

特集 自然言語による情報アクセス技術



3

テキスト自動要約

奥村 学

東京工業大学 精密工学研究所
oku@pi.titech.ac.jp

情報アクセス技術としてのテキスト自動要約

インターネットの普及に伴い、WWW上に膨大なテキスト情報が蓄積されるとともに、電子化されたテキスト情報の流通がますます拡大している。このような状況の中、ユーザが必要な情報に効率よくアクセスするためのテキスト処理技術が求められており、さまざまな広がりを見せている。大量のテキストからユーザにとって必要と思われるテキストを検索するテキスト検索技術、テキストから重要（必要）な情報を抜き出す情報抽出技術、テキスト情報から重要な情報のみを選択して提供し、要点の迅速な把握を支援するテキスト自動要約技術、ユーザの目的を反映した質問文に対してテキスト情報を利用した的確な答を自動的に返す質問応答（Q&A）技術など、さまざまな観点から研究が進んでいる。

本稿では、情報アクセス技術としてのテキスト自動要約に関して概観する。特に、現在テキスト自動要約技術が実際に情報アクセス技術としてどの程度利用可能なのかを示す実例を紹介する。

要約の種類

近年研究が活発化するとともに、要約を細分類して整理する傾向が強い。本章では、そのいくつかを紹介する。

要約を研究するにあたって考慮すべき要因として、以下の3つが提示されている。

- 入力の性質—テキストの長さ、ジャンル、分野、単一／複数テキストのどちらであるか、など
- 要約の目的—どういう人が（ユーザはどのような人か）、どういう風に（要約の利用目的は何か）、など
- 出力の仕方

たとえば、入力テキストのジャンルによっては、後述するような重要文抽出による要約が難しいものも考えられるし、また、要約というものの自体が考えにくいものもあり得る。ユーザの持つ予備知識の程度に応じて、要約に含める情報量は変えるべきであると考えられるし、また、利用目的が異なれば、その目的に応じた適切な要約が必要と考えられる。

これらの要因に伴ってまず、要約はいくつかの観点で分類することが可能である。要約対象のテキストが1テキストなのか複数テキストなのかにより、

- 単一テキスト要約
- 複数テキスト要約、

また、特定のユーザに特化した要約なのか、特定のユーザを想定しない要約なのかにより、

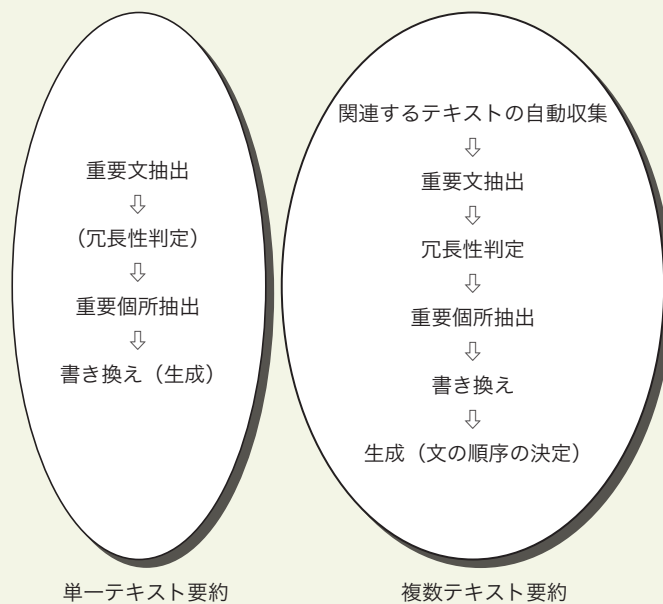


図-1 要約システムのシステム構成

- 'user focused' な要約^{☆1}
- 'generic' な要約

という区分がされる。

利用目的に応じて、要約を次の2つのタイプに分けることも多い。

- 指示的 (indicative) : 原文が読むべきものかどうか、自分の関心に合うかどうかを判断する、原文の適合性を判断するなど、原文を参照する前の段階で用いる。
- 報知的 (informative) : 原文の代わりとして用いる。

