次的に置き換えていけば、一応形にはなり、頭脳の負担はそれほど多くないと思われる。一方、要約をするためには、全体の意味がわかった上で、重要な部分を拾い上げ、さらに高度な要約になれば、原文になかった表現を使ってまとめるという仕事が追加されるわけで、頭脳にとってはかなり疲れる仕事であろうと思われる。

要約は、もっぱら、実用的な観点から言及されることが多いが、人間の頭脳の自然言語処理という観点からも興味のあるテーマであるといえる。前者については、最終的には「自動要約」のシステムの構築がめざされることになり、機械による自動翻訳とともに、コンピュータを活用した言語処理の研究目標のひとつといえよう。また、後者については、頭脳の中で自然言語がどう処理されているかということで、心理言語学、談話文法、自然言語処理、認知言語学などの知見を深めるきっかけになることが期待される。

2　要約ソフト

まず、コンピュータによる要約と、人間の直観による要約を対比することにより、コンピュータでの要約のためにどのような戦術が使われているかを確認する。要約用のソフトは、ワープロに内蔵されたもの、翻訳ソフトに内蔵されたものなど、いくつか見られるが、ここでは、手近の MSWord のツールのひとつとして内蔵されているソフトを使って検討を加えることにする。

2.1 Microsoft Word

このソフトの要約では、量的な制約が 2 種類ある。ひとつは、要約文の語数を指定するもので、もうひとつは、原文に対して何パーセントの要約をす

るかを指定するものである。ここでは、後者の制約をとり、基本的には、10%、20%、30%の要約の結果を比較している。また、どの文を要約文に選ぶかについての基準は公表されていないので、入力として与えた文の行頭の表現を変えることにより、要約の結果が、変わらないか、変わるかに注目し、このソフトが使用している戦術の一部を明かにするつもりである。

2.2 入力用の資料

Richard Carlson : *Don't Sweat the Small Stuff* の日本語訳である『小さいことにくよくよするな！』の最初の文章を取り上げる。この文章は、とくに日本語の場合、パラグラフの取り方が変則的であるので、要約がかならずしも簡単とはいえないものである。また、オリジナルも手元にあるので、日本語と英語での要約のしかたの違いにも注意を払うつもりである。なお、文中の左端の番号は、言及の便宜のために付加したもので、オリジナルにはない。また、入力する際には、タイトルは省いている。

[原文]

(001)「小さいことにくよくよするな ！」[001は原著にある番号]

(1) 私たちは少し頭を冷やせばなんとなく解決することに、つい大騒ぎしがちだ。ちょっとした問題や細かい心配事にいちいち過剰反応してしまう。

(2) たとえば、渋滞の道路で他の車に割り込まれたとする。気にせずほうっておけばいいのに、なんてやつだと怒るのが当然だと思い、頭の中でその相手をやっつける場面を思い描く。そのことが忘れられなくて、後でだれかにグチをいう人もいるかもしれない。

(3) そんな運転手には、どこかで勝手に事故を起こしていただこう。または彼に同情してみて、そんなに急がなくちゃならない事情というのはどんなにつらいものなのか想像しよう。そうすれば自分の幸せを改めて確認できるし、他人のせいで腹を立てることもない。

(4) こんな「小さなこと」は日常ひんぱんに起きる。

(5) 長い行列の順番待ち、身に覚えのないことで非難される、大きな仕事を任される…。そんなとき、くよくよしないコツを知っていれば生き方に大きな差がつく。

(6)「小さいことにくよくよする」ことに生命力を使い果たし、人生の楽しみに気づかない人がどんなに多いことか。

(7) そのコツさえ身につければ、人にもっとやさしくすると同時に、寛容になれるエネルギーが増大することに気づくだろう。

2.3 MS Word の要約方法

2.3.1 要約の結果

要約のしかたには、いくつかの選択肢があるが、本稿では、要約率の変化と、要約の結果を照合し、何を根拠に要約という処理を行っているかを観察することにする。要約率が低い段階で、選択される要約行は、全文のなかの最重要の部分を抽出しているはずである。

まず、原文の要約率を10%、20%、30%と変化させると第1表のような結果がえられる。アステリスクのついている行が要約を構成するための行である。

(第1表)

要約率-→	10%	20%	30%
(1)		**********	**********
(2) たとえば			
(3) そんな			
(4) こんな			
(5)			
(6)	**********	**********	**********
(7) その			**********

