日本語文章における表層的機械処理の有効利用に関する研究

1996年1月

博士（工学）

山本和英

豊橋技術科学大学

①

日本語文章における表層的機械処理の有効利用に関する研究

1996 年 1 月

博士 (工学)

山本 和英

豊橋技術科学大学

論文要旨

自然言語処理研究の一つの傾向として，処理の対象が単一の文から複数文へと徐々に移りつつあることが挙げられる．複数の文を対象にする談話処理の分野では，以前から文章の抄録／要約などの試みがなされてきたが，最近では，計算機及び機械可読文書の普及に伴うテキストの分類などの自動検索，あるいは対話処理などを対象にして，様々な研究が活発に行われている．本論文では今後ますます重要になると考えられる自然言語処理を対象にして研究活動を行ってきた．その中でも，従来ほとんど研究が行われていない段落分け，要約，文章の自動分類などの談話処理を取り上げ，検討を行った．また，対象言語としては日本語を取り扱った．

文章の段落分けは，段落のない文章 (文の羅列) を入力とし，段落に分けた文章を出力するという処理である．本論文では，段落分けに必要な要素として接続的語句と単語間の類縁性の二点を取り上げ，最終的には両要素を併用して考慮することによって段落分けを試みた．この手法で雑誌記事，新聞コラムの文章を段落分けした結果について人手による結果と比較し，作成した手法の有効性を支持する結果を得た．

文章の要約作成については，文章中から重要な文を単に抽出する「抄録」のように前後の文に全く結束性のない文章が生成されることを避け，要約文として出力された文章が最小限の結束性を保つように文の抽出を行った．さらに抽出した文に含まれる連体修飾語の一部を削除すること

により, 一段の文短縮を試みた. また, これらのシステムを実際に計算機上に実現することが重要と考え, 実在の論説記事 (新聞社説) を対象にした要約システムを構築した.

さらに, 従来ほとんど試みられていない, 何らかのつながりを持った複数新聞記事の要約も試みた. ここでは, 類似した複数文章の要約に特有の問題である類似記述部分の特定, およびその削除を試み, ある程度重複した記述の省略, 同じ語に対する修飾語句の削減などの点に着目することにより, 文章の短縮化を試みた.

本研究はまた, 文章の自動分類の研究にも取り組んだ. この研究では, 従来数多く研究が行われているベクトルモデルの分類手法のうち, 語に重みづけを行うこと (*term-weighting*) による手法を拡張し, 語をいくつかのグループに分類した上で重みづけを行う手法を提案した. この手法によって新聞の 10 種類のコラムを分類する実験を行った結果, 再現率, 適合率ともに高い結果が得られた.

Abstract

Studies on Automatic Non-Parsing Processing for Japanese Machine-Readable Texts

In the field of natural language processing (NLP), a large number of researches have been carried out in wide areas including related fields of NLP. The main subject of the research is gradually changing recently from one-sentence processing to that of plural-sentences. In the field of discourse processing, or plural-sentences processing, some researches on abstracting / summarizing sentences were made in the past, and in addition to them, there are various active researches at present, such as automatic document retrieval or dialogue processing.

Researches in this thesis belong to discourse processing in the area of NLP. Among several topics in discourse processing, we focus on topics of paragraphing, summarization and classification of documents, each of which is an important process that should be realized.

Discourse processing requires texts to be analyzed morphologically, syntactically and semantically in general. After technologies of these analyses have been advanced, discourse processing should be a research target, because these processes cause bad effects to discourse processing if their results are inaccurate. However, we dared to carry out researches

of some topics in the area of discourse processing with only morphological analysis. This is because some processes in discourse processing are feasible without syntactic or semantic processing. Another reason is that we need to deal with these problems on a long-term basis, because they are thought to have more difficult matters to advance, compared with the case of one-sentence processing. We also have a background that some processes like text summarization are so high and increasing in demand that their realization is strongly desired.

First this thesis explores the cohesion in Japanese sentences. When we analyze a text, a problem of how strong each sentence in the text connects each other arises. In this thesis, we focus on connectives and lexical cohesion which mainly affect cohesion in the text. First we inspect how strong connectives affects cohesion using paragraphing. We define cue words by investigating the beginning part of each sentence. Naturalness of paragraphing by cue words and length of sentences is evaluated by coincidence rate to the originals and consequence of a questionnaire. Then we attempt to paragraph using lexical cohesion. We use a thesaurus as the major knowledge-base for elucidating lexical structure in the text. We propose a graph model of cohesion structure of a text, called a cohesion graph, and formalize an evaluation function of paragraphing based on the cohesion graph. Texts are paragraphed by considering cue words and lexical cohesion using three combined methods. The experiments show that, compared with using lexical cohesion alone, the coincidence rate is improved. The consequence of the questionnaire also indicates that the proposed evaluation function is natural.

Then this thesis describes an experimental system, named GREEN, for summarizing Japanese texts. Analyzing several aspects of discourse phenomena is in general indispensable for making summaries of good quality because each linguistic phenomenon mutually affects in a complex manner. From this point of view, we have developed a system GREEN, which is capable of summarizing Japanese editorial newspaper articles. This system uses several surface linguistic characteristics appeared in Japanese texts. And in the proposed method, the sentences with some cohesion are choosed to avoid having little cohesion between the sentences in the output text, like the 'abstract.' The method proposed in this thesis not only chooses but shortens selected sentences by deleting three kinds of modifiers. We evaluate the performance of the system using a method of questionnaire where human subjects evaluate the quality of the summaries generated by the system.

We also attempt to summarize multiple Japanese articles into one document with non-parsing approach. We can hardly find studies on summarization of multiple texts. One of the reason is that the task is believed to be a next step of single text summarization. However, the processing of the multiple text summarization has some advantages to realize, one of which is that there are overlapped areas to delete if input texts are related each other. Thus we delete the overlapped parts among the input texts. Japanese grammar is considerably free in word order and allows high abridgement. This research also aims at coping with these phenomena. This thesis focuses on the following three points to sum up: same(or similar) clauses, noun modifiers, and change in wording.

We have implemented a prototype system of summarization and had some experiments.

