要約品質向上のための文再構成知識の自動獲得

竹内和広、松本裕治
奈良先端科学技術大学院大学 情報科学研究科
〒 630-0101 奈良県 生駒市 高山町 8916-5
Tel. 0743-72-5246
E-mail: fkazuh-ta, matsug@is.aist-nara.ac.jp

自動要約では文章中の重要文を抽出するだけでなく，抽出した重要文を再構成し，わかりやすい要約とする必要がある．このような文を再構成する操作には様々なレベルが存在するが，本稿では，人間が作成した要約データにもとづいて，文の文節を削除して文を簡略化する文再構成操作に焦点を合わせ，どのような性質をもつ文節が要約中で省略されやすいかを検討した．具体的には，人間が要約で要約元文から削除をおこなった文節を特定し，機械学習の一手法である SVM(Support Vector Machines) を用いて文再構成知識を学習させた．

KEYWORDS: 自動要約，文の簡略化，機械学習，SVM

Acquisition of Sentence Simpli¯cation Rules for Improving Quality of Text Summaries

Kazuhiro Takeuchi and Yuji Matsumoto
Graduate School of Information Science, Nara Institute of Science and Technology
8916-5 Takamaya, Ikoma, Nara 630-0101, JAPAN
Tel.+81-743-72-5246
e-mail:fkazuh-ta, matsug@is.aist-nara.ac.jp

Sentence simpli¯cation is one of the methods for improving quality of text summaries. This paper focuses on the way to select adequate segments from the sentences in a document for generating simpler sentences. We applied support vector machines (SVM) for acquiring the natural method to select segments. For the learning, human-written summaries are used. As a result we con¯rm SVM has capability to acquire the knowledge to make a sentence simpler.

KEYWORDS: automated summarization, sentence reduction, machine learning, SVM

1　はじめに

近年の電子化文章の増大により，計算機を利用して人間の文章処理を支援する研究がなされるようになった．そのような研究のひとつに，計算機に文章を要約させる試みがある．

計算機による要約の試みでは，文章中の重要と思われる部分を抽出することを中心に研究されてきた．しかし，要約は人間の高度に知的な作業であるため，計算機により重要と認定された重要部分を列挙するだけでは，要約文章の結束性，構成などの点で課題があることが認識されてきている [1].

重要文の抽出を前提に要約の品質を向上するために，抽出された文の集合をいかに再構成するかが課題となる．しかし，その再構成には，結束性，構成などが適切で人間にとって読みやすく，要点を適正に網羅しなければならないといった要件があり，これらの要件を満たすためには，表層的な表現の書き換えや，構文的書き換えなど様々な処理のレベルが考えられる．

そのような現状に対し，我々は，どのような要約文章なら読みやすく，適正であるかを判断するためには，人間が実際にどのような要約を作成するかを調査した上で，計算機でも実現が可能なレベルの要約操作に細分化し，整理することが必要であると考える．

本稿ではこのような立場から，人間が一定の制限のもとで作成した要約事例を収集・分析し，一般的な自然言語処理として，自動要約処理に適用できると思われる文の再構成操作を機械学習によって計算機により実現することを試みた．機械学習には，人間が作成した要約データから機械学習に適当であると考えられる事例を自動抽出し，その事例に基づいて、機械学習の一手法である SVM(Support Vector Machines) を適用する実験を行った．

2　対象とする要約操作

2.1　自動要約の過程と文再構成操作

自動要約における数多くの関連研究では，大まかな処理過程を，テキスト中から重要だと思われる文や節を抽出する過程と，抽出した文や節を再構成し出力する要約生成過程との２つに分類することが多い．

情報検索システム評価用テストコレクション構築 (NTCIR) プロジェクトのワークショップにおける自動要約タスクに参加するために実装した我々の自動要約システムでも，自動要約の過程は以下のような２つの過程に整理した [2].

