解説

電脳文章要約術
ー計算機はいかにしてテキストを要約するかー

北陸先端科学技術大学院大学／さきがけ21　佐藤 理史
北陸先端科学技術大学院大学　奥村　学

新聞を開くと，まず，見出しが目に飛び込んでくる．我々読者は，まず，これらの見出しを見て，その記事の内容に当たりをつけ，読むべき記事とそうでない記事をほとんど無意識のうちに，選別する．——何のことはない，我々が毎朝やっていることである．しかし，これが可能なのは，新聞の見出しが，記事の内容を端的に伝える「要約」となっているからである．

要約，要旨，抄録，見出し（目次）といったものは，すべて，伝達すべき内容を凝縮し，それをより短い形で表したものを指す言葉である（以下では，特別な場合を除いて，「要約」をこれらの総称として用いる）．ある程度以上の長さのテキストには，何らかの要約が付けられることが多く，読者の短時間でのおおまかな理解を助ける働きを担っている．要約は，現在，人間の手で作成されているが，これを機械化しようというのが，本稿の主題である自動要約（作成）である．

すでに40年以上に渡って研究されてきた自動要約が，広く脚光を浴び始めたのはつい最近のことである．まず，ワードプロセッサ（ワープロ）が広く普及し，多くの人々が電子的にテキストを作成するようになるにつれ，要約作成支援の必要性が潜在的に発生してきた．そして，自動要約の必要性を痛感させたのが，ワールド・ワイド・ウェブ（World Wide Web）の検索エンジンである．「検索結果として表示されるウェブページリストには，適切な要約がついていてほしい！」．ほとんどの人が，強くそう願っているに違いない．事実，検索エンジンが現れて以来，自動要約に対するニーズが急速に高まってきているのである．

本稿では，まず，現時点で実際的に使える自動要約手法である重要文抽出法について，歴史を振り返る形で述べる．次に，自動要約に関する最近の研究動向を，インターネット関連，放送関連，その他の3つに大きくわけて紹介し，今後の方向性を示す．なお，自動要約に関するより詳細な解説は，文献12）を参照されたい．

要約手法の歴史と現状
ー重要文抽出法を中心としてー

我々人間があるテキストの要約を作成する場合を考えよう．まず，そのテキストをじっくり読み，書かれている内容を十分に理解した後，必要ならば内容を再構成して，それを短い文章に書き出すという過程をとるのが普通であろう．しかしながら，このような「理解－再構成－文章生成」という過程をそのまま計算機でシミュレートすることは，現在のところほとんど不可能である．では，どうするのか．計算機には計算機なりのやり方があるのである．

■Luhnの研究

現在でもよく参照される自動要約の古典は，H. P. Luhnの研究[3]である．Luhnは，テキスト中の重要な文を抜きだし，それを出現順に並べることによってスクリーニング（そのテキストを読むべきか否かを判定する）のための要約が自動生成できることを示した．つまり，「理解－再構成－文章生成」の過程のうち，「再構成－文章生成」の部分をきっぱりとあきらめ，かつ，「理解≃重要部分の同定」と近似すれば，まがりなりにも使える要約が自動的に生成できると主張したのである．日本語の「抄録」という言葉は，「原文から要点を書きぬくこと，ぬきがき」という意味を持つので，この手法は，「自動抄録」と呼ぶのがふさわしいかもしれない．

このような枠組みを用いれば，自動要約の問題は，テキスト中の重要文をいかにして見つけ出すかという問題に帰着する．この問題に対するLuhnの解は，まず，重要語を決定し，次に，この重要語に基づいて重要文を決めるというものであった．

重要語の決定には，単語の頻度を用いる．あるテキスト中に現れる単語の頻度統計をとってみると，おおよそ**図-1**のようなグラフが得られる．ここで，左側に現れる高頻度語は，どんなテキストにもよく現れる，"the"，"a"，"is"などの一般的な語である．これに対して，右側に現れる低頻度語は，テキスト中に1，2回しか現れないため，

そのテキストにおいて，それほど重要ではないと考えられる．残った部分，すなわち，そのテキストにおいてしばしば現れる語（中頻度語）が，そのテキストの内容と密接に関連している重要語と考えるわけである．2つの閾値（図-1における線Cと線D）をどのように適切に決めるかという問題は残るにせよ，この方法を用いれば，単語の頻度を数えるだけで，そのテキストの重要語が決定できることになる．

