

テキスト自動要約技術の現状と課題

奥村 学, 難波英嗣

1998年2月

IS-RR-98-0010I

北陸先端科学技術大学院大学

情報科学研究科

〒923-1292 石川県能美郡辰口町旭台1-1

{oku,nanba}@jaist.ac.jp

©Manabu Okumura and Hidetsugu Nanba, 1998

ISSN 0918-7553

要旨

本稿では、これまでの(主に領域に依存しない)テキスト自動要約手法を概観する。
特に重要箇所の特定を中心に解説する。

1 はじめに

電子化されたテキストが世の中に満ち溢れ、情報洪水という言葉が使われるようになってからかなりの歳月を経ている。しかし、残念ながら、我々の情報処理能力は、たとえ処理しなければならない情報が増えたとしても、それほど向上はしない。そのため、自動要約技術などにより、読み手が読むテキストの量を制御できることが求められている。また、近年情報検索システムを利用する機会も増えているが、システムの精度の現状を考慮すると、ユーザは、システムの提示した候補が適切なものであるかどうかをテキストを見て判断せざるを得ない。このような場合、要約をユーザに提示し、それを見て判断を求めるようにすると、ユーザの負荷を減らす支援が行なえる。

自然言語処理の分野では、近年頑健な解析手法の開発が進み、これと、上に述べたような、自動要約技術の必要性の増大が重なり、自動要約に関連した研究は、90年代の中頃になって、再び脚光を集め始めている。市販ソフトウェアも続々と発売されており、アメリカでは DARPA 支援の Tipster プロジェクトで要約は新しい研究課題とされている [9]。また、ACL'97では要約に関する併設ワークショップが開催され盛況で活発な議論が交わされた。1998年3月にも、AAAI Spring Symposiumの一つとして、要約に関するシンポジウムが開催される予定である。

本稿では、このような現状を鑑み、これまでの(主に領域に依存しない)テキスト自動要約手法を概観する。また、これまでの手法の問題点を上げるとともに、最近自動要約に関する研究で注目を集めつつある、いくつかのトピックについてもふれる。

要約研究は時に、情報抽出 (Information Extraction) 研究と対で (あるいは、対比して) 述べられることがある。どちらも、テキスト中の重要な情報を抜き出すという点では共通するが、情報抽出は、あらかじめ決められた「枠」を埋める形で必要な情報を抜き出す。そのため、領域に依存してあらかじめ枠を用意する必要があったり、また、領域に依存したテキストの特徴を利用した抽出手法を用いたりするため、領域を限定することが不可欠となる¹。

要約は、原文の大意を保持したまま、テキストの長さ、複雑さを減らす処理とも言えるが、その過程は、大きく次の2つのステップに分けられるとされる: テキストの解釈 (文の解析とテキストの解析結果の生成) および (テキスト解析結果中の重要部分の) 要約文としての生成。しかし、これまでの研究では、これらのステップは、テキスト中の重要箇所の特特定およびその連結による生成として実現されること

¹情報抽出研究に関する解説としては、[5]を参照されたい。

が多かった。そのため、本稿では以後重要箇所の特定を中心に解説する。

2節では、重要箇所特定に用いられてきた、さまざまな情報を取り上げ、それぞれを用いた要約手法について述べる。3節では、それらの情報を統合して用いることで、重要箇所を特定する研究について概観する。4, 5, 6節では、これら以外に特記しておくべきトピックとしてそれぞれ、抽象化、言い換えによる要約、ユーザに適応した要約、複数テキストを対象にした要約に言及する。7節では、要約手法の評価方法について説明する。最後に、8節では、これまでの要約研究の課題および今後の展望について述べ、また、要約技術に関する他の解説へのポイントを示す。

2 重要文抽出に用いられるテキスト中の特徴について

1950年代まで歴史を遡ることができると言われるテキスト自動要約研究のこれまでの多くのものは、テキスト中の文(あるいは、形式段落)を1つの単位とし、それらに何らかの情報を基に重要度を付与し、その重要度で順序付け、重要な文(形式段落)を選択し、それらを寄せ集めることで、要約を生成する。本節では、この重要度評価の際に用いられている、テキスト中の(主に表層的な)情報について述べる。

Paice[29]はこの情報を7つに分類しているが、ここではそれも参考にした上で、以下の7つの情報を取り上げ、各小節で説明する。

1. テキスト中のキーワードの出現頻度を利用する、
2. テキスト中あるいは段落中での位置情報を利用する、
3. テキストのタイトル等の情報を利用する、
4. テキスト中の文間の関係を解析したテキスト構造を利用する、
5. テキスト中の手がかり表現を利用する、
6. テキスト中の文あるいは単語間のつながりの情報を利用する、
7. テキスト中の文間の類似性の情報を利用する

2.1 テキスト中の単語の出現頻度の利用

テキスト中によく出現する内容語はテキストの主題を示す傾向があるとの仮定が情報検索分野などではしばしば用いられる。この仮定に基づき、テキスト中に出

出現頻度の高い名詞をキーワードと考えたり (tf 法), また, これに合わせて, 出現するテキスト数も考慮することで, そのテキスト固有の出現の度合を計算したり (tf*idf 法) など, 情報検索分野では, さまざまな単語の重み付け技法が用いられている [30].

テキスト中の出現頻度に基づき単語に重要度を与えるという, このような考え方を利用し, 単語の重要度を元に, 文に重要度を付与するという重要文抽出手法が, 自動要約研究の開始当初である 1950 年代から提案されている [19, 6, 38, 46]. 単語の重要度から文の重要度を計算する手法はさまざま提案されているが, その一例としては, 文中に出現する単語の重要度の総和を文の重要度とするものがある.

Zechner[38] は, 単語を tf*idf 法で重み付けし, 文中に出現する単語の重みの総和を文の重要度とする重要文抽出手法の評価を行なっている. 人間の被験者の要約と比較し, recall/precision を計算した結果, 人間の被験者同士の比較による精度と大差ない結果を得ている. また, 新聞記事を対象としているので, 先頭数文を抽出する手法との比較を行っており, tf*idf 法を用いた手法の方が良い結果を得ているとしている.