- 重要文の抽出　→　要約文として再構成

我々の研究の最終的な目的は，計算機によって文章を読みやすく要約することである．我々の要約システムでは，重要文の抽出の処理過程は，文章における重要な情報を選択することに主眼をおき，それらの文を要約文として再構成する過程では，文章を読みやすくする役割を実現するように設計した．以下，本論文では特に，重要文抽出過程で抽出されたそれぞれの文を要約文として再構成する処理に研究の焦点を当てる．

重要文抽出に基づいて自然な要約生成を実現させるためには，例えば，難波ら [3] の研究のように，抽出した重要部分の集合を人間が読みやすいように適当に書き換えてやる必要性が指摘されている．しかし，残念ながら計算機で行える要約作成の操作には限界があるため，人間の行っている要約作成操作を，十分に細分化して整理し，少なくとも現状の計算機には不可能と考えられる高度な世界知識を必要とする操作を避けつつ，一定水準を満たす要約作成に必要な手順を明確にする必要がある．

人間が行う多様な要約処理を，計算機でも実現可能な要約作成操作として分析する前提の一つとして，人間が元テキストのどの部分を用いて，要約中の各文を作成したのかを対応付け，それを分析することは有益である．

対応付けの前提として，要約中のある文を作成するために，元記事中で要約に用いられる文は，重要文抽出の過程で得られたれるものと仮定し，それらの文を用いて，どのように要約が再構成するかを視点に分類してみると，以下のようなものが考えられる．

- 操作 1: 元記事の１文　→　要約中の１文
- 操作 2: 元記事の複数文　→　要約中の１文
- 操作 3: 元記事の１文　→　要約中の複数文

もちろん，要約は複数の文から成り立つのが普通であるから，一つの要約中でも操作１から３のすべてが使われる可能性が高い．さらに，この他にも，要約作成者が元記事を理解し新たに要約を作成するもので，深いレベルの意味理解を考慮しないと記事中のどの部分から要約が作成されたかわからないものが考えられる．

本稿では，まず，実際に人間が作成した要約を上の操作の観点から分析し，計算機によって要約する際の操作として，どの程度の文再構成を仮定することが適当かを検討する．

表 1: まとめあげのタイプ分類

主題化：	まとめあげられる文中の 1 文の要素がまとめあげられた文の主題となる.
元文 1	協定の効果を確実にするのが紛争処理案だ。
元文 2	紛争処理機構を設立したうえで、国家間紛争の処理を OECD での調停を含めた協議の場にとどめる案と、最終的な拘束力を持つ仲裁権限をもつ組織にする案を検討する。
要約例	紛争処理案は紛争処理機構を設立し、国家間紛争の処理を調停にとどめる案と、拘束力を待たせる案を検討する
連体化：	同一対象への言及がある文がまとめられ、まとめあげられる文の 1 つがまとめあげられた文の中で対象の修飾を行う。
元文 1	野洲工場に生産ラインを導入、 4 月から量産する
元文 2	野洲工場の半導体製造に使っている建屋を活用、カラーフィルター用のクリーンルームなどを設けた。
要約例	生産ラインを導入する野洲工場の半導体製造に使っている建屋を活用し、カラーフィルタ用のクリーンルームなどを設ける。
複文化：	連体修飾節以外の形をとる複文化。便宜上、重文も含める。
元文 1	冷戦時代を通じてインドは旧ソ連から武器の大半を調達するなど、親密な関係だった。
元文 2	これに対して、米国はパキスタンに軍事援助を進めていた。
要約例	冷戦時代インドは旧ソ連と親密な関係だったが、米国はパキスタンに軍事援助を進めていた。
その他：	人間が高度な書き換えを行っていると考えられ、上の分類から除外したもの。
元文 1	内国民待遇加盟国は、約束表に記載したサービス分野について他の加盟国のサービス提供者に対し、内国民待遇を与える義務がある。
元文 2	すなわち、その分野では他の加盟国のサービス提供者に対し、自国の同種のサービス提供者に与える待遇より不利でない待遇を与えることを盛り込んでいる。
要約例	また、内国民待遇加盟国は、他の加盟国のサービス提供者に対して、自国産業と同等の待遇を与えるサービス分野を約束表に記載する。