次は，こうして決定した重要語に基づき，重要文を決定する．基本的には，重要語をたくさん含む文が重要文であるという考え方に従うが，ここで，Luhnは，もうひとひねりし，「多くの重要語が比較的連続的に現れる」文を重要な文とみなす方法をとった．**図-2**は，それを模式的に表したものである．Luhnの論文には，この方法の適用例として，科学論文（The Scientific America）と新聞記事（The New York Times）の要約例が示されている．

Luhnの研究は，その後の自動要約の研究の方向を決定づけた研究といってよい．そのエッセンスは，以下のガイドラインである．

- テキストの一部分を抜き出して，それを並べて，スクリーニングのための要約（indicative abstracts）を作る．
- そのために，文の重要度を計算し，それの高いものを選ぶ．
- 対象テキストは，もっぱら，科学技術論文や新聞記事．

筆者らは，40年たった現在においても，多くの自動要約の研究の根底には，このガイドラインが生きているように思えてしかたがない．

■重要文決定のための各種方法

Luhnのガイドラインに沿うならば，取り組むべき問題は，いかにして文の重要度を計算するかという問題に絞られる．H. P. Edmundsonは，文の重要度を計算する新しい方法を提案し，実験を行った[1]．

ここで用いられた方法は以下の4つである．

1．手がかり語による方法（Cue Method）

重要な文には，それを示すような手がかり語（たとえば，英語では，"significant"，"impossible"，"hardly"など）がしばしば存在するいう事実に基づく．ボーナス（加点）を与える語と減点する語（その語が現れると重要文とはなりにくい）の両方を扱う．

2．重要語による方法（Key Method）

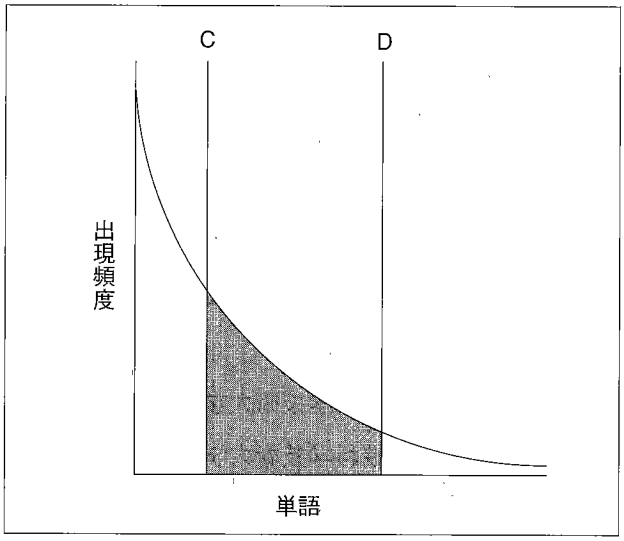
Luhnによって提案された方法の焼き直し．中頻度語は重要語となるという仮定に基づく．文中に重要語が現れたら加点する．

3．タイトルを利用する方法（Title Method）

タイトルや見出しに現れる語は重要語であるという仮定に基づく．文中にその語が現れたら加点する．

4．場所情報を利用する方法（Location Method）

見出しの直後の文は重要な文であることが多い，あるいは，重要な文はテキスト（あるいは段落）の先頭または末尾に現れることが多い，という仮定に基づく．これらの場所にある文に加点する．



X軸に単語を出現頻度順に並べ，Y軸にそれぞれの単語の出現頻度をプロットすると，図のような右下がりの曲線となる．ここで，線Cと線Dに挟まれる部分に入る中頻度語が重要語となる．

図-1　単語の出現頻度曲線

最終的には，以下のように，これらの4つの方法の素点に重みをつけて足し合わせ，各文の重要度の得点を計算する．

$$Score = a_1C + a_2K + a_3T + a_4L \qquad (1)$$

ここで，C，K，T，Lは，上記の4つの方法の素点で，a_iはそれぞれの方法に対する重みを表す．この得点は文抽出にそのまま用いられる．すなわち，要約度N％の要約を生成する場合は，重要度の得点が上位N％に入る文だけを抽出する．

この方法で各種実験を行ったところ，Cue-Title-Locationを組み合わせた方法が最も成績が良いという結果が得られた．また，単独手法では，Keyが最も悪いという結果が得られた．

■重要文抽出法の現状

Edmundsonの研究から，すでに30年が過ぎようとしているが，重要文抽出法は，さらにいくつかの手がかりを考慮するという形で強化され，現在も自動要約手法の主流の地位を占めている．上記の4つ以外の手がかりには，以下のものが用いられる[7]．