Fukumoto ら [7] は, これに対して, キーワードの重み付け方法として, 領域, 記事, 段落という 3 つの階層を用意し, 段落ごとの出現の度合の異なりよりも, 記事ごとの異なりの方が, また, さらに領域ごとの異なりの方が大きいような単語をキーワードとして抽出する手法を提案している. また, この重み付け手法を用いた重要段落の抽出手法を示している². Zechner の文抽出の手法を段落抽出に応用し, その結果と比較することで段落抽出の評価を行なっている. 評価用のデータのサイズが小さく有意な結果と言えるのかという議論はあるが, キーワードの重み付け手法が有効に働いていることが実証されている.

2.2 テキスト中での位置情報の利用

テキストは, ジャンルに依存して, ある程度構造に規則性が有ると通常考えられている. たとえば, 学術論文は, 序論, 本論, 結論のような構造を持つし, 新聞は, 見出し, 小見出しの後に, 本文が来ることが多い. このような, ジャンルにより決

²Fukumoto らの提案する, 段落の重要度計算手法は, 段落を単語のベクトルとみなし, ベクトル間の類似度により段落のクラスタリングを行ない, より緊密にまとまった段落ほど重要であると考え, 段落間の類似性を利用する」手法 (2.7 節参照) であり, ここで取り上げるべきでないものであるが, 出現するキーワードの重みを総和する手法による段落抽出実験も同様に行なっていること, また, 単語の重み付け手法の新規性が研究の特徴であると考えられるので, 本節で取り上げた.

まったテキストの構造を重要箇所抽出に利用する研究を本節と次節では紹介する。

テキストの構造から、テキスト中での重要な箇所の位置はある程度予測可能であると仮定して、テキスト中での文の位置情報をその文の重要度計算に利用する手法がいくつか考えられている。論説文の場合に、テキスト全体のまとめは書き出しや結び近くにあると仮定するものや、重要な文はテキストの先頭、最後、段落の先頭、最後、節の見出しの直後にあると考える [6] のものなどはその一例と言える。また、新聞記事を対象とした重要文抽出では、本文の先頭数文を抽出するのが良いとされるのも、新聞記事の構造 (本文中では大意をまず先頭に示す) に基づいた位置情報を利用した手法と言うことができる。

Lin と Hovy[18] は、ジャンルや領域に依存して、どの位置に重要な文が来るかを明らかにすることを目的とした研究を行ない、また、もっとも重要な位置を自動的に決定する手法を提案している。TIPSTER プロジェクトで用いられている Ziff-Davis コーパス 13000 新聞記事に対し、テキストごとに話題となるキーワード、要約を手で抜きだし、キーワード、要約の付与されたテキスト集合からの自動学習を試みている。訓練テキスト中の文に対して、含むキーワードの数に応じて重要度を計算し、重要度の順にソートした結果から、何番目の段落の何番目の文が重要であると判断できるかを求めている。タイトルの文、2 番目の段落の先頭文、3 番目の段落の先頭文の順であるという結果が得られている³。評価用のテキスト集合に対して、学習した情報を元に重要文を抽出し、人間の抽出した要約と比較を行ない、評価している。

2.3 テキストのタイトル等の情報の利用

ジャンルにより決まったテキストの構造から得られる、もう一つの情報として、本文以外に、テキスト中に付与されたタイトル、見出しの情報がある。たとえば、学術論文の場合は、テキスト自体がタイトルを持つ場合もあり、また、各章、節にも見出しが付与されることが多い。また、新聞には、見出し (headings)、小見出しが本文とは別に付与されることもある。

このタイトル、見出しは、ある意味で、テキスト本文の非常に簡潔な要約とも考えられる。そのため、タイトル、見出しに現れる内容語を含む文が重要であると考え、タイトル、見出し中の単語を重要文抽出に利用する手法がいくつか提案されている。

³段落に限定すると、2, 3 番目の順で先頭よりもむしろ重要であり、段落中の位置に限れば、先頭に近い文の方が重要であるという結果になっている。

[6]を始めとして、最近では、見出しに含まれる名詞を多く含む文を重要として抽出する [40, 49, 28] などのもその一例と考えられる⁴。

2.4 文間の関係を解析したテキスト構造の利用

自然言語処理の分野では、テキスト中の接続詞等の手がかり語情報などを基に、文間の構造を解析し、テキスト構造を得る研究がいくつか見られる (たとえば, [54, 34, 16])。このようにして得られたテキスト構造を利用して重要文を抽出する研究が近年見られるようになってきている [24, 22]。

Marcu[22]は、修辞構造解析の結果得られる核 (nucleus) がテキスト中の重要箇所検出に有効であるかどうかを実証するため、核と重要箇所間に相関関係があるかどうかを示す実験を行なっている。5テキストを13人の被験者に提示し、3段階評価で重要文の抽出を行なうと同時に、2人の計算言語学者がテキストを修辞構造解析し、構造木を生成した。そして、その構造木を元にそれぞれの文に重要度を付与した。結果として、被験者の生成した要約との比較で、計算言語学者の修辞構造を用いた要約は、再現率 67.5%、適合率 78.5%を得ている。Marcuはまた、450種類の discourse marker を用いた修辞構造解析器を作成し、それを用いた要約生成実験も試みている。結果として、被験者の生成した要約との比較で、修辞構造解析器を用いた要約が再現率 66%、適合率 68%を得ている。

日本語に対しては、Miikeらの研究 [24]がある。接続詞、照応表現などの手がかりを用いた規則集合により、文間の関係を解析し、テキスト構造を抽出するシステムを作成し、得られたテキスト構造に基づき、文に重要度を付与し、要約を作成している。人間の作成したテキスト構造と解析により得られたテキスト構造の比較、抽出した重要文と被験者の抽出した重要文の比較、抽出した重要文を用いて、検索テキストの妥当性判断を被験者に求めた際の、所要時間、判断の精度 (recall/precision) により、テキスト構造解析器およびそれを用いた要約手法の評価を行なっている。

解析により得られたテキスト構造を利用して重要文を抽出する手法の利点としては、

- 長さに応じた要約を、得られた構造木のそれぞれのレベルで生成できる、
- テキスト構造に基づいて重要文を抽出しているため、単語の頻度などを用いた手法に比べ、首尾一貫性の高い要約が生成できる可能性がある