2.2 要約操作に関する予備調査

要約における文再構成操作を分析するために，既に我々は，2 人の被験者に文や節での重要部分抽出を前提としない要約を作成してもらい (各被験者 32 記事)，自動要約を行う上でどのような操作が必要かを分析する予備調査を行った [4]．ここでは，その予備実験の結果を手短に紹介し，その結果から，本研究でどのような文再構成操作を仮定するかを述べる．

要約は 1 つの記事に対して，元記事の文字数の 40% までの量に要約してもらったものである．要約の指示として,「 全体のあらすじと著者の主な主張がわかるように要約する」,「 固有名詞はできるだけ原文の表現を用いる」という 2 つの制約を課した．なお，題名と形式段落は要約作成者が参照できないように元記事からこれらを取り除いて実験を行った.

要約率を 40%と高率にした理由は，文章を短くする時にしばしば必要な極端な言い換えを抑え，元記事から要約文がどのように作成されたか，その操作が分析しやすくなると考えたからである．実際，両要約作成者がほぼ指示通りに，記事を要約したことから，高度な言い換えと思われる例が少く，固有名詞を手がかりに元記事のどの文から要約中の各文が生成されているかを容易に見極めることができた．また，この分析では，人手で元テキストと要約文との対応付けを行った.

今回の予備実験で用いた要約は，ある程度の制約はあるものの，ほぼ自由要約と考えてよい．そのため，要約作成者間の重要部分選択に対してのゆれは非常に大きかったが，以下の 2 つの点においては両要約作成者に共通して見られる傾向があった.

一点目は，元記事で先に出現する文で述べられた情報が，要約中でも先に述べられる点である．我々の研究に関連する先行研究として，望主ら [5] らは，1 記事に対して 100 人分の要約事例を 3 記事分用意し，本研究と同様に文節レベルの対応付けを行う実験を行っている．その実験の結果から，元テキスト中に先に現れる構成要素は，要約中でも先に現れる線形的な順序付けの傾向が報告されている．本研究でも彼らの研究同様，要約には線形的な順序付けの傾向がみられている.

二点目は，要約元記事と比較して，要約では主題文が多い点である．これは，要約では，元記事に表れる主題と副主題を網羅的に簡潔に述べようとすることと関係があると思われる．元記事中での主題や副主題は，元記事中の 1 文だけで述べられていることもあるが，むしろ，元記事では主題・副主題に関連する情報を 1 文のみで叙述するのではなく，複数の文を用いて叙述されることが多い．すなわち，要約を作成する際，2.1 の操作 1 のような，元記事中の 1 文をそのまま，もしくは短くして述べられることも多いが，このような複数文を要約する場合，2.1 で分類した操作 2 を用いて，元記事中の複数文をまとめあげ，要約中の 1 文で述べることも行われている.

我々はこのような元記事中の複数文を要約中の 1 文にまとめあげる例にどのような類型があるかを整理した．その際の要約操作の分類と，それぞれの操作の例を表 1 に示す.

表 1 のまとめあげを行った際の主な観点は，要約元

記事中の主題・副主題が要約中でどのように扱われているかである．例えば，表 1 の複文化の例で示した要約文は，まとめあげ元文の各主題がまとめあげられた文でも従属節の主題と主節の主題で表現されている形である．

元文中で主題・副主題に関連する情報を複数文でまとめ上げた，次のような複文化の例も多数存在した．

元文 1　百貨店は売上不振が続いているが、初年度四百三十億円の売上高を見込んでいる。
元文 2　物流部門を外部委託する他、パート・アルバイトの比率を高めるなど、店舗運営の効率化を徹底する。
要約　百貨店は売上不振が続いているが、初年度４百三十億円の売上を見込み、店舗運営の効率化を徹底する。

この例は、まとめあげ元文の一方の主題がまとめあげられた要約文でも主題となり、もう一方の元文にはそもそも主題も主格主語も存在しない例である。

また，表 1 の主題化の例は，要約元文中で提題表現で扱われていない要素が提題化される事例を示したが，頻度的には，その例とは別の以下のタイプのものが多い．それは，例えば、上に示した複文化の事例事例をさらに短くした例である．