1．文間の関係（あるいは，テキスト構造）を利用する方法

接続詞や照応表現などを手がかりにして，文間の関係（理由，例示，逆接，並列，対比など）を決定し，利用する[5]．たとえば，前文と例示関係にある文は，重要文とはなりにくい，など．

2．文間のつながり情報を利用する方法

文や段落をテキストの一単位と考え，それらの間に関連の強さを定義する．関連の強さは，語彙的結束性などを用いて定義する方法[10]と，テキスト間の類似度を直接

用いる方法[9]などがある．多くの文（段落）と強く関連する文（段落）を重要な文（段落）とみなす．

使用する手がかりが増加するにつれて，それらをどのように組み合わせるかが問題となってくる．式（1）のように，重み付き和をとって最終的な得点を決定する場合は，それぞれの重みを適切に決定する必要が生じる．これらの重みを訓練例から学習する試み[11]や，決定木学習の枠組みを適用して有効な手がかりを決定する試み[15]もある．

最近になって，いくつかのワープロソフトや機械翻訳ソフトに，自動要約の機能が組み込まれ始めている．これらの機能がどのように実現されているかは，明確には示されていないが，それらのソフトによって生成される要約を見る限り，重要文抽出法を用いていると推察される．

重要文抽出法による要約の最大の問題点は，抽出された要約が「文章として自然ではない」という点である．元の文章のところどころを抜き出して，それをつなげたものに，文章としての自然さが備わるはずがない．そもそも，この要約法は，スクリーニングのための要約（indicative abstracts）生成のために考えられた方法であり，本文の代替物（要約を読めば本文は読まなくてもよい）としての要約（informative abstracts）を生成することを目的としていたわけではない．もちろん，つながりの深い前後の文はできるだけ一緒に抽出するといった方法を用いることによって，ある程度の軽減は可能であるが，重要文抽出法を用いる限り，この問題は避けられない問題であり，その点を十分認識しておく必要がある．

最近の研究動向と展望

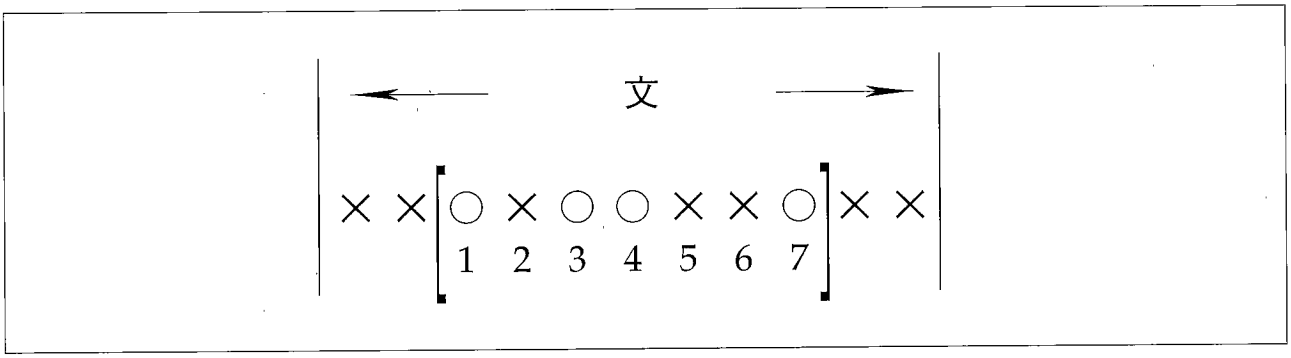
先に述べたように，この1，2年，自動要約の研究は急速な盛り上がりをみせている．自動要約に関するワークショップは，1997年の夏にACL（Association for Computational Linguistics）の大会に併せて開催された[4]のを皮きりに，1998年春には，AAAI（American Association for Artificial Intelligence）のシンポジウムの1つとして開かれ[2]，また，同時期，日本でも言語処理学会の大会に併せて開かれた[18]．さらに，Tipsterと呼ばれる米国のプロジェクトの一貫として，自動要約のコンテストSUMMACが1998年5月に開かれた．

これら最近の自動要約の研究において，筆者らが注目しているのは，作られた要約がどのような目的に使われるのかをより真面目に考慮しようという動きである．使用目的が異なれば，同一テキストに対しても望ましい要約は異なってくる．この点を考慮するならば，ある特定の応用を固定し，それに特化した形での自動要約を考える（応用指向の自動要約）か，あるいは，どのような観点からの要約が要求されているかを推定し，それに応じた要約を動的に生成する（観点を考慮した要約生成）ということになる．現在，研究が進んでいるのは，主に前者であり，インターネットの検索エンジンのための要約や，文字放送や字幕出力のための要約が，これに含まれる．以下では，これらの研究を中心に，現在の研究動向と展望について述べる．