⁴[49]は、見出し中の同じ名詞に関連する文を複数抽出しても冗長であるという考えに基づき、単純に文の重要度の順に文を選択せず、独自の選択手法を提案している。

点があげられる。

2.5 手がかり表現の利用

2.1 節で述べたような、テキスト、文の主題を表す内容語ではないが、テキスト中の重要箇所を指示すると考えられる手がかり語がいくつか存在する。たとえば、学術論文などでは、‘this report’, ‘in conclusion’, ‘our work’などの表現は、論文の主題を表す文とともに出現すると考えられる。このような手がかり表現を利用して、テキスト中の重要文を抽出する研究も存在する [6]。これとは逆に、重要文と負の相関関係にあると考えられる手がかり語を考慮することもできる。たとえば「などの例示を示す接続語で始まる文は重要度が低いと考えられるのはその一例である。

2.6 文間、単語間のつながりの利用

本節と次節では、テキスト中の文間のつながりの情報を重要文抽出に利用する手法について説明する。

Skorokhod’ko[33] は、文をノード、文間の関係をリンクとするグラフをテキストから構成し、多くの文と関係のある文が重要であるという考えに基づき、重要文を抽出する手法を示している。文中の単語が同一概念を参照しているような文間にリンクがあるとしている。

Halliday と Hasan[8] は、表層的な文間のつながりを表す指標として、5種類の結束性 (cohesion)、すなわち、指示 (reference)、代入 (substitution; たとえば、‘a new one’における ‘one’, ‘do so’における ‘do’ などを用いた照応)、省略 (ellipsis)、接続 (conjunction)、語彙的結束性 (lexical cohesion) をあげている。語彙的結束性は、関連性のある語彙が用いられることで、複数の文間の意味的なつながりが明示される場合であり、Skorokhod’ko が文間にリンクを与えたのはこの場合に相当すると考えられる。類義語 (synonym) や、連想により意味的な関連があると思われる単語が、前後する文で用いられ、意味的なつながりを表している。語彙的結束性の情報を得るため、類義語の情報を容易に得られるシソーラスが利用されることが多い。

Hoey[10] は、この語彙的結束性の情報を利用して、文間で単語によるつながりが多いほど、文間のつながりが強いと考え、他の文とのつながりの強さに基づき、要約を生成する手法を示している。

また、語彙的結束性の情報を、互いに関連のある単語のつながりである語彙的連

鎖 (lexical chain)[27]として計算し、それを要約の知識源として用いる研究としては、佐々木ら [43]⁵、Barzilay と Elhadad[2]、望月ら [55]がある。

2.7 テキスト中の文間の類似性の利用

情報検索の分野では、テキスト (や、その断片) を、その中に出現する単語の重みのベクトルとして表現することが多い。このような表現を用いると、テキスト間の類似度は、テキストを表現するベクトル間の内積等で計算することができる⁶。

これと同様に、テキスト中の文 (段落) を一単位として、それらの間の類似度を計算し、この類似度を文 (段落) 間のつながりの度合と考え、この情報を基に、重要と考えられる文 (段落) を抽出する手法がいくつか提案されている。これらの手法は、文 (段落) 間で共通の単語が出現する度合に基づき計算される文 (段落) 間のつながりにより、重要文 (段落) 抽出を行なっていると考えられる。

Salton ら [31] は、段落をノードとし、(ある閾値以上) 類似度の高い段落同士をリンクで結んだ text relationship maps をまず生成し、そこから重要段落を抽出する。重要段落抽出手法としては、次の 4 つを示している。

1. 他の多くのノードとリンクで結ばれているノード (段落) は、複数の段落にわたる主題について議論していると考え抽出する (bushy path)。
2. bushy path 法で抽出したノード同士は、必ずしもリンクで結ばれているわけではないため、それらを連結すると、必ずしも首尾一貫した要約ができるわけではない。depth-first path 法では、最初のノード (先頭段落) あるいは、他のノードとのリンクが多いノードから開始し、それよりもテキスト中で後ろにある、もっとも類似するノードを再帰的に抽出する。互いに類似するノードを抽出していくので、突然話題が変化するというようなことは避けられるが、テキスト全体をカバーする要約が生成できる保証はない。
3. テキストがセグメント (ある話題のまとまり) に分割できるのなら、セグメントからまんべんなく段落を抽出するのがテキスト全体をカバーする要約を作る上で良いと考えられる。しかし、bushy path 法では必ずしもそうはならない。segmented bushy path 法は、セグメントごとに bushy path を抽出する

⁵佐々木らは、語彙的連鎖ではなく、結束チャートと呼んでいる。

⁶ここで利用される単語の重み付け手法及び、ベクトル間の類似度計算手法の数々については [30] を参照して欲しい。

ことで、この問題に対処する⁷。

4. 新しい話題が始まると、通常その先頭の段落で、その主題について述べることが多い。正しくセグメントに分割できるのなら、この考えに基づき、セグメントの先頭の段落を抽出すると良いことになる。segmented bushy path 法は必ずしも先頭を抽出するわけではない。そのため、セグメント中の bushy path 以外に、先頭段落を必ず抽出する手法を augmented segmented bushy path 法と呼ぶ。

Mitra ら [26] は、この手法を、人間の抽出した重要段落との比較により評価している。百科辞典の見出し語に対する項目を対象に、2人の人間の抽出した段落の一致の度合 46% に対し、人間の段落とシステムの段落の一致の度合は、ほとんど変わらないとしている。先頭 20% の段落を抽出した場合が 4 つの手法を上回る精度を得ているが、これは、百科辞典が新聞同様、先頭に見出し語の定義など、主要な内容を含んでいるからと考えられる。

亀田 [40] は、2文間にどのくらい共通な単語 (キーワード) が現れるか⁸に基づいて計算した文間関連度 (の平均) と、ある文が他の文とどの程度広く関連があるかというカバレジに基づいて文の重要度を計算する手法を提案している。

福本 [53] も、文を単語の (重みの) ベクトルとして表現し、ベクトルの内積で文間の結合度を計算し、結合度の高い文を順に抽出する手法を提案している。

3 複数の表層的手がかりを統合して用いる要約手法

前節で述べたように、重要箇所特定には、これまでさまざまな情報が用いられてきている。これらの情報のうち一つを用いた手法も提案されているが、一般に複数の情報を同時に用いた方が精度を改善できるとの考え方に基づき、複数の情報を統合して重要箇所特定に用いる研究が数多く見られる。本節ではそのような研究を概観する。