A new method for automatic text classification is also proposed in this thesis by using class-weighting, or group-of-term-weighting approach to consider changes in wording. As a measure of importance, we introduce the product of the class frequency and the inverse document frequency (of class). We use a thesaurus to group all the terms appeared in the texts by their meaning. The result of open-test experiments is reported against Japanese columns which clarifies the efficacy of the proposed classification method.

発表論文リスト

A．学術論文

1. 山本和英, 増山繁, 内藤昭三：
   段落分けを用いた日本語文章における結束構造の検討,
   情報処理学会論文誌 Vol. 35, No. 10, pp.2029-2037
   (1994 年 10 月).
   (2章に対応)

2. 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 年 1 月).
   (3章に対応)

3. 山本和英, 増山繁, 内藤昭三：
   関連テキストを利用した重複表現削減による要約,
   電子情報通信学会論文誌 D-II (1996 年 2 月 21 日再提出).
   (4章に対応)

B．国際会議

1. Kazuhide YAMAMOTO, Shigeru MASUYAMA and Shozo NAITO: An Empirical Study on Summarizing Multiple Texts of Japanese Newspaper Articles.
   Proc. of Third Natural Language Processing Pacific-Rim Symposium, pp. 461-466 (Dec. 1995).
   (4章に対応)

2. Kazuhide YAMAMOTO, Shigeru MASUYAMA and Shozo NAITO: Automatic Text Classification Method with Simple Class-Weighting Approach†.
   Proc. of Third Natural Language Processing Pacific-Rim Symposium, pp. 498-503 (Dec. 1995).
   (5章に対応)

†Best Paper Awards 受賞

C．学会研究会講演

1. 山本和英, 増山繁, 内藤昭三：
   手がかり語を用いた日本語文章の段落分けに関する実証的考察,
   情報処理学会研究会資料 NL, Vol. 84, No. 9 (1991 年 7 月).

2. 山本和英, 増山繁, 内藤昭三：
   手がかり語を用いた日本語文章の段落分けに関する実証的考察,
   電気関係学会東海支部連合大会論文集, p. 549 (1991 年 11 月).

3. 山本和英, 増山繁, 内藤昭三：
   日本語文章における表層的段落構造の基本的検討,
   情報処理学会全国大会論文集, Vol.43, No.5H–3 (1991 年 11 月).

4. 山本和英, 増山繁, 内藤昭三：
   語の類縁性に基づく談話の結束性の評価尺度について,
   電気関係学会東海支部連合大会論文集, p. 559 (1992 年 10 月).

5. 山本和英, 増山繁, 内藤昭三：
   語の類縁性を用いた日本語文章の段落分けの試み,
   情報処理学会全国大会論文集, Vol.45, No.6G–8 (1992 年 10 月).

6. 山本和英, 増山繁, 内藤昭三：
   手がかり語及び語の類縁性を併用した段落分け,
   情報処理学会研究会資料 NL, Vol. 92, No. 6 (1992 年 11 月).

7. 山本和英, 増山繁, 内藤昭三：
   グラフ節点のある種の線形配列問題について,
   冬の LA シンポジウム, 京都大学数理解析研究所 (1993 年 2 月).

8. 山本和英, 増山繁, 内藤昭三：
   段落分けに関わる諸要素の評価について,
   情報処理学会全国大会論文集, Vol.46, No.7B–8 (1993 年 3 月).

9. 山本和英, 増山繁, 内藤昭三：
   グラフ節点の隣接枝のみを考慮した線形配列問題について,
   日本オペレーションズ・リサーチ学会研究発表会 (1993 年 3 月).

10. 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,
    情報処理学会研究会資料 NL, Vol. 99, No. 3 (1994 年 1 月).

11. 三輪倫子, 山本和英, 増山繁, 内藤昭三：
    「コ」「ソ」系指示語の用法に関する仮説とその検証,
    情報処理学会全国大会論文集, Vol.49, No.2K–4 (1994 年 10 月).

12. 山本和英, 増山繁, 内藤昭三：
    分類体系相互の関係を利用したテキストの自動分類,
    情報処理学会研究会資料 NL, Vol. 106, No. 2 (1995 年 3 月).

目 次

第 1 章

序　論

1.1　研究の背景と動機

我々人間の使う言葉をコンピュータに処理させよう，つまり計算機に言語能力を持たせようという試みは，計算機の誕生からそう遠くない時期に始まった．コンピュータで使用するプログラミング言語などとの違いを明確にするため人間の使用する言葉は自然言語と呼ばれ，これを計算機で処理するための研究は自然言語処理 (natural language processing, NLP) と呼ばれている[1]．人工知能 (artificial intelligence, AI) と認知科学 (cognitive science) の主要な研究課題でもある自然言語処理は，その後現在までに言語学など関連する分野を含めた広範な領域で活発に研究が行われている．例えば計算機を用いてある言語から他言語に翻訳させることを目的とした機械翻訳 (machine translation, MT) の研究は，この分野の一つの重要なテーマである．

自然言語処理の研究は，研究開始当初は様々な制約から一回の処理対象として単一の文[2]に制限することが多く，複数の文を同時に考慮するま

[1] 言語情報処理と呼ばれる分野もほぼこれに一致する．

[2] 文法上の単位．思想や感情を表す．完結した最小の言語表現．

でには至らなかった．ところが最近になって，解析対象の重点が単一の文から複数の文へと徐々に移動しつつある．自然言語処理のうち，複数の文を対象にする分野は特に談話処理 (discourse processing)，また場合によっては文脈処理と呼ばれる．この分野では，以前から簡単な文章生成や，文章の抄録／要約[3]生成などの試みがなされてきたが，それに加えて最近では対話の処理，解析なども広く研究されている．

また，最近の計算機の性能向上による処理環境の整備，及び機械可読文書の蓄積による必要性の増加に伴い，情報検索 (information retrieval, IR) の研究が現在活発に行われているが，この分野も検索対象の多くが言語情報，特に文章などの複数文であるため，必然的に談話処理との関連が強い．米国では，DARPA (Defense Advanced Research Projects Agency)[4]の研究プロジェクトである TIPSTER の一貫として，情報抽出の基本技術の向上を目的とした MUC(Message Understanding Conference) と呼ばれるコンテストが行われている ([Nom96], [Joh95]:pp.1613)．このコンテストは，日本語，英語の新聞記事を対象にして，特定のトピックをフレーム形式の抽出テンプレートで指定し，テキストからその指定された情報のみを抽出する精度を競うものである．