百貨店は、店舗運営の効率化を徹底する。

このタイプの主題化まとめあげは．複文化操作の特別な形と考えることもできる．

なお，予備調査では，表 1 に示した,「その他」のまとめあげに相当する事例は要約率を 40%としたことから数が少なく，本研究で対象とする要約操作からはこの操作は除外する．

以上のようなまとめあげの事例を眺めると，元記事中の文をどのように再構成する場合でも，要約に使用するとした文のいくつかの文節を削除して，文を簡略化していることがわかる．本論文では，この文から不要な文節を削除して，文を再構成する操作を，2.1 で分類した，すべての操作に共通する一般的な文の再構成操作としてとらえ，いかこの操作を計算機で実現するためには，どのような知識が必要かを検討する．

3　対応付け事例の自動抽出

Jing ら [6] や加藤ら [7] の研究では要約事例と要約元テキストの対から要約に用いられる知識を機械学習によって獲得する行う手法を提案している．我々の研究でも，機械学習を用いて文，学習事例から再構成操作に関する知識を自動的に獲得するアプローチを選んだ．

その際，学習に用いる学習事例を，予備調査で行ったような人手による対応づけによって抽出していたのでは，作業コストがかかりすぎる．そのため，人間が作成した要約データから，自動的に自動要約の研究に役立つ事例のみを抽出することが必要となる．このような研究流れの中に，例えば，Marcu [8] のコサイン類似度を用いて，要約文の各文が，要約元テキストのどの部分から作成されたかを自動的に推定する手法がある．本研究においても，学習の対象となるのは，どのような特徴をもった要約元記事中の文節が，要約中で用いられないかを判断する知識であり，人間が要約した文が要約元記事のどの文から作成されたかを対応づけた，要約文と要約元文との組が，学習事例となる．本研究ではこの様な学習事例を自動的に同定した．

本研究が対象とする，学習事例を抽出する要約データは，3 人の作業者に社説 90 記事に対して文字数で元記事の約 40 %の長さになる要約を作成してもらったものである（全 270 要約）．このデータは，2.2 で紹介した予備調査の後，さらに大規模に要約データを収集するため，予備調査と同じ条件を用いて 3 人の作業者が新聞の社説を要約する作業を続けてきた結果得られたものである．データ数は，90 記事 3547 文を 3 人の要約作成者が要約した，合計 2467 文 (763 文,773 文,931 文) である．要約元記事の 1 記事平均の文数は 39.4 文であり，それに対する要約の平均文数は 9.1 文となる．

このデータから要約事例を作成するための対応付けを自動的に抽出するため，まず，1 文から不要な文節を削除して短縮化する操作について、節レベルに対する操作と文節レベルに対する操作に分け、それぞれ以下の操作を仮定する．

節レベル	削除	許可
	順序入替え	許可
	新規挿入	不許可
文節レベル	削除	許可
	順序の入替え	許可
	新規挿入	特定の種類のみ許可
	書換え	特定の種類のみ許可
	それ以外	不許可

ここで，文節の新規挿入は，接続表現，副詞表現，呼応表現のみに限定した．また，文節内の名詞，動詞といった自立語の書き換えについては，複合名詞が短縮された文節など，文節内の情報のうちいくつかの単語が保存されているもののみに限定した．他方，助詞，助動詞，活用語尾などについてはゆるい制限で書き換えが可能であるとし，削除することも認めた．なお，今回は節レベルの操作の後に文節レベルの操作が行われると仮定した．

要約データより，この上の制約に基づいた以下のような手順を用いて，要約中の文とそれを作成するために用いられた文との対応付けを行い，そのような対応付けの中で，確実に，文節削除により，要約が作成されたと考えられる事例のみを文再構成知識を獲得するための学習事例とする．

1. 要約元記事中の文 O_i に対して，その記事から作成された要約の中で，文字ベースのコサイン類似度がもっとも高い文 S_i を抽出する．
2. 上で仮定した節レベル，文節レベルの書き換えに対する制約にもとづいて，O_i 中の文節のうち，S_i で用いられている文節を選びだす．O_i の文末述語 $pred_i$ が S_i で用いられていなかったり，$pred_i$ の文末述語以外の文節が S_i 上で $pred_i$ に対応付けられた文節よりも右に出現した場合は抽出事例から除外する．
3. O_i の文節数と S_i の文節数の比で閾値を設け，その閾値を調整することによりあまりにも多くの文節が削除されるような対応付けを事例として抽出することは避け，確信高く，O_i の文節が削除され，S_i になったと考えられるもののみを抽出する．