山本らのシステム GREEN[44] は、以下にあげる 2 つの情報を用いた要約システムである。

文の種類 人手で作成したパターンにより文の種類を分類し、著者の主張等を述べた文 (見解文) のみを抽出する、

⁷Salton らは、map から、内部では互いに緊密に結び付いているが、外部とはほとんど結び付いていない、連続した段落集合をセグメントとして抽出している。

⁸厳密には、どの程度単語の部分文字列に重複があるか

テキスト中での位置 文章，段落の先頭，最後の文を抽出する

この他に，重要文として抽出される文の先頭に指示語や接続詞が出現する場合，あるいは，抽出される文の主語が省略されている場合，前文との結束性が強く，単独で要約中に存在すると不自然であると考え，前文も要約に追加する処理を行なっている。また，重要文中で比較的重要度が低いと考えられる連体修飾句の削除を行ない，さらに要約を短くすることを試みている。要約の評価は，要約の自然さ，内容の適切さに関する被験者へのアンケートにより行なっている。

亀田 [41] は，類似した単語 (キーワード) が 2 つの文に共に出現する度合に応じて計算される文間関連度に基づいて，テキスト中の文の重要度を計算する，以前の手法 (2.6 節参照)[40] に，段落や見出しの情報などを追加して重要文の抽出を行なう手法を提案している。

この手法では，文間関連度に基づく文自体の重要度の他に，段落間関連度を用いて段落の重要度のランキングを行ない，段落の重要度を文の重要度に加味する (重要度の高い段落中の文に，より大きい重要度を付与する)。さらに，見出し文と関連する文 (2.3 節参照) や，重要性を示す機能語句を含む文 (2.5 節参照) を重要視する補正をこれに加え，最終的な文の重要度を計算する。被験者の抽出した重要文との一致の度合により評価を行ない，以前の手法，市販のソフトウェアとの比較の結果，提案した手法が優れていることを示している。

このように，テキスト中の箇所 (文，段落) に重要度を付与する情報を複数統合する際には，情報の統合の方法が問題となる。これまでの研究では，個々の情報により付与された重要度に，それぞれの情報の重みをかけたものを足し合わせ，全体としての重要度とする手法がよく用いられている⁹。

このような重要度計算の手法を用いる場合，それぞれの情報に対する重み付けは，これまで人手でその重みを調整する手法が取られることが多かった。

Edmundson[6] は，2 節で上げた，いくつかの情報を人手で組み合わせた重要文抽出手法による実験を行ない，複数の情報を組み合わせることで，より良い結果が得られることを示している。個々の情報を用いた手法では，「位置情報」，「手がかり表現」，「タイトル等の情報」，「出現頻度」の順に精度が良い。しかし，最初の 3 つの情報の組合せがもっとも良い結果を得られたと報告している。

間瀬ら [39] は，2 節で上げた情報 「タイトル等の情報」，「出現頻度」，「手がかり

⁹ある文に対して複数の情報が矛盾する判断をする場合や，また，複数の情報間に依存関係がある場合などを考慮すると，単純に複数の情報をスコアとして統合するのが適切なのかという疑問はある。

表現」, 段落中での位置情報」¹⁰に対応するパラメタおよび, 主題」(助詞「は」が後置する名詞), 文間の接続詞」などの情報を基にしたパラメタを組み合わせて利用する重要文抽出手法を示している. 文の重要度は, 各パラメタのスコアに, 人手で決定したパラメタの重みをかけたものを総和することで計算している. さらに, 抽出した重要文に指示語が存在する場合, 先行詞を含む文を推定し同時に抽出したり, 不要と考えられる接続詞や副詞を削除するなどの後処理を加えている.

これに対し, 近年その重みを, 要約文集合を訓練コーパスとして用いる自動学習により決定する手法が盛んに研究されるようになってきている. 自動学習による決定手法は, その学習法の違いから次の3つに大別できる.

- 確率を用いたベイズ推定手法 [15, 12, 35]
- 重回帰分析を用いた手法 [36]
- 決定木学習を用いた手法 [56, 1]

Kupiec ら [15] は, 重要文抽出を統計的な分類問題とみなし, あらかじめ人手で選択した重要文を訓練集合とし, 文が重要文集合に含まれるかどうかの確率を与える分類関数を求めておき, 重要文はこの確率により文を順序付けることで抽出する重要文抽出手法を提案している. 具体的には, 属性 (重要文抽出のための情報) 集合が与えられた時の, 文 s が要約 S に属す確率 $P(s \in S | F_1, F_2, \dots, F_k)$ を計算する.

$$P(s \in S | F_1, F_2, \dots, F_k) = \frac{P(F_1, F_2, \dots, F_k | s \in S) P(s \in S)}{P(F_1, F_2, \dots, F_k)}$$

属性間の独立性を仮定して,

$$P(s \in S | F_1, F_2, \dots, F_k) = \frac{\prod_{j=1}^k P(F_j | s \in S) P(s \in S)}{\prod_{j=1}^k P(F_j)}$$

ここで, $P(s \in S)$ は定数であり, $P(F_j | s \in S)$, $P(F_j)$ を訓練コーパスから求めておく. 訓練コーパスとしては, (テキストの著者以外の) 人手により作成した 188 テキストに対する要約を利用している.

属性集合としては, 2 節で述べた「出現頻度」, 「手がかり表現」, 「位置情報」に対応するものの他に, 「短い文は重要文になり難い」, 「固有名詞は重要であることが多い」という仮定を導入するための属性も利用している. 人手で属性の重みの決定を行なった Edmundson の研究と基本的に結果が一致している. また, テキスト

¹⁰間瀬らは, 段落を表層的な情報を基に分割した, セグメントという意味的なまとまりも考慮に入れている.