Salton らは，大量の機械可読テキストの中から使用者の要求するテーマのテキストを抽出するシステム「Smart システム」を 30 年以上に亘って開発した [Sal94]．このシステムは語のベクトル空間モデルを使用しており，その類似度を利用して，テキストとテキスト，あるいはテキストと部分テキスト (段落など) との類似度を計算している．また，この類似度を使用することによってテキストの内容同定や抄録も試みている．

[3]本論文では，抄録と要約を区別して別の意味で用いる．詳しくは第 3章参照．

[4]米国国防高等研究企画庁．国防総省の一部局で，先端軍事技術のプロジェクトを進めた．1993 年 ARPA と改称．

以上のように, 談話処理にはその対象, 目的などによって様々な処理を含む. 本研究ではこのうち, 以下の4項目について個別に検討を行った.

- 文章[5]の段落[6]分け
- 文章の要約作成
- 複数文章に対する要約作成
- テキスト[7]の自動分類

本研究では, 前述の項目のすべてについて, 日本語を処理対象言語として研究を行った. これらの研究が必要な理由, 及びこれらを研究対象とした理由は, それぞれ以下の通りである.

1. 文章の段落分け

   一般に, 文章は単なる文の羅列ではなく, 前の文と後ろの文との間に何らかの関係を持っている. この文章を正しく処理するためには, それらの文が文章全体から見た時にどのようにまとまっているのかという, 大局的なまとまりを把握する必要がある. さらに, 将来的には文章がどのような構造になっているかを把握する必要がある. 文章の構造把握に必要な処理の一つとして文章の段落分けを取り上げる.

2. 文章の重要部分の把握, 要約作成

---

[5]文が文脈をもって寄り集まり, それ自体が統一ある全体として完結している言語表現.

[6]文章を構成しているひとまとまりの文群を前後の改行によって視覚的にとらえやすく表示したかたまり.

[7]本研究では「文章」と「テキスト」をほぼ同義の語として使用する.

文章において，どの部分が重要であるかを把握する解析は，文章を理解する上での一つの重要な解析である．また，文章の要約作成，つまり文章において重要部分だけを抽出し生成する作業は，談話処理や文章生成の足掛かりとして重要な処理である．また最近，ネットワークを介してアクセスする電子図書館 (electric library) の実験プロジェクトが米国の NII(National Information Infrastructure) に刺激されて各地で動いている [Joh95] が，文章の要約作成は，このうち情報抽出処理の一部に該当し，電子図書館の実現に向けて重要な処理の一つに位置づけられる．このように，要約作成技術は自然言語処理の応用分野の一つとして一般社会からの潜在的需要が特に高く，近い将来優先して研究を進める必要があると考えられるため，本研究の検討対象とした．

3. 複数文章に対する処理

文章の要約処理をさらに一般化すると，複数の文章に対する処理が必要となってくる．この複数文章に対する処理は，現在までほとんど研究が行われてきていないが，この処理は今後談話処理における一つの重要な処理となる．また前述したように，電子化された文書の情報抽出処理の一部として実現が待たれる技術の一つとなり，ますます重要となる．このため本研究ではこれらの基礎的研究として，その重複部分を除去することによる複数文章の要約作成を検討の対象にした．

4. テキストの自動分類

テキストの自動分類，つまりテキストをその内容によっていくつかのカテゴリーに分類する処理は，談話処理の面からも重要な処理と

考えられる．テキストを分類するためにはそのテキストの内容把握が必要であるからである．

ところで，談話処理は形態素解析，構文解析などの処理後に，これらの処理結果を利用して行うと考えるのが一般的である．そのため，これらの処理結果に誤りや曖昧性が含まれていると談話処理にも悪影響を与えるので，本来ならばその基礎技術が確立された後に研究を進めることが望ましい．しかしながら，

- 談話処理のうち一部の処理は，表層処理のみで可能である
- 談話処理は文を単位とする処理と比較して同等，あるいはそれ以上に困難な問題を多く抱えると考えられるため，長期的に取り組む必要性がある
- 要約などの処理は，現在すでに潜在的需要が高いと考えられ，一刻も早い実用化，または一部実用化が待たれる

ことから，構文解析などの比較的深い解析を要求しない処理，つまり表層処理によって，談話処理固有のいくつかの問題を検討した．

1.2　本研究の特徴

談話処理にはいろいろな要素を考慮することが必要となってくる．例えば前後の文の論理的関係，語彙面での類縁性などである．これらの要素が相互に，複雑に影響しあってはじめて，文章に文の羅列ではない「文章らしさ」が形成される．このため文章を解析するためには，

- 各要素が，どのように，またどの程度全体に影響するのか

を考慮しなければならないと同時に，

- 各要素間が，相互にどのような影響を与えるのか

を考慮しなければならない．本研究では全体にわたってこの点に留意して研究を行った．

本研究ではまず，文章の結束性，つまり文章内の個々の文同士のつながりを把握することを目的にして，文章の段落分けを試みた．段落分けとは，段落に分けられていない文章，つまり文の羅列を入力として，それを段落に切る処理である．この研究は短期的には今後社会的需要が増加すると考えられる文章の推敲支援の一部として，長期的には将来の文章生成の基礎研究と位置付けられる．一般に，段落は意味のまとまりと考えられており，段落分けした結果の段落は何らかのまとまりを持っていなければならない．そこでこの研究では接続詞などの文章の部分的構造を明示する語句と，文章内の語彙的なつながりの２点に着目し，これらによる結束性が希薄な部分を段落の切れ目と判断することによって段落分けを行っている．また，実用の可能性を考慮し，実際の文章 (科学雑誌記事，新聞コラム) を対象にして，計算機上で実験，検証を行っている．