この手順により，1057 例の要約文とその元文の対を要約事例として抽出した．要約を行う操作のうち，最も基本的であるのは，要約元記事のある文を直接抽出し，そのまま要約で用いる操作であるが，しかしこのような操作は要約のべ 2467 文中 779 文しか存在せず，その他多数の要約文は人間が何らかの文の再構成・生成操作を行って作成されている．本節で説明した手順により抽出された 1057 事例は，そのような要約元文をそのまま要約でも用いるような操作ではなく，元記事中の文の文節をいくつか削除することにより簡略化し，それを要約でもちいる操作を計算機で実現するための足掛かりとなる．

なお，この要約元文それに対応する要約文の対 1057 事例に基づいて、要約元文から削除される文節数は 2942 である（抽出され要約事例の要約元文の全文節数は 10856）．

4　文再構成知識の自動獲得

前節で抽出した，データを用いて，要約作成操作に関する知識を機械学習する実験を行う．

4.1　学習手法

計算機に自動要約を行わせるという我々の最終目的を考えると，削除文節の特徴を人間が個別に分析した結果をルール化するのではなく，手本となるデータを機械学習させ，計算機が新たな要約を作成する際にその学習結果が妥当に反映されるようになるのがよい．この学習手法として我々は Vapnik [9] によって提唱された Support Vector Machines (SVM) を採用した．

機械学習において，事例 x_i は，その事例を表現する特徴ベクトル f と，その事例が分類されるクラス c の組 $< f; c >$ によって記述する．SVM は上記のような組で表現された事例から各クラスに分類される事例の特徴を統計的に一般化された分類器を獲得する（学習する）手法である．この分類器を使用することにより，未知のデータを，その特徴を与える f で表現してやれば，そのデータが属するクラスを予想することができる．SVM は，特徴ベクトル f が高次元で表現されていても，学習に用いた事例に特化した分類器を避け，一般的な分類器を獲得できるとされている．

本実験の場合，分類する事例は元記事中の文に存在する文節 b_i であり，b_i が分類されるクラスは，b_i が要約中で用いられる, 用いられないの 2 値とした．

4.2　素性の整理

文節を特徴付けるベクトル f は，文節がある特定の特徴を表した素性の n 個の組

$$< a_1; a_2; ¢¢¢; a_n >$$

で表現する．すなわち，文節の特徴づける素性の数はベクトルの次元数 n に相当する．本実験で使用した，素性は，当該の文節が, その素性のあらわす性質を持つか否かの２値の値をとる素性である．文節の特徴をこのような，複数の素性で表現できるように整理することは，機械学習において本質的に重要な作業である．

SVM は先に述べた通り，素性による特徴ベクトルが高次元であっても，過学習を起こしにくい特徴がある．本実験で採用した文節の特徴付けは，この SVM の特徴を生かし，文節を表現する特徴ベクトルの次元数は 2000 次元を越える．そのため，すべての素性を逐一表記し, 説明することは難しいため，どのような観点から，文節 b_i のを特徴づける上で素性を選択・整理したかを述べる．今回の実験で文節 b_i を特徴付ける上で用いた観点は以下のようなものである．

- 観点 1: b_i の意味的主辞による分類
- 観点 2: 文節 b_i の末尾の付属語による分類
- 観点 3: b_i に含まれる語の頻度情報
- 観点 4: 文節 b_i の係り受け構造上の位置
- 観点 5: 文節 b_i が修飾する文節の情報

文節の特徴ベクトルを決定する前提として，文内の統語構造を考慮することが重要であると考え，要約元文の文節の修飾関係は，係り受け解析ソフトウェア [10] によって，あらかじめ解析している．