の先頭の文を抽出する手法をベースラインとすると、ベースラインの精度 24% に比べ、42% 平均の適合率を得ている。

Watanabe[36] は、属性集合として、2 節で述べた「出現頻度」、位置情報」に対応するものの他に、「時制情報」(現在/過去)、「文のタイプ」(事実/筆者の主張/推測)、「前文との接続関係」(理由/例示/逆説/並列/...) を利用している。これらの属性集合それぞれの各文に対するスコアに属性集合の重みをかけたものの総和を文の重要度とするが、属性集合の重みを訓練データから重回帰分析を行ない求めている。

Nomoto[56], Aone ら [1] はそれぞれ、人手による要約文を訓練データとし、決定木学習アルゴリズム C4.5[13] を用いて、そこから学習した決定木による文分類により重要文抽出を行なう手法を提案している。Kupiec らの手法と同様に、テキスト中の文が重要文集合に属すかどうかを分類する分類器を学習するが、ここでは、確率ではなく決定木を学習する。決定木学習は、あらかじめ分類済みの訓練データに属性情報を付加しておき、そのデータを正しく分類できるようなルール集合を決定木の形で学習することになる。

4 抽象化, 言い換えによる要約手法

これまで述べてきた研究はどれも、要約文 = 抽出したテキスト中の重要箇所の「連結」という考え方に基づいていた。これは、要約を「原文から(何も変えずに)抽出した extract(抜粋)」と見なしているとも言うことができる。これに対し、言い換えたり、新たに合成したりすることで、原文の内容を表現する abstract を要約として生成する試みが近年いくつか見られるようになってきた。

この abstract の生成のためには、通常の extract(テキスト中の重要概念の発見) 以外に、発見した概念の統合、生成の過程が必要である。概念の統合は、抽出された複数の重要概念を、何らかの知識を用いて、より高い階層の概念にまとめることである。これにより、テキスト中の重要概念は、より少ない数の概念で表されることになる。概念の統合には、概念階層やスクリプトといった知識が必要となる。Hovy らの SUMMARIST[11, 17] システムは、WordNet を概念階層として利用し、このような概念統合を実現している。Hovy らは、

John bought some vegetables, fruit, bread, and milk.

のような文を、概念階層を用いて、

John bought some groceries.

のように言い換える処理を，概念階層を用いた概念統合の例として示している。

KondoとOkumura[14]は，Hovyらの用いている概念階層以外に，EDR単語辞書中の単語の定義文をスクリプト知識と見なし，定義文を利用して，概念統合を実現する手法を示している。定義文は見出し語の説明であるため，逆に見出し語は定義文中に出現する動作系列の簡潔な言い換えになっていると考えられる。たとえば，

説得する：よく話して納得させる

納得する：物事を理解して承認する

承認する：相手の言い分を聞き入れる

のように，それぞれの単語の定義文が与えられている場合，

私は彼女に事情を話した。

彼女は私の言う事を理解し，

聞き入れてくれた。

のような文は，上の3つの定義文を利用し，定義文中の動作系列を再帰的に見出し語に言い換えることにより，

私は彼女を説得した。

のような文に要約できると考えられる。

5 ユーザに適応した動的な要約手法

これまでの要約研究は主に，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進められてきたように思われる。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生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 [28]。

たとえば，情報検索において，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必要があり，これまでのように，テキストの内容のみから作成していたのでは必ずしも十分とは言える。

動的な要約手法に関する研究は，まだ始まったばかりの段階と言えるが，そのための基礎的な調査を難波ら [50] が行なっている。難波らは，被験者が重要文抽出を行なう際に，あらかじめ観点を与えられている場合と与えられていない場合とで，抽出される重要文にどのような違いがあるのかを調べる心理実験を行っている。この調査の背景には，

一般にテキストの主題は複数あり得る。また、テキストを書き手が1つの主題で書いたとしても、読み手によって主題と受け取るものは様々である可能性があり、個々の文の重要度は読み手の観点によって変わってくるであろうと考えられる。よって、テキストに対する観点を与える場合と、与えないで、読み手が自由に解釈する場合は、テキストのどこが重要かということが異なるはずである。また、観点を与えない場合よりも与えた場合の方が、被験者が重要と考える文のばらつきは減るはずである。

という仮説がある。

日本語情報検索システム評価用テストコレクション BMIR-J1[52]中の25テキストを用い、6人の被験者に対して、対象テキストを正解(検索されるべきテキスト)とする検索クエリを観点として実験を行った結果、被験者にあらかじめ観点を与えるのと与えないのでは抽出される文に差が生じることが確認されている。このことから、情報検索システムにおける利用などでは、対象テキストの内容だけでなく、検索クエリ等の、その他の利用可能な情報も用いた動的な要約生成手法が必要であると言える。

6 複数テキストを対象にした要約手法

これまで、単一テキストの要約生成に関する様々な手法について述べてきた。要約対象のテキストが複数で、対象テキスト間で重複する記述がある場合、単一テキストからの要約生成手法で、個々のテキストから要約を生成して並べただけでは、要約中に重複した内容が含まれる可能性があり、冗長で、要約として適切ではないと考えられる。そこで、関連するテキストをまとめて対象とし、そこから要約を生成する手法が重要となる。以下では、複数テキストからの要約生成のいくつかの手法について述べる。

Yamamotoら[37]、船坂ら[48]、McKeownら[23]、Maniら[21]、柴田ら[45]はいずれも、複数の新聞記事を対象に研究を行っている。複数の記事を対象とする場合、ある記事で述べられていることが続報記事で再度述べられており、その冗長な重複箇所をどのように検出し、削除するかが問題となる。

船坂らは、このような冗長な箇所を導入部と呼び、「したが」、 「事件で」といった導入部特有の表現に着目して導入部を特定し、導入部中の名詞の7割以上が古い記事中の名詞と一致すればその導入部は冗長であると考え削除している。

柴田らは、Fit という検索システムに文章融合機能を埋め込み、自動分類されたテキスト (新聞記事) の融合を試みている。出現頻度の低い名詞が異なる記事の文中に出現する場合、それらの文は重複文である可能性が高いと考え、重複文を同定し、その片方を用いて要約生成を行っている。

Mani らは、個々のテキスト (記事) を、関連する語句 (ノード) の間にリンクを張ったグラフで表現している。そして、活性伝播により、テキストの話題と関係するノード集合 (サブグラフ) を検出し、複数のテキスト間でそれらのサブグラフを照合することで、テキスト間の類似箇所と相違箇所の抽出を行っている。

McKeown らは、MUC 用に開発されたシステムが出力した枠 (template) から情報をまとめ上げ、複数の新聞記事の要約を試みている。記事はテロリストに関するもので、枠により犯人、犠牲者、事件のタイプなど、25 種類の情報が抽出される。McKeown らは、要約コーパスを経験的に分析し、どの情報を要約中に含め、どのように情報をつなぎ合わせ、要約を生成するかを決定する「要約オペレータ」および、要約中で用いる語、構文を同定し、抽出された情報を 7 種類のオペレータで取捨選択し、要約の生成の際には、要約コーパスから同定した語、構文を用いている。