■インターネットにおける要約ー対象テキストの種別に応じた手法ー

検索エンジンの検索結果として得られるウェブページリストに適切な要約を付加して出力することは，今，最も求められている自動要約の応用である．多くの検索エンジンは，現在，そのページの最初のNバイト（タグなどを除く）を出力する，といった単純な方法（場所情報のみを利用した要約）をとっているが，それが十分な品質の要約を提供できていないことは明らかである．

検索エンジンのための要約生成においては，(1) 対象となるテキストが膨大である，(2) 多種多様なテキストが対象となる，(3) テキストの品質が低いものがある，といった特徴があるため，高速かつロバストな自動要約手法でなければ適用できない．ここで，最も大きな問題としてたちはだかるのが (2) である．今までの要約システムが処理対象として想定したテキストは，新聞記事なら



○は重要語，×は非重要語を表す．まず，4語以上離れないで出現する重要語の並び（大括弧の部分）を見つけ，その部分に含まれる単語数をN，その部分に含まれる重要語数をnとするとき，その文の重要度をn^2/Nとして計算する．

図-2　文の重要度の計算

ば新聞記事というように，1つの種別のテキストであったのに対し，検索エンジンのための要約においては，個人のホームページから政府の公文書まで，あらゆる種別のテキストの要約を作り出さなければならない．常識的に考えて，すべての種別のテキストに対して適切な要約を生成できる汎用的な方法があるとは思えない．そのため，まず，種別判定（テキスト分類）を行い，種別ごとにそれ専用の要約方法を適用するといった方法を考える必要があるだろう．

1つの種別に限定するならば，大量かつ低品質なテキストに対しても，既存の要約技術は，比較的うまく動作する．たとえば，ネットニュースの質問記事とそれに対する応答記事から，質問応答集を自動生成するシステムにおいては，質問記事から要約（質問）を自動的に抽出することが実現されている[14]．

■放送における要約－文中の重要個所の抽出－

これまでの要約作成手法の多くは，テキスト中の重要な文を抽出することで実現されていた．しかし，文単位の抽出では，重要でないとして捨てられる情報の単位が文であることから，要約を作成する際に，情報が大きく欠落する可能性がある．このような要約は，indicativeな要約（原文を参照する前段階で用いる）としては問題が小さいかもしれないが，原文の代わりとなるinformativeな要約としては問題が大きい．そのため，文単位で抽出することでテキストを短くするのではなく，一文ごとに重要でない個所を削り（あるいは，重要な個所を抽出し），情報をなるべく減らさずに，テキストを短く表現し直す要約作成手法が近年提案され始めている．これらの手法は，段落，文，節を単位とした重要個所抽出ではなく，句，文字列を単位とした重要個所抽出（不要個所削除[16]）ということができる．

これらの手法の1つといえるのが，文字放送，ニュース番組の字幕を作成することを想定した要約である[17), 13)]．文字放送，字幕の自動作成は，近年需要が大きくなり出していることから，注目を集めている研究である．どちらも，原文の内容を簡潔に画面上に表示する技術であり，要約作成の1つの応用と考えられるが，文字放送，字幕のみで情報として完結している必要があるため，原文の代わりとなる要約を作成する技術が必要である．

文字放送，字幕を作成することを想定した場合，文字放送，字幕では体言止め，漢字熟語などを多用した，固有の表現が可能であること，また，通常の要約と比べると，要約の長さをそれほど短くする必要がないことなどから，不要と考えられる文字列を削除したり，表現をより簡潔な別の表現に言い換えるなど，表層の文字列に関する処理で，ある程度文を短縮することが可能である．文末のサ変動詞を体言止めにする（「7月中に解散します」→「7月中に解散へ」），文末の丁寧の助動詞は削除する（「余震が相次ぎました」→「余震が相次いだ」）などのような変換規則を用意し，文に対し変換規則を繰り返し適用することで，文はより短い文に変換される．

近年モバイルコミュニケーションが脚光を浴びているが，限られた通信・表示リソースしか持たないモバイル端末へのテキスト表示のための要約作成技術の研究も開始されている．この場合も，重要文抽出ではなく，情報をなるべく欠落させず表示する必要があることから，字幕作成の場合と同様な技術が用いられる．