以下，それぞれの観点を用いて，具体的にどのような形の素性に整理したかを説明する．

観点 1

文節の意味的な主辞情報については，文節内の自立語に基づいて決定する．動詞，形容詞などの用言については，各語彙をそれぞれを独立した素性として用いた．例えば，ある文節に，「絡む」「下ろす」「慣れる」といった語が含まれるどうかの 2 値をもつ素性としてそれぞれ表す．接続詞，代名詞，副詞，連体詞についても，語彙をそのままそれぞれを独立した素性とする．

名詞については，語彙をそのまま使うと，あまりにも多様になりすぎてしまうと考え，名詞の接続形，固有名詞か否かなどに基づいて特徴化した素性に分類した．

観点 2

文節の統語的性質を特徴づける情報として整理する．助詞については，語彙をそのまま独立した素性として扱う．用言については，用言の活用形，それに接続する付属語や接尾辞の「動詞-自立-基本形:だろ-助動詞-未然形:う-助動詞-基本形」といった品詞分類の並びで抽象化した素性で表す．サ変接続の名詞に動詞がつく場合は，動詞の語彙も考慮した．

複文を形成する時など，文節末に読点「、」がつく場合も，接続助詞などの素性だけでなく，文節末に読点がつくかいなかも素性としている．

観点 3

文節 b_i が，文末文節を 0 とする係り受け構造を木構造で表現した木の深さがどの程度にあるかを素性とした．また，この構造木を用いて，文節 b_i が，どの文節からも修飾されない葉であるか，いくつの文節から修飾されているか，直接すぐ右の文節を修飾するか否かも素性とした．

観点 4

語の頻度情報は，観点 1 で整理した文節の意味的な特徴を補足する役割がある．特に，名詞を意味的な主辞とする文節を特徴づけることが，この観点を導入した目的である．素性としては，要約元記事中のすべての名詞について出現頻度をとっておき，文節内の名詞の最も出現頻度大きいものを代表させ，文節の語頻度情報とした．この情報は，特定の名詞の当該文章における重要性を反映させる意味ももつ．

観点 5

文節 b_i がどのような文節を修飾しているかを，修飾先の文節の観点 1 と 2 で整理した属性を用いて表現した．文節 b_i を修飾する文節の特徴を素性として用いなかった理由は，文節 b_i を修飾する文節が 1 つとは限らず，複数存在するため，情報を表現する適当な方法を考えつかなかったからであり，現状では，観点 3 で整理した，文節 b_i をどれだけの文節が修飾するかで代替している．

5 結果と考察

5.1 要約率と正解率

前節で説明した特徴に基づき 1057 例の文簡略事例中の各文節を特徴づけし SVM を用いて学習した分類器を得た．

学習で得られた分類器に要約前の文を入力すれば，獲得された要約知識に基づいて文中から不要文節を削除し，文簡略したデータを作成することができる．この分類器の評価として，学習に用いた事例を 10 分割し，交差検定を行った時の，文節位での平均要約率と精度を以下の表 2 に示す．正解率は以下の式で計算したものをパーセント表示で示した．

$$\text{正解率} = \frac{\text{データの文節の削除判断と一致した数}}{\text{データの文節の数}}$$

この結果から，今回整理した特徴を用いることにより，精度 77.4%程度であることは，要約を作成したす

表 2: 要約率と正解率

要約率	正解率
76.5 %	77.4 %

べての被験者がある元文に対して全く同じ要約文を生成するわけでないことに象徴されるように，人間がどのような要約を良いとするかは不透明であり，この精度はあくまでも参考程度の指標である．また，この交差検定は，3 節で説明した手法により，要約データより自動的に機械学習に適しているとして自動的に抽出した事例に対して行ったものであり，要約元記事中の任意の文に対して予想される要約率・正解率ではない．そのため，得られた分類器が持つ性質を実際に人間が見て評価することが重要であり，獲得した文節分類器に，学習させたデータ以外の文を入力して得た文簡略データをもとに，獲得させた要約知識がどのようなものであるかを分析する必要がある．