これらとは異なり、難波ら [51] は、特定の研究分野のサーベイを自動作成することを目指し、複数の学術論文を要約対象にし、それらから 1 つの要約を作成する研究を行なっている。難波らは、論文間の参照関係に着目し、要約対象論文を参照している論文中の、要約対象論文について記述している箇所 (参照箇所) を抽出し、それらを要約の原型としている。抽出した参照箇所から要約を生成するには、複数の参照箇所間の類似点、相違点を明確にする必要があるが、それを実現するまでには至っていない。

新聞記事と比較して、難波らが対象とする学術論文では、テキストごとに書き手のスタイル等が異なる度合が顕著であり、複数テキストを融合して要約生成を行う際には、スタイル等の統一に関する処理が別途必要となる可能性がある。

7 要約手法の評価方法について

これまでの単一テキストを対象とする要約研究の多くは、人間の被験者の生成した要約文と、システムの生成した要約文を比較し、システムの要約文の再現率、適合率を評価尺度とした評価を行なっていた。

しかし、人間においても要約というタスクは必ずしも容易ではなく、人間の被験者による要約が必ずしも高い割合で一致するとは言えない。また、この評価法の前

提とする「ただ一つ正しい要約が存在する」という仮定が不自然であるという批判が以前からあり、要約システムの評価法は再検討される段階にあると言える。

これに対して、Miikeら [24] は、要約を利用して人間がタスクを行なう場合の、タスクの達成率が間接的に要約の評価となるという考え方に基づき、評価を行なっている。具体的には、情報検索における検索テキストの適切性の判断をする際に要約を用いることで、要約を評価し、タスクに要する時間と検索の再現率、適合率で評価を行なっている。

DARPA Tipster プロジェクト (Phase III)¹¹の評価 [9] においても、同様に、上の仮定の不自然さから、タスクに基づく評価方法が採用される予定である。Tipster プロジェクトでは、テキストの分類、情報検索における検索テキストの適切性の判断それぞれに要約を利用し、被験者のタスクに要する時間 (要約しないテキスト全体を用いた場合とも比較する)、タスクの精度により要約を評価する。

また、要約を用いて人間の被験者が何らかのタスクを実行する際の精度等を問題にするのではなく、要約を利用して何らかのタスクを実行する応用プログラムの精度を示すことで、間接的に要約の評価を行なうという試みも見られる。隅田ら [42] は、抽出した要約文のみを索引およびスコアづけの対象としたテキスト検索システムの評価を行ない、テキスト全体を索引等に用いた場合に比べ、精度の向上が実現できることを示すことで、抽出した要約文がテキストの大意の把握に成功していることを間接的に実証している。良い要約が得られれば、重要な概念や単語のみが索引語として利用されるので、検索の精度が改善されるはずであるという仮定にこの研究は基づいている。

間瀬ら [39] は、原文を読んだ後および、その要約だけを読んだ後、原文の内容を問うテストを被験者に行ない、テストの得点比で要約の評価を行なっている。テストの問題作成の困難さが問題点として残るが、原文を伴わない状況での利用を想定した要約の内容の十分性の評価としては興味深い手法である。

さらに、要約文の内容に関しては問わず、「文章としての読み易さ」のみを評価する評価法も考えられる [25]。

一方、複数テキストを対象とする要約研究の評価は、単一テキストの場合と比較して、研究が始まったばかりでもあり、また、難しくもあるので、十分な議論がなされてきていないと言って良い。

6 節で述べたように、複数テキストを対象とする場合、冗長な重複箇所を検出し、削除することが必要となるため、「冗長箇所をどの程度正しく削除できているか」

¹¹1998 年春に予定されている。

[48], 「テキスト間の類似箇所と相違箇所をどの程度正しく抽出できているか」 [21] という観点での評価は行なわれている。また, 難波ら [51] は, 「要約に必要な記述内容 (参照箇所) をどの程度正しく抽出できているか」を評価している。

しかし, 複数テキストから作成された要約文全体に関する評価はこれまでなされておらず, どのような点を評価すべきかということも明らかではない。今後, 生成された要約全体の評価について検討していく必要があると考えられる。

8 おわりに

これまでの要約研究として, テキスト中の重要箇所の特定とその連結による生成に基づくものを主に説明してきた。この手法の問題点としては, 1) 抽出した文中に代名詞などが含まれている場合, その先行詞が要約文中に存在する保証がないこと, 2) テキスト中の色々な箇所から抽出したものを単に集めているため, 抽出した複数の文間のつながり (首尾一貫性) が悪いことが指摘されている。

Paice[29] は, テキスト中のキーとなる重要な文を抽出するために用いられる表層的な情報の概説 (2 節) の後, 重要文抽出による要約生成の問題点として, 上のような点を指摘し, その問題点の解決法について述べている。照応詞を含む文の前の数文を要約に追加したり, 接続詞は削除したり, 動詞の時制や態は調和がとれた流れにしたりすることで, 部分的な解決が実現できることを示している (3 節)。また, テキストの構造の把握が要約文生成に重要であることを 4 節で述べている。

3 節で紹介した, 山本らのシステム GREEN では, 上に述べた問題点に対し, 重要文として抽出される文の先頭に指示語や接続詞が出現する場合, あるいは, 抽出される文の主語が省略されている場合, 単独で要約中に存在すると不自然であると考え, 前文も要約に追加することで対処している。

Mahesh[20] は, 要約過程を, テキストからの hypertext の構成過程ととらえ, 重要な箇所が前面にあり, そこからリンクをたどることで, より詳細な情報が段階的に得られるような枠組を提案している。従来の研究が (ある決まった要約率の) 要約と, (要約する前の) テキスト全体という 2 つの要素しか提示しなかったのに対し, ユーザが自分の関心に応じてリンクをたどることで, さまざまな要約率の要約を段階的に参照可能である枠組であり, 上の問題点に対しても, 要約の表示法で部分的な解決を試みているという点で, 興味深い。テキストからの重要文抽出は従来の手法を用いているが, その後処理として, 抽出した文を (部分的に) 構文解析し, 埋め込み文, 従属節などを削除することで, さらに要約文を短くすることを試みている。