いろいろなテキスト自動要約手法

どのような手法がテキスト自動要約にはあるのか。この問への回答となる解説はすでに数多く書かれているので、伝統的なテキスト自動要約手法や、その研究動向に関しては、これまで書かれてきた解説等へのポイントを参照しつつ概説することにする。

テキスト自動要約の手法としては、伝統的に、テキスト中の重要な文を抜き出す重要文抽出法が用いられてきた。1990年代に入り、研究が活発化するとともに、その研究の方向性も多様化し、以下のようないくつかのトピックスでテキスト自動要約手法に関する研究は行われるようになっていく。

- 1) 文から文中の重要個所を抽出することによる要約手法、
- 2) 単一テキストではなく、複数テキストを対象にした要約手法、
- 3) ユーザに特化した要約を動的に作成する要約手法、

1) は、本来は「文字放送、字幕を出力として想定した要約」の項で述べるように、報知的要約を作成するために研究されてきた要約手法であるが、テキストから重要文を抽出する伝統的手法と組み合わせて使用することで、現在の要約システムではほぼ定着しつつある手法といえる。3) については、「テキスト検索時の利用を想定した要約」の項で述べるように、テキスト検索のクエリが与えられて、そのクエリに特化した要約を作成する手法が現在広く使われるようになってきている。2) については、「ナビゲーションのための複数テキスト要約」の項で述べるように、テキスト検索時のユーザナビゲーションに用いる試みが始まりつつある。

単一テキスト、複数テキストそれぞれを対象にした要約システムの構成 (要約プロセス) は概略図-1 のようになる。現時点までのテキスト自動要約に関する研究動向、伝統的な要約手法の詳細に関しては、これまでの解説^{2) ~5)} を参照していただければ幸いである。また、

^{☆1} テキスト検索システムと併用する場合のように、ユーザの情報が問合せ (クエリ; 'query') として明示される場合は、'query-biased' な要約などとも呼ばれる。

自動要約に関する Mani の教科書の翻訳が出版されている¹⁾。さらに、自動要約に関する日本語の(初めての)教科書の出版も予定されている。今年中に出版されることを期待したい。

テキスト自動要約技術の応用

最近ますます要約が我々の身近で活躍する場面は増えてきている。近年検索エンジンが広く利用されるようになってきているが、システムが提示する検索結果には、(リンク先の) Web ページ(テキスト)の内容を短く紹介したものが合わせて提示される場合もある。これは、リンク先のページが、ユーザの欲しいものかどうかを要約を見て判断してもらおうという趣旨で付けられている。また、インターネット上や、携帯電話への配信サービスもある。ニュースの文字放送では、ニュースの原文自体ではなく、その要約といえるようなかたちでニュースが配信されている。携帯端末における情報提示のため、テキスト情報を携帯端末向けに変換(トランスコーディング)するのにも要約技術が用いられている。さらに、市販のワープロソフトの中には、要約機能をつけたものもかなり見受けられるようになってきた。

このように、テキスト自動要約は現在、人間の知的情報アクセス技術として重要視されていることは間違いないであろう。ここで一口に「情報アクセス」といっても、状況により要約の使われ方が異なり、そのため、どういう要約が望まれているかが大きく2つに分けられることに注意していただきたい(図-2 参照)。

上述した最初の例や検索の過程でのユーザのナビゲーション(たとえば、大量の検索結果をクラスタリングなどして、各クラスタに内容の紹介をつけ、ユーザにそれを見て判断してもらうことでユーザを欲しい情報に向けてガイドする)などでの要約は、原文の重要な情報をもれなく含んでいる必要はなく、また、テキストとして読みやすくこなれたものである必要も必ずしもなく、ユーザが検索結果のテキストなり、テキストのクラスタなりの内容を適切に判断できるような要約でありさえすれば(極端な話)それで構わない。また、要約の長さもあらかじめ決まっているわけではなく、「内容を適切に判断できるような」長さを適宜選択できる。