次に文章の要約処理について検討を行った．要約に関しては，現在までに類する研究のほとんどが文章中から重要な文を単に抽出する「抄録」の研究 (例えば Luhn[Luh58]，間瀬 [Mas89])，または実在する文章を対象にしていない思考実験上の要約 (例えば田村など [Tam92]) もしくは処理対象が限定されているシステムによる要約 (例えば中澤など [Nak91]) であることが予備調査で明らかになった．また，特に日本語を対象にした要約の研究では，要約処理以前の解析を必ずしも十分な精度で行えないことからこの傾向が一層強かった．このため，本研究では段落分けと同様にあえて実際の文章 (新聞社説) をその要約対象にして研究を進め計算機上にシステムを構築した．

また，要約は本当に重要な文を抽出することも必要であるが，要約文章を独立した一編の文章として読んだ時に全体として何らつながりを持たなければ要約したとはいえない．そこで要約の研究では，指示詞の出現に着目して，結束性を保ったまま文を抽出することで，全体として不自然な文章とならないよう試みた．さらに，単なる抄録では余分な部分の削除が十分でないと考え，抽出文の一部の修飾語句を除くことを試みた．

これらの研究では，表層に表出した現象を可能な限り利用したヒューリスティックなアプローチで研究を行っている．これには，現時点で構文解析 (特に日本語の構文解析) や意味解析が十分な精度で結果が得られていないこと，あるいは解析ツールなどの入手が困難なことが背景にある．むしろ現状では，表層の表現を十分に利用したヒューリスティックなアプローチの方が実際の文章を対象にした実験では有効な結果が得られる場合も多いと考えられる．

談話処理には文単位での処理が前提となるが，日本語をその対象言語とした場合，現状で構文解析など，一連の日本語文処理を行うことは未だ

容易ではない．また，それぞれの問題は単独でもかなり大きな問題である．このため，本研究では表層構造をできる限り利用してその解決を試みた．

さらに，何らかのつながりを持った複数新聞記事の要約も試みた．ここでは，類似した複数文章の要約に特有の問題である類似部分の特定，およびその削除を試み，ある程度重複した説明の省略，同じ語に対する修飾語句の削減などの点に着目することにより，文章の短縮化を試みた．

本研究ではまた，文章の自動分類の研究にも取り組んだ．この研究では，従来からある統計情報を用いた分類手法 [Tam88] と，分類体系に依存した情報を用いた手法 [Suz88] の中間的な手法を用いることで，前者の分類精度の低さと後者の汎用性の低さを同時に解決した．本研究が行った手法はベクトルモデルに属する手法で，出現語の頻度に何らかの重みづけすることによりベクトルの要素を決定する *term-weighting* と呼ばれる手法 [Sal88] の拡張であるが，ベクトルの要素を語から語群に変更することで語の言い替えなどの現象に対応した手法となっている．

1.3　本論文の構成と内容

本章に続いて第2章では, 文章の段落分けについて述べる. 複数文からなる文章の解析における基本的な問題は, 文間に存在する結束性を見い出すことである. 本研究では結束性を構成する要因の中から, 「手がかり語」, すなわち接続的語句と, 単語間の類縁性, すなわち出現した語の意味的な類似性の二つに着目する.

まず, 接続的語句が結束性に与える影響を, 段落分けを行うことによって検証する. 実際の文章の各文頭に出現する語の傾向に基づいて手がかり語を定義し, 各語の統計的特徴と段落長の要素を用いて, 計算機によって段落分けを試みる. 段落分けの自然さを, 原文の再現性及びアンケート結果の2種類の基準により評価する. 次に, 語彙的要素を用いた段落分けを試みる. シソーラスを使用して語句の類縁性を数値化し, これに基づき文章の結束構造をグラフによりモデル化したものを結束グラフと定義し, 段落の設定を評価する関数を, それに基づいて定義する. さらに, 前述の手がかり語と語句の類縁性の二つの要素に基づく3種類の方法により, 実際に段落分けを試みる.

第3章では, 要約処理を試みる. 一般に, 質の良い文章要約を行うためには, ある一つの言語現象だけをとらえた談話解析だけでは不十分である. なぜなら, 談話に関わる言語現象は相互に関連しているからである. 本研究ではこの観点から, 日本語での様々な表層的特徴をできるだけ多く利用して, 日本語文章の要約を試みる. 本稿では実際に計算機上で試作した論説文要約システム *GREEN* に関して, これで用いられている論説文要約の手法の紹介と, これによって出力された文章の評価を行う.

第4章では, 複数のテキストに対する要約について述べる. 日本語新聞

記事を対象として，単一のテキストの要約にはない，重複部分の把握，及びその除去という固有の問題に対して，連体修飾語，類似節，名詞句の言い替えを利用した要約手法とその実験結果について述べる．

第5章では，分類体系相互の関係を利用した日本語テキストの自動分類手法を提案する．従来の分類手法は，表記の統計情報を用いた手法と分類体系に依存した情報を用いた手法の二つに大別できるが，本手法ではシソーラスによる統計情報から分類体系相互の関係を自動学習するという両者の中間的な手法を用いることで，前者の分類精度の低さと後者の汎用性の低さの問題点を同時に解決する．また，語を基準としてベクトルモデルを形成するのではなく，意味的な語群によってベクトルの要素を構成することで語の言い替えなどによって頻度が低下することを防止している．

以上の各章の手法は独立したものであり，個別に適用することができる．第6章では，第5章までに検討してきた結果を総合的にまとめ，今後の課題について述べる．

第 2 章

段落分けを用いた日本語文章における結束構造の検討

2.1　はじめに

構文規則に従って, 品詞を並べても必ずしも意味のある文が成立しないのと同様に, 相互に関連のない文を羅列しただけでは意味的にまとまりのある文章は必ずしも成立しない. すなわち, 文章中での文の出現順序には, 意味的なつながりをつけるための何らかの制約が存在し, 逆に, この制約を満たす文の順序により, 文章に意味的なまとまりが生じていると考えることができる. 本章の目的は, 文章の意味的なまとまり, つまり文章内の結束性を解析することにある. また一般に, 段落は文章の意味的なまとまりを示す一つの要素とされている. 本研究では, 文章の結束性が表出した結果が段落であると考え, 実際に文章を段落に分けることによって, 文章の結束構造を解明しようと試みた.