自然言語処理の分野では、近年頑健な解析手法の開発が進んでいるが、これらの手法を用いた要約研究が今後数多く提案されるようになると思われる。2.4節で述べた「解析したテキスト構造を利用した」要約手法も、テキスト構造を解析する頑健な手法が開発されて初めて実現可能な手法であり、また、照応解析を利用した要約手法 [3] など、頑健な文脈処理を利用した要約手法が今後盛んに研究されることと思われる。頑健な文脈処理を利用した手法は、上に示した問題点の解決にも貢献できる可能性が高いと言える。この他にも、複合語 (“Bill Clinton”) を抽出しそれを利用する、また、固有名詞を抽出し、そのタイプ (人名, 場所, 会社名など) わけを利用するなどして、要約手法をこれまでの単語に基づく単純なものから、より詳細な情報に基づくものに拡張し精度向上を図る試み [1] も増えていくと思われる。また、山本ら, Mahesh が行なっているような、要約文中の不要箇所を削除し、さらに要約文を短くする処理には、頑健な (部分) 構文解析手法の利用が不可欠であると考えられる。

また、聴覚障害者のための、ニュース番組の字幕用要約生成など、具体的な応用を想定した要約研究も活発になりつつあり、注目される。

若尾ら [47] は、実際のニュース番組中の字幕を用いて、人手で作成されている字幕とニュース原稿を比較することで、字幕用の要約手法の分析を行なっている。要約手法として、表層の文字列の情報のみで可能な手法のみを取り上げ、5つに分類している。また、各手法の使用頻度、削減される文字数なども調査している。

最後に、本稿以外の過去の解説を紹介しておく。

[4] の 7.4 節に Spark Jones の簡単な解説がある。

Paice[29] も、対象が論文中心ではあるが、上に述べたように、これまでの手法の解説を含んでいる。

Information Processing & Management, 31(5), 1995 は自動要約 (Automatic Summarizing) の特集号である。本稿では述べなかった、要約の生成過程に関する研究として、3編の論文が収録されている。

[32] にも Alterman の解説がある。この解説は、対象が物語中心であり、領域知識を用いた手法に関してのみが説明されている。

謝辞

本稿をまとめるに当たっては、博士課程 2 年、望月源君、近藤恵子さん、徳田昌晃君の協力が大きな助けとなった。ここに記し、感謝します。

参考文献

- [1] C. Aone, M.E. Okurowski, J. Gorlinsky, and B. Larsen. A scalable summarization system using robust nlp. In *Proc. of the ACL Workshop on Intelligent Scalable Text Summarization*, pp. 66–73, 1997.
- [2] R. Barzilay and M. Elhadad. Using lexical chains for text summarization. In *Proc. of the ACL Workshop on Intelligent Scalable Text Summarization*, pp. 10–17, 1997.
- [3] B. Boguraev and C. Kennedy. Saliency-based content characterisation of text documents. In *Proc. of the ACL Workshop on Intelligent Scalable Text Summarization*, pp. 2–9, 1997.
- [4] R.A. Cole, J. Mariani, H. Uszkoreit, A. Zaenen, and V. Zue, editors. *Survey of the State of the Art in Human Language Technology*. <http://www.cse.ogi.edu/CSLU/HLTsurvey/>, 1996.
- [5] J. Cowie and W. Lehnert. Information extraction. *Communications of the ACM*, Vol. 39, No. 1, pp. 80–91, 1996.
- [6] H.P. Edmundson. New methods in automatic abstracting. *Journal of ACM*, Vol. 16, No. 2, pp. 264–285, 1969.
- [7] F. Fukumoto, Y. Suzuki, and J. Fukumoto. An automatic extraction of key paragraphs based on context dependency. In *Proc. of the 5th Conference on Applied Natural Language Processing*, pp. 291–298, 1997.
- [8] H. A. K. Halliday and R. Hassan. *Cohesion in English*. Longman, 1976.
- [9] T.F. Hand. A proposal for task-based evaluation of text summarization systems. In *Proc. of the ACL Workshop on Intelligent Scalable Text Summarization*, pp. 31–38, 1997.
- [10] M. Hoey. *Patterns of lexis in text*. Oxford University Press, 1991.
- [11] E. Hovy and C. Lin. Automated text summarization in SUMMARIST. In *Proc. of the ACL Workshop on Intelligent Scalable Text Summarization*, pp. 18–24, 1997.

- [12] D. Jang and S.H. Myaeng. Development of a document summarization system for effective information services. In *Proc. of the RIAO97*, pp. 101–111, 1997.
- [13] Quinlan J.R. *C4.5: Programs for Machine Learning*. Morgan Kaufman Publishers, 1993.
- [14] K. Kondo and M. Okumura. Summarization with dictionary-based paraphrasing. In *Proc. of the Natural Language Processing Pacific Rim Symposium'97*, pp. 649–652, 1997.
- [15] J. Kupiec, J. Pedersen, and Chen F. A trainable document summarizer. In *Proc. of the 18th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, pp. 68–73, 1995.
- [16] S. Kurohashi and M. Nagao. Automatic detection of discourse structure by checking surface information in sentences. In *Proc. of the 15th International Conference on Computational Linguistics*, pp. 1123–1127, 1994.
- [17] C. Lin. Knowledge-based automatic topic identification. In *Proc. of the 35th Annual Meeting of the Association for Computational Linguistics*, pp. 308–310, 1995.
- [18] C. Lin and E. Hovy. Identifying topics by position. In *Proc. of the 5th Conference on Applied Natural Language Processing*, pp. 283–290, 1997.
- [19] H.P. Luhn. The automatic creation of literature abstracts. *IBM Journal of Research and Development*, Vol. 2, No. 2, pp. 159–165, 1958.
- [20] K. Mahesh. Hypertext summary extraction for fast document browsing. In *Proc. of the AAAI Spring Symposium on Natural Language Processing for the World Wide Web*, 1997.
- [21] I. Mani and E. Bloedorn. Multi-document summarization by graph search and matching. In *Proc. of the 14th National Conference on Artificial Intelligence*, pp. 622–628, 1997.
- [22] D. Marcu. From discourse structures to text summaries. In *Proc. of the ACL Workshop on Intelligent Scalable Text Summarization*, pp. 82–88, 1997.