このためテキスト自動要約手法としても、従来伝統的に用いられてきた重要文抽出で十分であり、近年研究が活発であるような「人間にとってより自然な要約」を作成する手法や「原文の情報をもれなく含む要約」を作成する手法など、きめの細かい要約手法を用いる必要は(必ずしも)ないといえる。これは、「要約の種類」の項での分類でいえば、「指示的」要約の範疇といえる要約

である。

一方、テレビのニュース番組における字幕作成や、ニュース記事の文字放送化においても、要約技術が用いられているが、ここでの要約には対照的に「情報の網羅性」と「テキストとしての可読性」が要求される。これらの情報アクセスの状況では、ユーザは、字幕なり、文字放送なりといった要約しか参照できず、原文を参照できない。そのため、要約は原文の代わりという重要な役割を果たす必要があるからである。これは、「要約の種類」の項での分類でいえば、「報知的」要約の範疇といえる要約である。

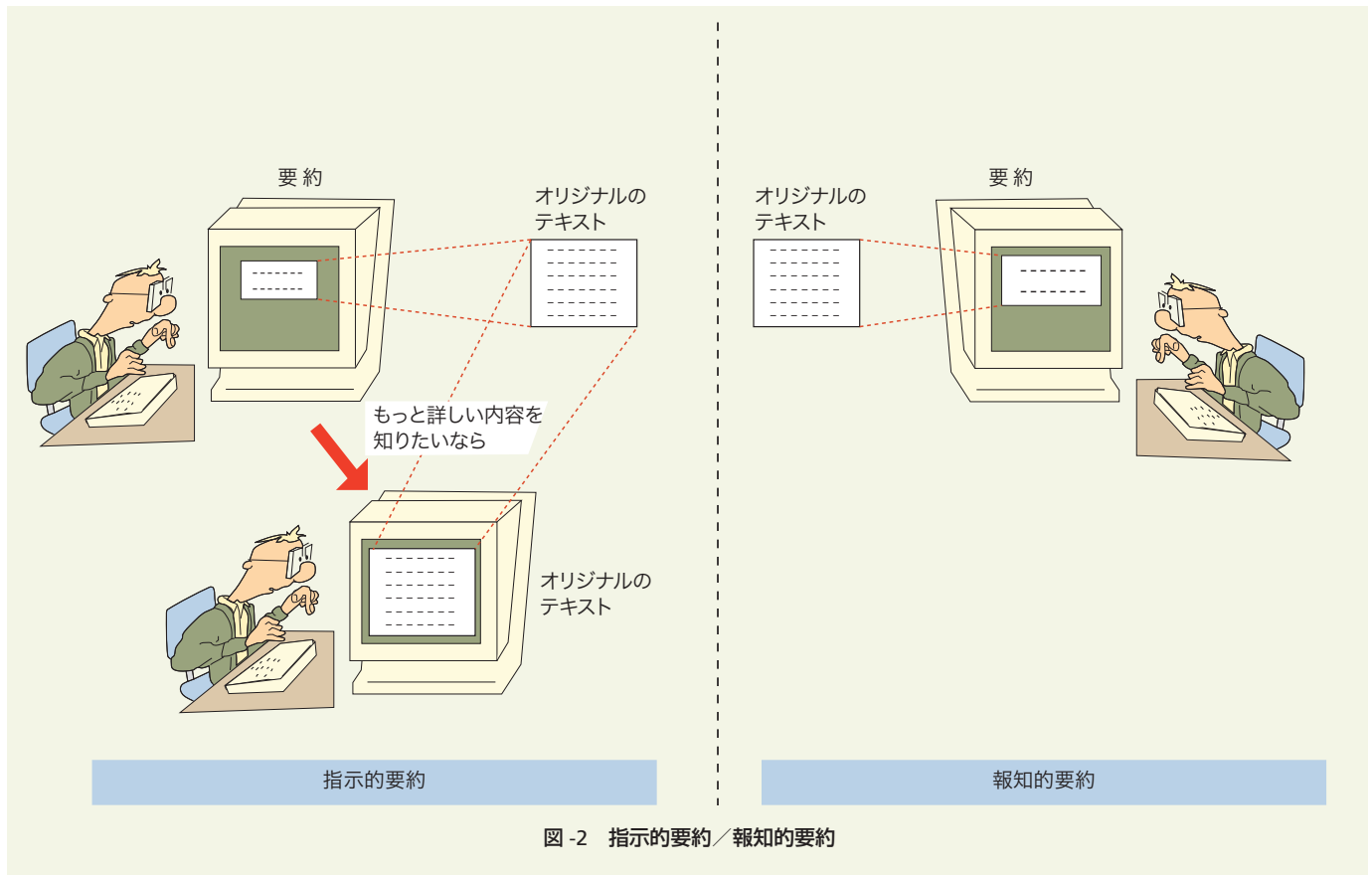
このような状況では、上述した「指示的」要約の場合とは本質的に異なる要約手法を用いる必要が出てくる。以下、それぞれの要約における要約手法について概説する。