上表のように、要約率10%では、(6) 行がこの文章の要点とみなされ、20%以上の要約率では、(1) 行も要点に関与する重要な部分であるとし追加され

る。この結果は、主観的な要約の結果に近いものと考えられる。

コンピュータが要約する場合、default 値は、文章のはじめと [または] おわりになると考えるのがふつうであるが、最後の行である (7) をさしおいて (6) が選択されているのには特別な理由があるはずである。このことは、(7) 行の行頭の「その」の存在が影響しているのではないかと思われる。事実、(7) の行頭の「その」を削除すると、結果は、第2表のように、最初の要約行は、(6) から (7) に変わる。つまり、(7) 行に「その」が付けられると、その行は、その上への結束性を示唆するので、かならずしも最重要の行とは感じられない。この場合、(6) が要約行に選択されるのは、要約率25%を超えてからである。

(第2表)

	10%	20%	30%
(1)		**********	**********
(2) たとえば			
(3) そんな			
(4) こんな			
(5)			
(6)			**********
(7)	**********	**********	**********

同じく、オリジナル文において、(3) の行頭には、「そんな」という、先行文への結束性を示す表現が含まれるために、この行が要約行として選択されるのは、要約率80％になってはじめてである。しかし要約率80%というのは、ほとんど全文を要約文として拾っている状態に近く、そもそも「要約」とはよべないものである。一方、「そんな」を削除すると、要約率65％で、要約行として選択される。　また、(4) の行頭には、「こんな」という表現があるために、要約文として選択されるのは、要約率85%であるが、「こんな」を削除すると、要約率75%で選択される。

以上の点から、このソフトでは、照応関係などの結束性を示す表現が、要約のキューのひとつとして認められていると思われる。

一方、(7) の行頭に「結論として」、「最後に」という、通例結論行を示唆する表現を挿入しても、少なくとも要約率10%から30%に限っていえば、要約行として選択される行に変化は生じない。このことは、(5) の行頭についてもいえる。

それに対して、先行部分への従属性を示す「たとえば」という表現を (6) の行頭に入れると、第3表のような変化が生じる。つまり、このソフトでは、「たとえば」という表現は意味のあるキューになっているといえる。

(第3表)

	10%	20%	30%
(1)		**********	**********
(2) たとえば			
(3) そんな			
(4) こんな			
(5)			
(6) たとえば			**********
(7) その	**********	**********	**********

また、オリジナルの (2) 行の行頭に「たとえば」があるときは、要約行として選択されるのは、要約率60%であるが、この「たとえば」を削除すると、要約率45%で、要約行として選択される。

結論として、このソフトは、「結論として」、「最後に」などの表現には無反応であるが、結束性を示す表現には敏感に反応をするように思われる。また、「たとえば」という表現には、反応がみられることから、これは、「結論として」、「最後に」などとは違って、結束性に影響する表現としてカウントされているようである。

2.3.2 日本文と英文における要約のちがい

上記の文章のオリジナルの英文をつぎに示す。パラグラフのとりかたが、日本語の場合とちがうが、日本語版につけた番号を、英文版に挿入し、日英での要約のしかたに注意してみたい。なお、日本語では、(1) は、二つの文からなっているが、全体としてまとまって解釈されている。一方、英語では、二つの文が、それぞれちがう扱いをうけるので、(1)、(1') と表示している。

[原文]

---

(1) Often we allow ourselves to get all worked up about things that, upon closer examination, aren't really that big a deal. (1') We focus on little problems and concerns and blow them way out of proportion. (2) A stranger, for example, might cut in front of us in traffic. Rather than let it go, and go on with our day, we convince ourselves that we are justified in our anger. We play out an imaginary confrontation in our mind. Many of us might even tell someone else about the incident later on rather than simply let it go.

(3) Why not instead simply allow the driver to have his accident somewhere else? Try to have compassion for the person and remember how painful it is to be in such an enormous hurry. This way, we can maintain our own sense of well-being and avoid taking other people's problems personally.

(4) There are many similar, "small stuff" examples that occur every day in our lives. (5) Whether we had to wait in line, listen to unfair criticism, or do the lion's share of the work, it pays enormous dividends if we learn not to worry about little things. (6) So many people spend so much of their life energy "sweating the small stuff" that they completely lose touch with the magic and beauty of life. (7) When you commit to working toward this goal you will find that you will have far more energy to be kinder and gentler.