本章では, 手がかり語と語彙的な結束性の二つの要素を考慮して段落分けを行う. このうち手がかり語に関しては, 例えば福本らが, 手がかり語の他に文末表現などを利用して文章の構造化を試みている [Fuk91]. また, 語彙的な関係を利用したものとして, Morris and Hirst によっ

て英語を対象にして語の類縁関係から文章構造を解析した試みがある [Mor91]. この文献では, シソーラスに *Roget's International Thesaurus* を用いて, 語の類縁性に注目した文章の構造解析を行っている. また最近では, 野本 [Nom94] が談話セグメントという用語で, 望月ら [Mot95] がテキストセグメンテーションという用語で文章の構造化を試みている.

段落分けに関しては, 山崎が文章の構成単位として「意味的段落」というものを認め, 異なり語数などの話題の展開を計る尺度を利用して, 実際にその単位を切り出す方法を提案している [Yam83]. この研究は, 日本語を対象にして段落分けを試みている点において本章の内容に最も類似する研究であるが, 以下のような問題点, 及び本研究との相違点がある.

1. 「用意」と「準備」のような, 同義語の語の置き換えについても全く異なる語として計算してあり, 同義語などの類縁性に対する適切な考慮が行われていない.
2. 「語の持つ意味情報の殆どを捨象している」 [Yam83] ため, 解析の深さに限界がある.
3. 段落の分割基準が主観的であり, 不明確である.
4. 「データ数が少なく充分な考察とは言えない」 [Yam83].
5. 計算機を用いて解析していない.

本研究は段落分けそのものが目的ではないが, 段落分けが利用可能な応用として, 文章作成の支援や推敲の支援がある. これらの研究を進めていくことによって, 我々が文章を作成する場合に, たとえば, 適当な段

落分けの位置を計算機が指示することにより文章全体の構造を明確にし, 読者にとって読みやすい文章を書くための支援を行うための基礎データを得ることができる.

2.2　手がかり語を用いた段落分け

本節では, 文章の結束性の表現手段の一つである「手がかり語」, すなわち接続詞, 副詞などの接続的語句が結束性, 及び段落の設定に与える影響を, 段落分けを行うことによって検証する.

2.2.1　文頭の単語出現調査

実際の文章で, 段落の初めや段落内の文頭に, どのような単語が使用されているのかを知るために, 文頭の語の統計調査を行った. 調査の対象とした文章は, 科学雑誌「日経サイエンス」と朝日新聞のコラム「天声人語」の 2 種類である. 統計調査は, 2 種類のテキストに対して, 個別に行った.

調査は, すべての文頭について, どの文章にも出現する可能性のある単語を予め選び, それらの頻度を調べるという方法で行った. これらの中には, 文間に出現して陽に文と文を接続する役割をもつ接続詞, 副詞の他にも, (代) 名詞, 連体詞の中で照応によって文間の接続機能を果たす単語も含めて調査を行っている. この調査の, 抽出対象の語の例を表 2.1に示す. その他の専門用語などの語については, 予備調査の結果, テキストに依存しないで統計的に大きな割合で出現する単語は存在しなかったので, 今回の調査対象には含めなかった.

表 2.1: 調査対象単語の例

品詞	単語例
接続詞	そして, しかし, 一方, そこで
連体詞	この, その, そういう, ある
副詞	たとえば, もし, つまり, なぜ
代名詞	私, 彼, われわれ, それ, ここ
名詞	最初, 筆者, 現在, このよう
(その他)	したがって, とすると, 本稿

実際に調査した文章は, 日経サイエンスの記事 13 編, 天声人語 505 編である. また調査は, 計算機を使用して最長一致法によって対象となる単語を検出した[1]. 調査結果の統計データを表 2.2に示す. ただし, 表 2.2で抽出語の割合とは, 調査を行ったすべての文の内で, 文頭に調査対象単語が出現していた文数の割合を示している.

調査結果の品詞別データを表 2.3と表 2.4に示す. ただし, この表にある「指示語」とは, いわゆる「こそあど言葉」であり, 品詞とは別の概念であるが, 別途に集計を行い, 算出した.

品詞別の割合では, 接続詞が全体の 4 分の 1 程度しかないことがわかる. この結果は, 文と文を接続する役割を果たしているのが接続詞だけではないことを端的に示している [Nag86]. また接続詞については, 段落内の文頭に比較的多く使用される傾向にあり, 段落頭での使用頻度との

[1]現在, 一般的な形態素解析として最長一致法はほとんど使われていない [Joh95] が, ここでは, 文頭からの一語を抽出することだけが必要であるため, 最長一致法でも十分な精度を得ることができると考えられるのでこれを用いた.

表 2.2: 調査を行った文章

項目	日経サイエンス	天声人語
調査対象	13 編	505 編
総段落数	824	3039
総文数	3501	10233
総文字数	212403	416006
段落当たりの平均文字数	257.8	136.9
その標準偏差	130.6	62.3
抽出語の割合	60.0 %	30.2 %

差は日経サイエンスにおいて顕著である.

また, 品詞分類とは別に, 指示語について統計をとった結果をみると, 天声人語で 3 分の 1, 日経サイエンスでは抽出した語の 4 割近くに達するという結果を得た. 段落頭と段落内文頭との比較では, 接続詞, 連体詞, 代名詞は段落内文頭の方が割合が高く, 副詞, 名詞については段落頭の方が割合が高くなっている. この傾向は, 日経サイエンス, 天声人語という, 文章の種類に関係なく見られる. その理由として, 連体詞や代名詞の使用による前方の文との間の照応関係により, 文間の結合が強くなることが考えられる.