テキスト検索時の利用を想定した要約

インターネットの普及により、Web ページが急増するとともに、検索エンジンの数も急速に増大している。通常、検索エンジンを利用する場合、ユーザがクエリを入力すると、検索結果として Web ページのリストが表示され、ユーザはそのリストの中から適切と思われるページを選択するという過程をたどる。だが、「適切なページを選択する」という作業はそれほど容易ではない。従来の検索エンジンでは、ページの内容を判断する情報としてページの先頭部分を表示するという程度の機能しか提供していないからである。このような背景から、検索エンジンには現在、Web ページの内容を簡潔に伝え、ユーザの選択を支援するような要約を提示する機能が求められ、実際に提供されるようになってきている。たとえば、Alta Vista Discovery 検索エンジンのインタフェース中には、InXight 社の要約器が搭載されているなど、検索エンジンの出力表示に要約を用いるというのも現在ではもう珍しくないかもしれない。

テキスト検索において、ユーザがクエリを入力し、検索されたテキストが適切かどうかを判断する際に要約を用いる場合を考えると、要約はユーザが入力したクエリに即したものになっていることが望ましい。しかし、従来の要約作成手法は主に、対象となるテキストの情報を基に、要約は静的に決定できるという考え方で実現されてきたように思われる。したがって、これまでの手法で作成した要約は、このような利用状況では、必ずしも十分ではない可能性がある。

これに対して、近年、要約の利用される状況でユーザの要求に適合した要約を動的に作成する必要があるという考え方に基づいた研究が開始されている。上で述べた、



テキスト検索時の利用を想定した、クエリに即した要約はその一例であるし、ユーザの持つ予備知識の程度に応じて、出力する要約の詳細さ、長さを可変にする試みなどもその一例であると考えられる。

クエリに即した要約を作成する、単純な手法は、重要文抽出の際、クエリと関連する文に、より大きい重要度を与えるものである。提案されている手法はいずれも、クエリ中の単語が文中に出現する頻度に応じたスコアを文の重要度に加味することで、クエリに依存した重要文抽出手法を実現している。

Google では、検索結果としてそのページの snippet (抜粋) を表示するが、その snippet をクエリによって動的に変化させて作成している⁷⁾。これは、「要約の種類」の項の分類でいえば、query-biased 要約といえる。

ナビゲーションのための複数テキスト要約

テキスト検索の結果得られたテキスト集合には、ユーザの目的と合致しないテキストが数多く含まれている可能性がある。このような場合、目的のテキスト集合へユーザをナビゲートする支援システムは有用であり、そのようなシステムでは、テキスト集合を自動的にクラスタリングし、クラスタごとに、クラスタのテキスト集合の