上の文章を要約すると第４表のようになる。

（第４表）[英語]

要約率-→	10%	20%	30%
(1)			
(1')			**********
(2)			
(3)			
(4)			
(5)		**********	**********
(6)	**********	**********	**********
(7)			

対比しやすいように、日本語の場合を下に再録する。

（第１表）（再録）[日本語]

要約率-→	10%	20%	30%
(1)		**********	**********
(2) たとえば			
(3) そんな			
(4) こんな			
(5)			
(6)	**********	**********	**********
(7) その			**********

上のふたつの表を対比してみると、(6) 行が要約文の第一候補であるという点は共通している。が、要約率が上がっていくと、英文では (5) が、日本文では (7) が選ばれている。さらに、文章の最初の部分の処理には大きな違いが認められる。日本語で、(1) 行とした部分は２文からなっており、要約率20%以上で選ばれるが、英文では、(1) の後半のみが取り上げられている。

2.4 まとめ

要約という処理が、現実にどのように行われているかを見るために、最初の手順として、手元にあるソフトを調べてみたわけであるが、上の要約の結果を、不出来とするか、案外うまくいっているとするかは、個人の要求水準の問題であろう。ただ、このソフトの要約の方式は、かなり初歩的なものといわざるをえない。つまり、表示された文を変えることなく、重要な文を拾うという操作をしているにすぎない。したがって、要約に貢献する複数の行をまとめて列挙しても、首尾一貫性は認められないという事態も生じることになる。ましてや、複数の行をまとめて、原文に含まれない表現を使ってまとめるということは、まったく不可能である。

要約というのは、主に、実践的な観点から、出来・不出来に注目する傾向があるが、心理言語学、談話文法、自然言語処理、認知など、学問研究面からも考察を加えられるべき問題である。わが国においては、要約はまじめな研究対象になっていないようであるが、そのなかにあって例外的な研究成果として、佐久間まゆみ（編）(1989)：『文章構造と要約文の諸相』（くろしお出版）を忘れてはならない。これは、主に、記述的観点から、要約文の諸相を研究した論文を集めたものである。一方、外国では、記述的な観点よりも、言語処理、認知、人工知能を含むコンピュータ処理などの分野からも多くの発言があり、学際的なひとつの研究分野を構成しているといえる。1998年刊行の Brigitte Endres-Niggemeyer : *Summarizing Information* (Springer) などは、網羅的に要約という現象をとらえており、また、邑本俊亮 (1998) には、これまでの要約についての研究が要領よくまとめられている。本稿の以下の概観では、後者２文献に負うところが大きい。さらに、インターネット上の amazon. com にも要約関係の文献が多くみられる。

3 種々の要約の試み

MSWord の要約ソフトは、ごく初歩的なものであるとのべた。つぎに、これまでにどのような試みがなされたかを、断片的な資料によることになるが、

概観してみたい。

3.1 Brown, Ann L. & Jeanne D. Day (1983)

テキストの要約の成熟度に注目し、被験者を三つにわけてその特徴を明かにしている。第一は、5年生、7年生レベルが示す“copy-delete strategy”による要約で、重要でない表現や冗長な表現を削除し、その残りを要約とする。MSWord の要約は、このレベルに属すると思われる。第二は、高校高学年や大学生レベルで、invention や integration を含む洗練された凝縮 (condensation) 規則を使用する。第三は、“experts, college rhetoric teachers”のレベルで、パラグラフを越えて、「変形規則」を使い、自分のことばでシノプシスを作成できる。

要約に際しては、マクロ規則が使用されるとし、その規則には、削除 (deletion)、一般化 (generalization)、統合 (integration) が含まれている。Kintsch & van Dijk (1978) も、同じくマクロ規則をあげているが、その規則は、削除 (deletion)、上位化 (superordination)、選択 (selection)、創造 (invention) からなっている。

3.2 Rumelhart (1975), Thorndyke (1977)

要約について論じたごく初期の文献だといえる。いわゆる物語文法 (story grammar) の観点から、文章の中に階層を認め、物語の上位レベルは、目標 (goal)、企て (attempt)、結末 (outcome) からなるとし、階層の上位の事柄を拾っていけば、それが要約になるというものである。対象は、“simple action-based folk tales”が中心で、適用範囲が限定されている。