以上の調査結果から, 抽出された語の多くは段落頭よりも段落内の文頭に多く出現することがわかる. このことは, これらの語が文頭に出現した場合には, 比較的前の文とつながりやすいことを示している. そこで, 抽出した語でどの文章にも出現する可能性のある語のうち, 頻度の低い語 (本研究では頻度 1 の語とした) を除いた語すべてを「手がかり

表 2.3: 抽出語の品詞別割合 (日経サイエンス)

	全文	段落頭	段落内文頭
接続詞	25%	17%	27%
連体詞	20%	15%	21%
副詞	20%	25%	19%
代名詞	15%	13%	16%
名詞	14%	25%	11%
その他	6%	5%	6%
指示語	38%	32%	40%

表 2.4: 抽出語の品詞別割合 (天声人語)

	全文	段落頭	段落内文頭
接続詞	23%	21%	24%
連体詞	21%	19%	21%
副詞	15%	17%	14%
代名詞	12%	8%	13%
名詞	8%	10%	7%
その他	21%	25%	21%
指示語	33%	28%	35%

語」と定義する[2].

2.2.2　計算機による段落分けの試行

以上の調査結果に基づき, 計算機によって実際に段落分けを試みた. 段落分けアルゴリズムの設計は, 次のような方針に従った.

段落分けは, On-line 的方法, すなわち順番に入力文を読み込んでいき, 文毎に段落分けを行うかどうかを決定するという, 逐次的な方法で行う. 段落分けには, 手がかり語に関する情報と文長の情報を使用する. 文長は, 文の情報量を反映していると考えることができるので, 段落分け要因のひとつとみなし, 導入した.

まず, 文頭に出現した語によって, 前文との基本的な結合の強さを決定する. それを文長の要素によって修正する. また, 文長の要素は, 結合の強さに対して線形に影響すると仮定する.

これらの方針に基づいて, 以下のような段落分けのアルゴリズムを導入した. 2.2.1節の調査結果から, それぞれの手がかり語に対して, 以下の式に従い, 段落内での結合性の強さを表す数値 (これを原結合度と定義する) を割り当てた.

$$C_0 = 100\frac{A}{B} \tag{2.1}$$

ただし,

C_0：その手がかり語の原結合度

A：その手がかり語の段落内の文頭での出現数

[2]「手がかり語」という語は他の研究でも使用されるが, 何を手がかり語とするかの合意は得られていない [Joh95]. 本章では, 以上のように定義した.

B：その手がかり語の全出現数

である．手がかり語の抽出対象に含まれない，専門用語等の語については，全文に対して，平均的な結束性を示すと仮定し，それらの語全体を一つの手がかり語のように取り扱って，処理を行った．また，原結合度は 2 種類の文章それぞれについて別に算出する．

次に，段落分けを行う際に，段落の長さも考慮に入れるため，ある文までのその段落長を，段落の初めからその文の前の文までの文字数と定義した．つまり，段落を構成する文を順に $S_1, S_2, \cdots, S_n, \cdots$ とする．現在，段落分けの判断を行おうとする文 $S_n (n > 1)$ において，文 S_n までの段落長 L を，

$$L = \sum_{i=1}^{n-1} l_i \ (l_i：文S_i の文長) \tag{2.2}$$

と定義する．ただし，段落の第一文 $(n = 1)$ のときは形式的に $L = 0$ と定義する．以上の準備に基づき，その文の後で段落を分けるかどうかを判断するための文の結合度 C を，次のように定義した．

$$C = C_0 + \alpha_1 L_m - \alpha_2 L \tag{2.3}$$

$$= C_0 + (\alpha_1 - \alpha_2) L_m + \alpha_2 (L_m - L) \tag{2.4}$$

ここで，L_m：平均段落長，α_1, α_2：定数である．

各文末で算出した結合度が 50 以下ならば，その位置で段落を分ける．以後，この操作を反復し，文章全体の段落分けを行う．

日経サイエンスと天声人語の文章のうちで, 2.2.1節で手がかり語の調査を行ったのとは異なる文章を対象にして, 前述のアルゴリズムを用いて実際に段落分けを行った. 対象は, 日経サイエンス 6 編, 天声人語 40 編である. 対象となる文章から段落分けを除去した文章を入力とし, このアルゴリズムにより段落分けを行わせた. 言語は, *Common Lisp*(KCL) を用い, *Sun SPARC Station I* 上で実行した. 実験に使用するパラメータ α_1 や α_2 の値は, あらかじめ推定を行うことが困難であるため, 先に α_1 を仮に固定し, 出力段落数が原文の段落数に近くなるように α_2 を決定するという方法を繰り返す, という試行錯誤によりパラメータを決定した.

2.2.3　原文との比較による自然さの検証

出力された文章の段落分けの自然さの評価基準のひとつとして, 原文の段落との一致度をあげることができる. 比較的良好な結果を得られたいくつかのパラメータ対についての, 原文段落分けとの一致した割合を表 2.5に示す.

表 2.5では, 原文とどの程度一致したかを, 再現率, 適合率の二つの評価基準で表現している. すなわち, 再現率 R_r, 適合率 R_p は次式で定義される.

$$R_r = \frac{N_{ab} - 1}{N_a - 1} \tag{2.5}$$

$$R_p = \frac{N_{ab} - 1}{N_b - 1} \tag{2.6}$$

ただし,

表 2.5: 原文に対する一致の割合

文章	α_1	α_2	R_r	R_p
日経サイエンス	0.02	0.1	40.3%	36.8%
	0.02	0.15	48.4%	33.5%
	0.03	0.15	44.2%	32.8%
天声人語	0.01	0.1	38.1%	34.6%
	0.1	0.2	37.1%	35.7%
	0.3	0.5	34.1%	30.2%
	1.5	2.0	35.1%	31.8%

N_{ab}：原文の段落と出力段落とで共通する段落数

N_a：原文の段落数

N_b：抽出した段落数

である．再現率, 適合率共に高い程, 原文に対する再現性が高いことを示す．

計算機で出力された結果の再現性が低いことの理由の一つとして, 2.2.1節での調査において, 手がかり語が日経サイエンスでは約 6 割の文頭にしか出現していないことをあげることができる．このことは, 手がかり語情報としては, 全体の 6 割しか使用していないことに対応する．したがって, 手がかり語情報だけでは, 再現率は最大 6 割であると考えられる．

2.2.4　アンケートによる自然さの検証

文章の段落分けの自然さの判断基準は一般には複雑であり, 本アルゴリズムによる段落分けの自然さの評価を数値化することは容易ではない. 本章では, 段落分けアルゴリズムの評価基準のひとつとして, 人間がその出力結果を読んで, 原文の段落分けと区別ができないかどうかということを採用した.