■その他の方向性

上で述べた2つの応用指向の方向性以外で特筆する動きは，2つある．1つは，先に述べた「観点を考慮した要約」であり，もう1つは「複数テキストからの要約生成」である．

観点を考慮した要約[6]とは，「1つのテキストに対して適切な要約はただ1つ存在する」という仮定は成り立たないという主張を含んでいる．たとえば，機械翻訳ソフトの販売開始に関する新聞記事の要約を作る場合を考えよう．「機械翻訳の方式」に興味がある研究者にとっては，そのソフトがどのような方式を用いているかが分かる要約が望ましい．一方，機械翻訳ソフトの購入を考えている人にとっては，ソフトの価格がいくらであるかが分かる要約が望ましい．すなわち，要約の読み手が何に興味を持っているか（ユーザの観点）が異なれば，必然的に異なる要約が必要になるだろうという主張である．情報検索と組み合わせた形で，検索に対する入力（クエリ）を考慮した形の要約を生成するという問題は，興味深い問題であり，今後の研究が期待される．

複数テキストからの要約生成は，関連したテキスト群からそれらに対する要約を1つ作成しようというものである．複数新聞記事を対象とした要約[8]では，(1) ある事件について書かれた記事とその続報記事から要約を作成することと，(2) ある記事に関する複数の情報源（新聞社）の記事から要約を作成すること，の両方が試みられている．このような要約問題においては，もはや重要文抽出法は使えず，まず，テキストをあるレベルで理解し，その結果から要約を生成するという「理解－再構成－文章生成」のアプローチをとらざるを得ない．この理解のレベルとしては，情報抽出の手法によって得られるテンプレート情報が用いられるのが普通である．現時点で実現されていることは，かなり限定されているが，より人間に近い要約方法への方向性を持った研究ということができる．

参考文献

1) Edmundson, H. P.: New Methods in Automatic Extracting, Journal of the Association for Computing Machinery, Vol.16, No.2, pp.264-285 (1969).

2) Hovy, E. and Radev, D. (Eds) : Intelligent Text Summarization, Technical Report, SS-98-06, American Association for Artificial

Intelligence, AAAI Press (1998).
3) Luhn, H. P.: The Automatic Creation of Literature Abstracts, IBM Journal of Research and Development, Vol.2, No.2, pp.159-165 (1958).
4) Mani, I. and Maybury, M. (Eds) : Proceedings of the ACL Workshop on Intelligent Scalable Text Summarization, Madrid, Spain (1997).
5) Marcu, D.: From Discourse Structures to Text Summaries, Proceedings of the ACL Workshop on Intelligent Scalable Text Summarization, pp.82-88 (1997).
6) Ochitani, R., Nakao, Y. and Nishino, F.: Goal-Directed Approach for Text Summarization, Proceedings of the ACL Workshop on Intelligent Scalable Text Summarization, pp.47-50 (1997).
7) Paice, C. D.: Constructing Literature Abstracts by Computer: Techniques and Prospects, Information Processing and Management, Vol.26, No.1, pp.171-186 (1990).
8) Radev, D. R. and McKeown, K. R.: Generating Natural Language Summaries from Multiple On-Line Sources, Computational Linguistics, Vol.24, No.3, pp.469-500 (1998).
9) Salton, G., Singhal, A., Buckley, C. and Mitra, M.: Automatic Text Decomposition Using Text Segments and Text Themes, Proceedings of the 7th ACM Conference on Hypertext, pp.53-65 (1996).
10) Skorokhod'ko, E. F.:Adaptive Method of Automatic Abstracting and Indexing, Information Processing 71, pp.1179-1182, North Holland (1972).
11) Watanabe, H.: A Method for Abstracting Newspaper Articles by Using Surface Clues, Proceedings of the 16th International Conference on Computational Linguistics, Vol.2, pp.947-979 (1996).
12) 奥村　学, 難波英嗣: テキスト自動要約技術の現状と課題, JAIST Research Report, IS-RR-98-0010I, 北陸先端科学技術大学院大学情報科学研究科 (1998).
13) 加藤直人: ニュース文要約のための局所的要約知識獲得とその評価, 情報処理学会研究報告, 98-NL-126-10, pp.69-76 (1998).
14) 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, pp.1225-1234 (June 1997).
15) 野本忠司, 松本裕治: 人間の重要度判定に基づいた自動要約の試み, 情報処理学会研究報告, 97-NL-120-11, pp.71-76 (1997).
16) 山本和英, 増山　繁, 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, 自然言語処理, Vol.2, No.1, pp.39-55 (1995).
17) 若尾孝博, 江原暉将, 白井克彦: テレビニュース番組の字幕に見られる要約の手法, 情報処理学会研究報告, 97-NL-122-13, pp.83-89 (1997).
18) 言語処理学会第4回年次大会ワークショップ「テキスト要約の現状と将来」論文集, 言語処理学会 (1998).