5.2 NTCIR-2 要約データとの比較

本研究が使用した要約データと同様に，自動要約研究を発展させる言語資源としての基礎データとして国立情報学研究が公開する NTCIR-2 要約データ[1]がある．NTCIR-2 の要約データでは，いくつかの形式の要約データを公開しているが，人間が重要個所を選択する形で得られた形式の要約データがある．

ここでは，この要約がなされた同じ元記事を，我々が機械学習によって獲得した文節選択分類器に入力し，その実験結果を NTCIR-2 要約データと比較する．比較に用いるのは，この NTCIR-2 要約データのうち，毎日新聞 95 年の 60 記事に対して作成された，要約率 40%及び 20%の計 120 要約である（以下，本稿では，この比較対照のみを NTCIR 要約と呼ぶ）．NTCIR 要約の対象となった文を機械学習された分類器によって文中の文節を削除した文再構成の例を表 3 に示す．分類器によって削除された文節は＜＞によって表した．

なお，今回学習に利用した学習事例は，文から平均 2.7 文節程度を落としている事例であり，獲得された分類器もそのような学習事例に対して 1 文あたり 2.4 文節程度を削除するものである．そのため，この分類器を未知の一般の文に適用した場合に，文から全く文節を削除しない場合もあり，NTCIR 要約と定量的に比較することには課題が残るため，本稿では，表 3 に比較の対象として当該文に対する NTCIR 要約を示すにとどめた．

表 3 のすべての例から言えるように，我々の分類器では，接続詞，副詞，文末の定型表現，特定の接続助詞等の特徴をもつ文節を削除する要約知識が獲得できていることがわかる．

文として原文の意味が通らなくなってしまう例は，(6)(7) などがあり， 複数文節が連続して削除されるような文節の削除については今回整理した文節の特徴付けに，文再構成知識を獲得する上で改良の余地があることが認められる．

また，すべての例にいえることであるが，文節の削除には文脈依存の部分があり，例えば，(2) の例の「金原学芸員は」の削除など，先行する文脈情報も考慮に入れ文節を特徴付ける必要がある．

このように文節の特徴付けは，今後，改善の余地はあるが，人間による要約作成における削除文節に一定の性質が存在し，複数の人間が作成した要約文章を学習データとして機械学習を適用することにより，どのような文節を削除すればよいかという知識を獲得可能であることが示された．

6 まとめと今後の課題

本稿では，計算機による自動要約作成を実現する基盤作りの一環として，人間の作成した要約文章を学習データとして，文を再構成する操作の一つである不要文節削除に関する知識を機械学習させる実験を行なった．その際，各文節の諸種の特徴づけを利用することによって，一定水準の機械学習が実際に可能であることを確認した．

今後の課題としては，複数文節の削除を可能にすること，文脈情報を素性として導入することが挙げられる．複数文節を削除する知識を獲得するためには，本稿で要約事例を抽出した対応付け手順を用いるのではなく，複数文節の削除のみを抽出するような手順を用いて学習事例を抽出し，複数文節の削除のみを判断するような分類器を学習させることが方向性として挙げられる．

文脈情報の素性としては，文間の連接関係，当該文の文脈における中心要素の表現などを考えてゆきたい．また, ある特定の要約誤りだけに注目して，その誤りを修正し，ある特定の要約知識の獲得に特化した要約データを蓄積してゆくことも有益であろう．