ここでは, 主観性を取り除いて人間に判断してもらうために, 次のような方法でアンケートを行った. すなわち, 被験者に対して, 文章の原文を 1 編以上と, 計算機によって段落分けした文章を 1 編以上含む, 合計 5 編の文章を提示する. 5 編の記事の内訳は被験者には示さない. そしてこれらの文章のそれぞれについて, 原文かどうかを当ててもらう. このアンケートを日経サイエンスと天声人語について行う. ただし, 天声人語は提示する合計編数を 7 編とする.

以上のような方法によって, 実際にアンケートを工学系の大学生, 大学院生 11 人 (日経サイエンスは 10 人) に対して行った. その結果を表 2.6に示す.

表 2.6: アンケート調査の結果

文章	提示のべ文章	原文認識率
日経サイエンス (原文)	27 編	67 %
日経サイエンス (出力文章)	23 編	26 %
天声人語 (原文)	34 編	76 %
天声人語 (出力文章)	43 編	44 %

表 2.6に示すように, 天声人語に対する計算機による出力結果は, 4 割以上が原文と認識された. この結果は, 天声人語の原文に対して, 原文再認識率が 4 分の 3 であることを考慮に入れると, 比較的高い数値であると考えることができる. 日経サイエンスについては, 天声人語よりも低い認識率となった. この理由としては, 日経サイエンス一編の段落数が比較的多いために, 明らかに不自然と思われる段落の分け方が出現する可能性が高くなることが考えられる.

2.3　語の類縁性の導入とそれを追加した段落分け

2.3.1　語彙的手段による結束性

本節では, 前述の手がかり語の要素の他に, 語彙的手段による結束性にも着目し, この両者が文章の結束構造に与える影響を考察する.

語彙的手段による結束性は, 同一語句を含む, 意味の類似性 (similarity) による結束性と, 意味の近接性 (contiguity) による結束性に分類できる [Ike83]. 前者はさらに,

男の子 が立っていた. その 少年 は泣いていた.

のように, 類義語 (synonym) の場合と,

男の子 が立っていた. その 子供 は泣いていた.

のような上位語 (superordinate) や下位語 (hyponym) の場合に分類できる. 一方, 後者の意味の近接性の例は以下である.

空はとても青い. 今日は雲一つない天気である.

前者の要素をとらえるための資料として, シソーラス (thesaurus) が利用できるが, 後者をとらえるためには大規模な知識ベースの構築が必要である. 本論文では語彙の近接性は扱わず, 同一語句, 上位・下位語, 類義語, 対義語のみを考慮し, 以後この関係をまとめて語の類縁性と呼ぶ. また, 語間の類縁性を測定する基準として, 本研究では角川類語新辞典 [Oon81] をシソーラスとして使用する.

2.3.2　結束グラフ

文章中の文間の結束度を表現するために結束グラフを構成する. 結束グラフ $G = (V, E, w)$ を次のように定義する.

$$
\begin{aligned}
V &= \{v \mid v \text{は文章中の一つの文} s \text{に対応する}\} \\
E &= \{(u, v) \mid u \text{に対応する文中の語と} v \text{に対応する} \\
&\quad \text{文中の語でシソーラスの中分類, 小分類での} \\
&\quad \text{同一の分類に属する, または同一の語である} \\
&\quad \text{ものが存在する}\} \\
w &: E \to \Re\ (\Re : \text{実数の集合}), (u, v) \in E \text{に対し,} \\
&\quad \text{下の式(2.7)で節点} u \text{と} v \text{の結束度} w(u, v) \text{を定義する}
\end{aligned}
$$

$$
\begin{aligned}
w(u,v) &= \exp\{\lambda d(w_1x_1 + w_2x_2 + w_3x_3)\} \qquad (2.7)\\
d &: u \text{ と } v \text{ 間の距離}\\
x_1 &: u \text{ と } v \text{ での同じ語の組数}\\
x_2 &: u \text{ と } v \text{ での小分類一致の組数}(x_1 \text{を除く})\\
x_3 &: u \text{ と } v \text{ での中分類一致の組数}(x_1, x_2 \text{を除く})\\
& w_1 > w_2 > w_3 > 0\,,\ \lambda < 0
\end{aligned}
$$

すなわち, 二つの文の語彙的結束度は, 前述した 3 段階の枝の強弱を定数 w_1, w_2, w_3 とし, それぞれに枝の本数を掛けたものの総和として定義している. 結束度が文間の距離に関して単調非増加となる性質を持たせるために, 指数関数を採用した.

2.3.3　結束度の評価関数

与えられた文章に対する段落分けの評価関数を, 以下のように定義する.

$$
\begin{aligned}
&\max \text{ 評価関数}\\
&= \alpha \sum_{p_i \in T} \left(\sum_{s_j \neq s_k \in p_i} \frac{w(s_j, s_k)}{|p_i|} - \sum_{s_j \in p_i, s_k \notin p_i} \frac{w(s_j, s_k)}{|p_i|} \right)\\
&- \beta \sum_{p_i \in T} \left(\sum_{s_j \in p_i} l(s_j) - \frac{\sum_{p_j \in T} \sum_{s_k \in p_j} l(s_k)}{|T|} \right)^2 \qquad (2.8)
\end{aligned}
$$

ただし,

$T = \{p_1, p_2, \cdots, p_n\}$: 段落 $p_1, p_2, \cdots, p_n$ からなる文章

$p_i = \{s_j, s_{j+1}, \cdots, s_k\}$: 文 $s_j, s_{j+1}, \cdots, s_k$ からなる第 i 段落

$l(s_i)$：文 s_i の文字数
$w(s_i, s_j)$：文 s_i と文 s_j の語彙的結束度
$|S|$：集合 S の要素数
$\alpha > 0$, $\beta > 0$：定数

である.

式 (2.8) の二項は, それぞれ文間の語彙的結束度と段落長 (段落内の文字数) の要因を定式化したものである. 文間の語彙的結束度は, 段落内の文間の結束度はプラス要因, 段落間の文間の結束度はマイナス要因とした. また, 段落長は, バランスの良さをプラス要因とした.