物語文法で扱えるような構造や階層をもつ文章の場合は、ある程度有効だと思われるが、すべての文章が物語を構成するとはいえない。

3.3 Van den Broek, Paul & Tom Trabasso (1986)

要約に際しては、文に内包される命題関係の階層レベルよりも、むしろ、その因果連鎖 (causal-chain) や、連続性 (connectivity) が重要な要因になって

いることを、実験的に明かにした。階層の上位に含まれる項目は、下位の項目よりは、要約に含まれる可能性が高いとはいえるが、階層と因果関係とを比較すると、要約に関係する決定的要因は、因果関係であるとした。

問題点は、すべての文章に因果関係があるとはいえないということである。

3.4 Alterman, Richard (1986)

大要約 (summarization-in-the-large) と小要約 (summarization-in-the-small) に区分する。前者は、テキストを全体として扱い要約を作成し、重要性、興味、顕著などに関する複雑な問題を扱う。一方、小要約では、短縮、ダイジェスト、繰り返しが扱われる。

小要約は、表面的な操作だけを考えているようにみえるかもしれないが、MSWord の要約よりははるかに高度な操作を考えている。つぎの例では、原文にない単語を使用して要約を行っている。

[原文]

Just then the clatter of horses' hooves was heard. And Gessler, the governor general, galloped into the square. His military retinue followed him. He reined his horse to a stop before the pole.

[sum-in-the-small]

The governor general rode into the square. (Alterman 1986 : 73)

このように、原文にない表現を使用するためには、少なくとも ride が gallop よりも上位語であること、また、より一般的な語であることがわかるシステムが必要になる。つまり、語彙項目の概念辞書とよべるようなものが必要になる。さらに、clatter、 horses' hooves、reined his horse などの表現をまとめて処理するいわゆる frame, script などの考え方にも関係してくるものと思われる。

3.5 Alterman, Richard & Lawrence A. Bookman (1990)

要約に関してのコンピュータにもとづいた実験で、濃密さ (thickness) という概念を導入している。濃密さとは、談話の中における概念の間の推論され

る関係の濃度 (p. 143) と説明されている。テキストの濃密さを数量的に測定し、濃密さが大きいほど負荷が多くなるので解釈の負担が大きくなり、要約にはたす役割が大きくなる。

3.6 Lehnert (1982) : Kay & Black (1986)

物語はプロット・ユニットを含む。プロット・ユニットは、種々の状態、事態からなり、物語の流れを維持する。よい要約は、より多くの重要なプロット・ユニットを、また、それらの間の高度の連続性 (connectivity) を含む。Kay & Black は、基本的にはプロット・ユニットを認め、それらを結合する frame を認めている。ユニット間の相互関係が重視されるので、このモデルは、hierarchical ではなく、heterarchical とよばれている。(Kay & Black 1986 : 217) Frame には、総称的 (generic) なものと、領域特定的 (domain-specific) なものがある。

3.7 DeJong (1976) (1982)

FRUMP (Fast Reading Understanding and Memory Program) で有名。トップダウン方式で新聞記事の要約を行う。具体的には、スクリプトを構成する事象のスロットを埋めることによって要約を完成する。スクリプトとしては、demonstrations, earthquakes, visits of statesmen, political conventions, baseball matches (Endres-Niggemeyer 1998 : 313) などがあげられ、それぞれのスクリプトのなかには、複数の event が含まれる。たとえば、demonstrations にはつぎのような events が含まれているという。(Endres-Niggemeyer 1998 : 313)

Event 1 : The demonstrators arrive at the demonstration location.
Event 2 : The demonstrators march.
Event 3 : The police arrive on the scene.
Event 4 : The demonstrators communicate with the target of the demonstration.
Event 5 : The demonstrators attack the target of the demonstration.

Event 6 : The demonstrators attack the police.

Event 7 : The police attack the demonstrators.

Event 8 : The police arrest the demonstrators.