要約を作成しラベルとして付与する。ユーザは、自分の必要なテキストがクラスタに含まれているかどうかを付与されたラベルを見て判断する。

これまでの複数テキスト要約研究では、あらかじめ人間が用意した比較的小規模なテキスト集合をシステムの入力として要約を作成するのが中心的であったといえる。しかし、近年、テキスト検索システムの検索結果を直接要約システムの入力に用いるなど、より大規模なテキスト集合を要約対象とする実用性の高いシステムがいくつか提案されてきている。

上田ら⁶⁾は、クラスタリングによりある程度同じ話題でまとめられたテキスト集合を対象に、各クラスタの特徴を表す文を自動的に作成する手法を提案している。上田らは、テキスト中の各文を構文解析し、テキスト間で構文木同士を比較することで、テキスト間の共通個所を同定するという手法を提案している。

文字放送、字幕を出力として想定した要約

従来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、指示的要約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報知的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、1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開発されている。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）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。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。

対象テキストの性質に特化したテキスト自動要約

「要約の種類」の章で述べたように、テキスト自動要約技術では、対象とするテキストのジャンルを考慮して、要約手法を決定する必要がある。そのため、テキスト自動要約技術を利用した、商用ソフトウェア、利用に供されているサービスなども、対象とするテキストを特化したかたちで提供されていることが多い。本章では、そのような商用ソフトウェア、サービスのいくつかを紹介す

る。なお、ここで挙げるものは、筆者が知る範囲で特徴的なものに限定しており、網羅的なものではないことをお断わりしておく。

テキスト自動要約の対象のテキストのジャンルとしては、新聞記事、学術論文、メールなどがある。新聞記事は、テキスト自動要約の対象テキストとしてこれまで最も活発に研究されたものであり、そのためもあり、現在利用可能なシステムも他のジャンルのものに比べ、質の高い要約を作成できる可能性が高いといえることができる。

ここでは、前回の解説²⁾で詳細に説明した、Web上で利用可能な2つの複数テキスト要約システムについて再度触れる。どちらもWeb上の英文ニュース記事を対象にしている。GoogleでもGoogle News (<http://news.google.com/>) というサービスを始めており、複数のニュース記事をまとめて1つの要約として作成する、これら

のシステムは、Web上で実際に利用可能であることも含め、大変興味深いといえる。

Newsblaster (<http://www.cs.columbia.edu/nlp/newsblaster/>) は、Columbia大学のMcKeownらのグループが開発した要約システムである。このシステムは、CNN, Reuters, Fox News, NY Post, USA Today等の17のニュースサイトから新聞記事を収集し、要約を自動作成する。

NewsInEssence (<http://www.newsinessence.com>) は、Michigan大学のRadevらが開発した要約システムである。このシステムはBBC, CNN, MSNBC, USA Today, Yahoo!の6つのニュースサイトから新聞記事を収集し、要約を自動作成する。

メールを対象とするシステムに関しては、ニーズが高いのか、かなり研究も見られるが、実際に要約を行うシステムが開発された事例は少ない。その中で、沖電気工業がメールを要約して携帯電話等に転送するというシステムを発売している。「電子メールの冒頭の挨拶文や文末などメール本来の主旨から判断し不要な部分を省き、メール本文の中で要点と思われる文章のみを抜き出す」とのことである。小さな画面の携帯電話等でもストレスなくメールを読めるという利点がある。

このように多様な種類の入力を自動要約システムで扱おうとする試みは、自動要約システムがアプリケーション指向でこれからも開発されていくであろうことを考えると、今後も続くと思われ、注目されるトピックということができる。



自動要約ソフトウェア

市販のソフトウェアの中で、要約だけを専門に行うものはあまり見受けられないように思われる。その中で、中国語と英語の要約システムが販売されているようである (<http://naxun.sjtu.edu.cn/paojia.htm>)。上海交通大学の王永成氏のグループのシステムであり、中国語の新聞記事の要約は、Web上で試すこともできるようである (<http://naxun.sjtu.edu.cn/se.asp>)。