なお，本稿では，時間的制約から NTCIR-2 で公開されたデータを学習データとして用いていないなど，

[1] URL http://research.nii.ac.jp/ntcir/index-ja.html

表 3: 自動文再構成した事例と，NTCIR 要約

(1)	分類器の要約	＜それ以降、＞ 県警は 二つの 「アジト」を 二十四時間監視下に 置いた。
	NTCIR 要約	二つの アジトを 監視下に 置いた。
(2)	分類器の要約	＜金原学芸員は＞「役人は ＜せっかくの＞ 地元特産の 味に なじめず、 都の 食べ物を 持参したり 送らせたりして 食べていたらしい」と 話している。
	NTCIR 要約	役人は せっかくの 地元特産の 味に なじめず、 都の 食べ物を 持参したり 送らせたりして 食べていたらしい。
(3)	分類器の要約	＜しかし、＞ 東京、 大阪の 春の 知事選に みられた 有権者の 政党不信は ＜なお＞ 強く、 政界液状化は とどまりそうに ない。
	NTCIR 要約	しかし、 有権者の 政党不信は なお 強く、 政界液状化は とどまりそうに ない。
(4)	分類器の要約	＜逆に、＞ 秋田城跡の トイレ遺構から 見つかった 寄生虫の 卵は、 藤原京や 平城京で 見つかった トイレ遺構の 分析結果と ＜ピタリと＞ 一致。
	NTCIR 要約	秋田城跡の トイレ遺構から 見つかった 寄生虫の 卵は、 藤原京や 平城京で 見つかった トイレ遺構の 分析結果と ピタリと 一致。
(5)	分類器の要約	捜索は、 ＜東京都内の＞ ＜教団施設や＞ ＜静岡県富士宮市の＞ 富士山総本部でも 行われた。
	NTCIR 要約	捜索は、 東京都内の 教団施設や 富士山総本部でも 行われた。
(6)	分類器の要約	一九九三年 十月、 国と 都の 課長ら 十七人が 出席したはずの 港区赤坂の かっぽうは 「客の ことは 言えないが、 ＜ウチの＞ 座敷は 六畳一間だけ。 ＜一度に＞ ＜十七人なんて、＞ ＜とても＞ 入れない」と いぶかる。
	NTCIR 要約	十七人が 出席した はずの かっぽうは 「ウチの 座敷は 六畳一間だけ。 十七人なんて、 とても 入れない」と いぶかる。
(7)	分類器の要約	＜今後、＞ 日蓮宗系の 全国 十七の 寺で ＜オウムについて＞ ＜相談に＞ 応じていく。
	NTCIR 要約	日蓮宗系の 十七寺でも 相談に 応じていく。

このデータに関して詳細な分析を行ったとはいえない．今後，NTCIR-2 のデータを有効に活用してゆくことも課題である．

謝辞

本研究のでは分析データとして毎日新聞の記事を使用させていただいた．毎日新聞社に謝意を表す．また，NTCIR-2 要約データの使用許諾をいただいた国立情報学研究所および，NTCIR-2 要約データの公開に御尽力された方々に対して感謝します．

参考文献

[1] 奥村学, 難波英嗣．テキスト自動要約に関する最近の話題．Technical report, 北陸先端科学技術大学院大学情報科学研究科, 2000.

[2] T.Hirao, M.Hatayama, S.Yamada, and K.Takeuchi. Text Summarization based on Hanning Window and Dependency Structure Analysis. In Proc. of the Second NTCIR Workshop Meeting on Evaluation of Chinese & Japanese Text Retrieval and Text Summarization, National Institute of Informatics, 2001.

[3] 難波英嗣，奥村学．書き換えによる抄録の読みやすさの向上．情報処理学会研究報告 (自然言語処理研究会）, 99-NL-133, 1999.

[4] 竹内和広，松本裕治．テキスト構造に基づく要約生成制約条件の検討．情報処理学会研究報告 (自然言語処理研究会）, 00-NL-138, 2000.

[5] 望主雅子他．重要文と要約の際に基づく要約手法の調査．情報処理学会研究報告 (自然言語処理研究会）, 00-NL-135, 2000.

[6] H.Jing and K.R.McKeown. The Decomposition of Human-Written Summary Sentences. In SIGIR'99 *Proc. of the 22nd International Conference on Research and Development in Information Retrieval, 1999.

[7] 加藤直人，浦谷則好．局所的要約知識の自動獲得手法．自然言語処理, Vol.6 No.7, 1999.

[8] D.Marcu. The automatic construction of large-scale corpora for summarization research. In SIGIR'99 *Proc. of the 22nd International Conference on Research and Development in Information Retrieval, 1999.

[9] Vladimir N. Vapnik. Statistical Learning Theory. John Wiley & Sons, 1998.

[10] 工藤拓，松本裕治．チャンキングの段階適用による係り受け解析．情報処理学会研究報告 (自然言語処理研究会）, 01-NL-142, 2001.