- [23] K. McKeown and D.R. Radev. Generationg summaries of multiple news articles. In *Proc. of the 18th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, pp. 74–81, 1995.
- [24] S. Miike, E. Itoh, K. Ono, and K. Sumita. A full-text retrieval system with a dynamic abstract generation function. In *Proc. of the 17th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, pp. 152–161, 1994.
- [25] J. Minel, S. Nugier, and G. Piat. How to appreciate the quality of automatic text summarization? examples of fan and mluce protocols and their results on seraphin. In *Proc. of the ACL Workshop on Intelligent Scalable Text Summarization*, pp. 25–30, 1997.
- [26] M. Mitra, A. Singhal, and C. Buckley. Automatic text summarization by paragraph extraction. In *Proc. of the ACL Workshop on Intelligent Scalable Text Summarization*, pp. 39–46, 1997.
- [27] J. Morris and G. Hirst. Lexical cohesion computed by thesaural relations as an indicator of the structure of text. *Computational Linguistics*, Vol. 17, No. 1, pp. 21–48, 1991.
- [28] R. Ochitani, Y. Nakao, and F. Nishino. Goal-directed approach for text summarization. In *Proc. of the ACL Workshop on Intelligent Scalable Text Summarization*, pp. 47–50, 1997.
- [29] C.D. Paice. Constructing literature abstracts by computer: Techniques and prospects. *Information Processing and Management*, Vol. 26, No. 1, pp. 171–186, 1990.
- [30] G. Salton. *Automatic Text Processing*. Addison-Wesley, 1989.
- [31] G. Salton, A. Singhal, C. Buckley, and M. Mitra. Automatic text decomposition using text segments and text themes. In *Proc. of the 7th ACM Conference on Hypertext*, pp. 53–65, 1996.
- [32] S.C. Shapiro, editor. *Encyclopedia of Artificial Intelligence*. John Wiley and Sons, Inc., 2nd edition, 1992.

- [33] E.F. Skorokhod'ko. Adaptive method of automatic abstracting and indexing. In *Information Processing 71*, pp. 1179–1182. North Holland, 1972.
- [34] K. Sumita, K. Ono, T. Chinoto, T. Ukita, and S. Amano. A discourse structure analyzer for japanese text. In *Proc. of International Conference of Fifth Generation Computer Systems*, pp. 1133–1140, 1992.
- [35] S. Teufel and M. Moens. Sentence extraction as a classification task. In *Proc. of the ACL Workshop on Intelligent Scalable Text Summarization*, pp. 58–65, 1997.
- [36] H. Watanabe. A method for abstracting newspaper articles by using surface clues. In *Proc. of the 16th International Conference on Computational Linguistics*, pp. 974–979, 1996.
- [37] K. Yamamoto, S. Masuyama, and S. Naito. An empirical study on summarizing multiple texts of japanese newspaper article. In *Proc. of the Natural Language Processing Pacific Rim Symposium'95*, pp. 461–466, 1995.
- [38] K. Zechner. Fast generation of abstracts from general domain text corpora by extracting relevant sentences. In *Proc. of the 16th International Conference on Computational Linguistics*, pp. 986–989, 1996.
- [39] 間瀬久雄, 大西昇, 杉江昇. 説明文の抄録作成について. 電子情報通信学会言語理解とコミュニケーション研究会報告, pp. 5–12, 1989. NLC89-40.
- [40] 亀田雅之. 疑似キーワード相関法による重要キーワードと重要文の抽出. 言語処理学会第2回年次大会発表論文集, pp. 97–100, 1996.
- [41] 亀田雅之. 段落間及び文間関連度を利用した段落シフト法に基づく重要文抽出. 情報処理学会自然言語処理研究会報告, pp. 119–126, 1997. 121-17.
- [42] 隅田英一郎, 飯田仁. 統計的な抄録法を使った情報検索. 言語処理学会第3回年次大会発表論文集, pp. 353–356, 1997.
- [43] 佐々木一朗, 増山繁, 内藤昭三. 語彙的結束性に着目した文章抄録法の提案. 情報処理学会自然言語処理研究会報告, pp. 65–72, 1993. 98-9.
- [44] 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.

- [45] 柴田昇吾, 上田隆也, 池田裕治. 複数文章の融合. 情報処理学会自然言語処理研究会報告, pp. 77-82, 1997. 120-12.
- [46] 若尾孝博, 江原暉将, 村木一至, 白井克彦. テレビニュース番組電子化原稿を題材とした自動要約手法の大規模評価. 情報処理学会自然言語処理研究会報告, pp. 31-36, 1997. 119-6.
- [47] 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. 情報処理学会自然言語処理研究会報告, pp. 83-89, 1997. 122-13.
- [48] 船坂貴浩, 山本和英, 増山繁. 冗長度削減による関連新聞記事の要約. 情報処理学会自然言語処理研究会報告, pp. 39-46, 1996. 114-7.
- [49] 仲尾由雄. 見出しを利用した新聞・レポートからのダイジェスト情報の抽出. 情報処理学会自然言語処理研究会報告, pp. 121-128, 1997. 117-17.
- [50] 難波英嗣, 奥村学. 観点に基づいた新聞記事の重要文抽出に関する心理実験と考察. 言語処理学会第4回年次大会併設ワークショップ「テキスト要約の現状と将来」論文集, 1998.
- [51] 難波英嗣, 奥村学. 論文間の参照情報を考慮した学術論文要約システムの開発. 言語処理学会第4回年次大会発表論文集, 1998. A6-3.
- [52] 福島俊一, 他. 日本語情報検索システム評価用テストコレクション BMIR-J1. 自然言語処理シンポジウム「大規模資源と自然言語処理」論文集, 1998.
- [53] 福本淳一. 文の結合度に基づく内容抽出手法. 言語処理学会第3回年次大会発表論文集, pp. 321-324, 1997.
- [54] 福本淳一, 安原宏. 日本語文章の構造化解析. 情報処理学会自然言語処理研究会報告, 1991. 85-11.
- [55] 望月源, 岩山真, 奥村学. 抄録を利用したテキスト検索. 言語処理学会第4回年次大会併設ワークショップ「テキスト要約の現状と将来」論文集, 1998.
- [56] 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. 情報処理学会自然言語処理研究会報告, pp. 71-76, 1997. 120-11.