(平成10年12月10日受付)

■自動要約の実行例

本解説の最初の16文（囲みの部分）を対象テキストとして，2種類のワープロソフトで自動要約を実行した結果を以下に示す．どちらのソフトも，重要文抽出法を用いていると思われる．

• ワード98（マイクロソフト）

10％要約（1文）

要約は，現在，人間の手で作成されているが，これを機械化しようというのが，本稿の主題である自動要約（作成）である．

25％要約（4文）

要約は，現在，人間の手で作成されているが，これを機械化しようというのが，本稿の主題である自動要約（作成）である．

すでに40年以上に渡って研究されてきた自動要約が，広く脚光を浴び始めたのはつい最近のことである．まず，ワードプロセッサ（ワープロ）が広く普及し，多くの人々が電子的にテキストを作成するようになるにつれ，要約作成支援の必要性が潜在的に発生してきた．事実，検索エンジンが現れて以来，自動要約に対するニーズが急速に高まってきているのである．

50％要約（8文）

我々読者は，まず，これらの見出しを見て，その記事の内容に当たりをつけ，読むべき記事とそうでない記事をほとんど無意識のうちに，選別する．

要約は，現在，人間の手で作成されているが，これを機械化しようというのが，本稿の主題である自動要約（作成）である．

すでに40年以上に渡って研究されてきた自動要約が，広く脚光を浴び始めたのはつい最近のことである．まず，ワードプロセッサ（ワープロ）が広く普及し，多くの人々が電子的にテキストを作成するようになるにつれ，要約作成支援の必要性が潜在的に発生してきた．そして，自動要約の必要性を痛感させたのが，ワールド・ワイド・ウェブ（World Wide Web）の検索エンジンである．「検索結果として表示されるウェブページリストには，適切な要約がついていてほしい！」．事実，検索エンジンが現れて以来，自動要約に対するニーズが急速に高まってきているのである．

本稿では，まず，現時点で実際的に使える自動要約手法である重要文抽出法について，歴史を振り返る形で述べる．

• 一太郎8（ジャストシステム）

7％要約（1文）

要約は，現在，人間の手で作成されているが，これを機械化しようというのが，本稿の主題である自動要約（作成）である．

25％要約（4文）

我々読者は，まず，これらの見出しを見て，その記事の内容に当たりをつけ，読むべき記事とそうでない記事をほとんど無意識のうちに，選別する．

要約，要旨，抄録，見出し（目次）といったものは，すべて，伝達すべき内容を凝縮し，それをより短い形で表したものを指す言葉である（以下では，特別な場合を除いて，「要約」をこれらの総称として用いる）．要約は，現在，人間の手で作成されているが，これを機械化しようというのが，本稿の主題である自動要約（作成）である．

まず，ワードプロセッサ（ワープロ）が広く普及し，多くの人々が電子的にテキストを作成するようになるにつれ，要約作成支援の必要性が潜在的に発生してきた．

50％要約（8文）

我々読者は，まず，これらの見出しを見て，その記事の内容に当たりをつけ，読むべき記事とそうでない記事をほとんど無意識のうちに，選別する．しかし，これが可能なのは，新聞の見出しが，記事の内容を端的に伝える「要約」となっているからである．

要約，要旨，抄録，見出し（目次）といったものは，すべて，伝達すべき内容を凝縮し，それをより短い形で表したものを指す言葉である（以下では，特別な場合を除いて，「要約」をこれらの総称として用いる）．要約は，現在，人間の手で作成されているが，これを機械化しようというのが，本稿の主題である自動要約（作成）である．

すでに40年以上に渡って研究されてきた自動要約が，広く脚光を浴び始めたのはつい最近のことである．まず，ワードプロセッサ（ワープロ）が広く普及し，多くの人々が電子的にテキストを作成するようになるにつれ，要約作成支援の必要性が潜在的に発生してきた．そして，自動要約の必要性を痛感させたのが，ワールド・ワイド・ウェブ（World Wide Web）の検索エンジンである．

本稿では，まず，現時点で実際的に使える自動要約手法である重要文抽出法について，歴史を振り返る形で述べる．