同様に、Web上で試すことができる要約ツールとしては、英語を対象にした、MITREの要約ツールがある (<http://complingone.georgetown.edu/~sbj3/compling.html>)。

また、フリーの要約システムとして利用可能なものに（英語を対象にした）Meadシステムがあり、<http://www.summarization.com/mead/> から入手可能である。日本語については、Meadシステムと比べtoyシステムであるが、我々のグループで以前開発したPosumが東京外国語大学の望月源氏のサイト (<http://www.cl.tufts.ac.jp/pub/tools/index-j.html>) から入手可能である。日本語テキストに対する重要文抽出ライブラリとしては利用できるものと思う。

他の情報アクセス技術との統合を目指して

本稿では、情報アクセス技術としてのテキスト自動要約について概観した。本稿では、テキスト自動要約技術を単体として、あるいは、テキスト検索技術との関連で、テキスト検索結果の提示時に利用する話題に終始した。しかし、テキスト自動要約やその関連技術の利用は、これらの状況に限られるものではない。

近年、テキストを利用してさまざまな質問に自動的に答える質問応答技術に関する研究が活発に行われている。質問応答はある意味で「究極の」情報アクセス技術であり、テキスト検索、情報抽出、自動要約など、現在研究されている情報アクセス技術は、質問応答技術の要素技術の位置づけともいえないわけではない。現状の質問応答では、対象とする質問の解答は、固有名や名詞が中心であり、質問応答技術は、テキスト（パッセージ）検索技術と、情報抽出技術の統合技術として実現されている感がある。しかし、質問は現状で扱われている以外に多様なものがあり、ある単語の意味、説明を尋ねる質問（「DNAって何ですか？」）、何かの仕方を探る質問（「シュークリームの作り方を教えて」）、人の意見を尋ねる質問（「自衛隊のイラク派遣についてどういう意見があるの？」）、何かの原因、理由を尋ねる質問（「バブルはどうしてはじけたの？」）等では、解答を含むテキス

ト（集合）から、固有名や名詞よりも長いテキストの断片（パッセージ）を抽出し、（必要なら）それらをまとめて1つのテキストを作り出す（集約）処理が必要になる。これらの処理は、テキストをより短くすることを自動要約というなら、要約とはいえないだろうが、「テキスト検索時の利用を想定した要約」の項で触れたような、クエリに関連したテキストの断片を抽出する処理まで要約に含めるなら、要約技術の範疇に入る処理といえる。

また、テキスト中の固有名の情報を利用したテキスト自動要約手法が一定の成果を上げている。今後、このようないろいろな情報アクセス技術の統合、連携による研究、技術開発が進められていくものと思われる。期待したい。

参考文献

- 1) Mani, I. (著), 奥村 学, 難波英嗣, 植田禎子 (共訳): 自動要約, 共立出版 (2003).
- 2) 難波英嗣, 奥村 学: ここまで来たテキスト自動要約, 情報処理, Vol.43, No.12, pp.1287-1294 (Dec. 2002).
- 3) 奥村 学, 難波英嗣: テキスト自動要約に関する研究動向, 自然言語処理, Vol.6, No.6, pp.1-26 (1999).
- 4) 奥村 学, 難波英嗣: テキスト自動要約に関する最近の話題, 自然言語処理, 「自動要約」特集号, Vol.9, No.4, pp.97-116 (2002).
- 5) 佐藤理史, 奥村 学: 電脳文章要約術—計算機はいかにしてテキストを要約するか—, 情報処理, Vol.40, No.2, pp.157-161 (Feb. 1999).
- 6) 上田良寛, 小山剛弘: 共通意味断片の抽出による複数文書要約, 言語処理学会 第6回年次大会, pp.360-363 (2000).
- 7) 山名早人, 近藤秀和: サーチエンジンGoogle, 情報処理, Vol.42, No.8, pp.775-780 (Aug. 2001).

(平成16年4月19日受付)