新聞の記事のようにステレオタイプの形式の場合は、ある程度活用できるが、自然言語における要約を考える時には、その目的があまりにも限定されているといわざるをえない。

4 インターネット上の情報

インターネット上で、「要約」、"summarization" をキーワードとして検索をしてみると、主なものとして、次のようなサイトがみつかる。

4.1 東京基礎研究所 (IBM)

自然言語処理のもとに、機械翻訳と自動要約のプロジェクトがあり、自動要約の研究項目として、重要文抽出手法と文章タイプ別要約手法があげられている。

4.2 北陸先端科学技術大学院大学

自然言語処理学講座のサイトで、テキスト自動要約に関する研究というタイトルのページに、研究の概要、発表文献、公開ツールなどが収録されている。

4.3 豊橋技術科学大学

知識情報工学系のサイトで、情報検索とテキスト要約というタイトルでページが設定されている。新聞論説文を含む新聞記事や、TV ニュース文の要約が扱われている。

4.4 University of Ottawa

Knowledge Acquisition & Machine Learning Research Group の Text Summarization Project のサイトで、テキスト要約に関連する文献、研究者、サイト、会

議、プロジェクトなどの情報が収録されている。

5　あとがき

要約には、不確定な要因が多く含まれている。まず、原文に対して、要約がどの程度の縮約を行うかという問題がある。MSWord では、縮約の度合いによる要約のちがいが示されており、適切な配慮がされているといえよう。また、縮約の質的なちがいも多様であり、そのちがいは、どのようなストラテジーで要約を行うかによって決まる。一番単純なものは、MSWord にみられるように、原文にある文のなかから重要な文を、そのまま抜粋するものである。高度なストラテジーをとるものには、本論にあるようにいくつかの方針が区別できる。対象になる原文には、いくつかの種類があり、これまでの試みは、ある特定の種類の原文を想定して行われているといえる。逆にいうと、どのような原文が与えられても、適切な要約を作成するようなモデルは、まだないというしかない。また、インターネット上の情報を調べてみると、自動要約というタイトルのもとで、種々の試みが行われていることがわかる。
(1999年10月29日)

参考文献

Alterman, Richard (1986): Summarization in the small. Sharkey, N. (ed.) (1986): *Advances in Cognitive Science*. Ellis Harwood.

Alterman, Richard & Lawrence A. Bookman (1990): Some computational experiments in summarization. *Discourse Processes*13.143-174.

Brown, Ann L. & Jeanne D. Day (1983): Macrorules for summarizing texts: The development of expertise. *Journal of Verbal Learning and Verbal Behavior* 22.1-14.

DeJong, Gerald (1982): An overview of the FRUMP system. Lehnert, Wendy G. & Martin H. Ringle (eds.): *Strategies for Natural Language Processing.* Lawrence Erlbaum Associates. Pp. 149-176.

Endres-Niggemeyer (1998): *Summarizing Information*. Springer.

Galambos, J. A., R. P. Abelson & J. B. Black (eds.) (1986): *Knowledge Structures*. Lawrence Erlbaum Associates.

Kay, D. S. & J. B. Black (1986) : Explanation-driven processing in summarization : The interaction of content and process. In Galambos, J. A., R. P. Abelson & J. B. Black (eds.) (1998) Pp. 211-236.

Kintsch, Walter & Teun A. van Dijk (1978) : Toward a model of text comprehension and production. *Psychological Review*. 85.5.363-394.

Lehnert, Wendy G. (1982) : Plot units : A narrative summarization strategy. Lehnert, Wendy G. & Martin H. Ringle (eds.): *Strategies for Natural Language Processing*. Lawrence Erlbaum Associates. Pp. 375-414.

邑本俊亮 (1998):『文章理解についての認知心理学的研究――記憶と要約に関する実験と理解過程のモデル化』風間書房

Rumelhart, D. E. (1975) : Notes on a schema for stories. In Bobrow, D. G. & A. Collins (eds.) : *Representation and Understanding*. Academic Press.

佐久間まゆみ（編）（1989）:『文章構造と要約文の諸相』 くろしお出版

Thorndyke, Perry W. (1977) : Cognitive structures in comprehension and memory of narrative discourse. *Cognitive Psychology* 9.77-110.

Van den Broek, Paul & Tom Trabasso (1986) : Causal networks versus goal hierarchies in summarizing text. *Discourse Processes*.. 9.1-15.

分析資料

Carlson, Richard (1997) : *Don't Sweat the Small Stuff···and it's all Small Stuff*. Hodder & Stoughton.

リチャード・カールソン著　小沢瑞穂（訳）(1998):『小さいことにくよくよするな！』サンマーク出版