2.3.4　両要素の併用法

2.2.2節での式 (2.1) の原結合度 C_0 と, 2.3.3節での評価関数式 (2.8) の両要素を併用するためには, 以下の方法が考えられる.

1. 一つの数値に集約する方法 (集約法)
   原結合度と評価関数式の両者に対して加算, あるいは乗算等の演算を施して一つの数値に集約する方法である.

2. 原結合度を先に考慮する方法 (閾値法)
   まず原結合度に着目して, その数値がある上限閾値以上の個所では段落をつなぎ, 下限閾値以下の個所では段落を分ける. 残りの個所に関しては, 評価関数式を考慮して段落分けを行う.

3. 集約法＋閾値法
   前述の集約法と閾値法は, 別の部分で統合を行っているために, 同時に実現することも可能である. これを同時に行うこの方法は, 原

結合度の値が両端に近い場合は閾値法を採用して, 中間の部分は集約法を採用することによって実現される.

文頭に現れる手がかり語の情報は, 類縁性などによる文章の結束構造を, 明示的に表現したものとみなすことができる. この観点からは, 手がかり語の情報は語の類縁性の情報の一部であるともみなすことができるので, 評価関数式を先に考慮する方法は除外した.

2.3.5　システムの構成

上記のモデルに基づき, 実際に段落分けを行うシステムを作成した. システムはすべて, *Common Lisp* (KCL) を用い, *Sun SPARC Station I* 上で作成した. システムは 3 つの部分から構成されている.

1. 形態素解析部

   この部分では, 語尾変化処理と簡易な形態素解析を行う. 実験ではシソーラスに掲載されている語を切り出すことを目的としているので, 掲載されていない語, つまり, 多くの付属語部分については正しい形態素に分解される必要はない. このため, 本研究では文節数最小法による形態素解析を行った. 文節数最小法は, 文の最後まで見て, 文節の数が最小であるような解釈をとるものである. この方法は, 必ずしも正しいわけではないが, 簡単な割にはかなりの精度が期待できる [Joh95]. また, 本辞書には, シソーラスと同じ角川類語新辞典 [Oon81] に固有名詞の辞書を付加したものを採用した.

2. 評価関数値の算出

   形態素解析の終了後, すべての文の組合せについて, 前述した計算

式で評価関数値を計算する[3]．実験では，3 種類の枝の重みの比 w_1 : $w_2 : w_3$ を，10 : 8 : 3 としている．

3. 段落分け

Off-line による方法，つまり文章全体がすでに入力されてから段落分けを行っている．具体的な段落分けは，以下の通りである．

(a) 初期設定として，1 段落 1 文とする．閾値法では，まず原結合度と閾値に基づき，段落に分けない個所と分ける個所を決定する．

(b) ある位置で隣接する段落を併合したとして，その時の評価関数の減少量が最も小さい位置 (あるいは最も増加する位置) を段落のつなぎ目とする．

(c) この操作を繰り返し行い，目的関数がそれ以上改善されなくなった段落の分け方を最も自然な段落分けの候補とする．

段落分けの近似方法については，順次併合する方法と，順次分割する方法が考えられる．本研究では，予備実験で相対的に結果の良かった前者を採用した．後者が相対的に悪かった理由は，初期状態から数段落を構成するまでの段落決定において，後者では段落長の要因が大き過ぎるためと考えられる．

2.3.6　実験結果

実験は，朝日新聞「天声人語」 50 編を用いて，

[3]文間の結束度の計算は文数に対して二乗に比例した計算量，記憶量となるが，実際のテキストでは比較的少ない文数でくくられており，十分実用可能であると考えられる．また，ある距離以上の文間の結束度はすべて 0 と近似した修正を行えば，長いテキストでも計算量，記憶量共に軽減が可能であり，さらに実用的になる．

1. 類縁性のみを考慮した方法

2. 集約法

3. 閾値法

4. 集約法＋閾値法

の 4 種類について行った．実験結果を図 2.1, 図 2.2に示す．ただし閾値法では, 原結合度 80 以上の場合に段落をつなぎ, 原結合度 20 以下の場合に段落に分けるとして実験を行った．また, 集約法による実験では, 集約演算に乗算を用いた．

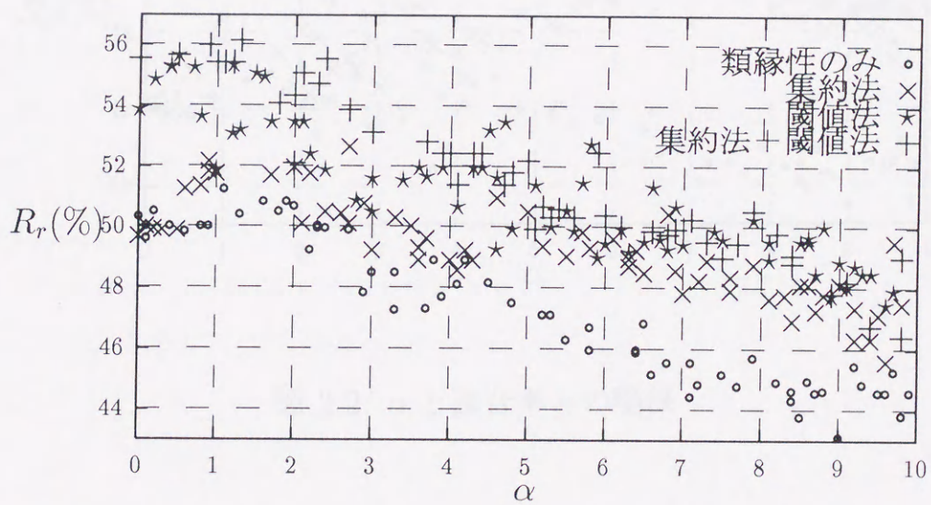


図 2.1: α と再現率との関係

図 2.1, 図 2.2の縦軸の再現率, 適合率は, それぞれ式 (2.5), (2.6) に従う．また, 横軸の α は, (2.8) 式において β を 0.001 に固定した時の α で

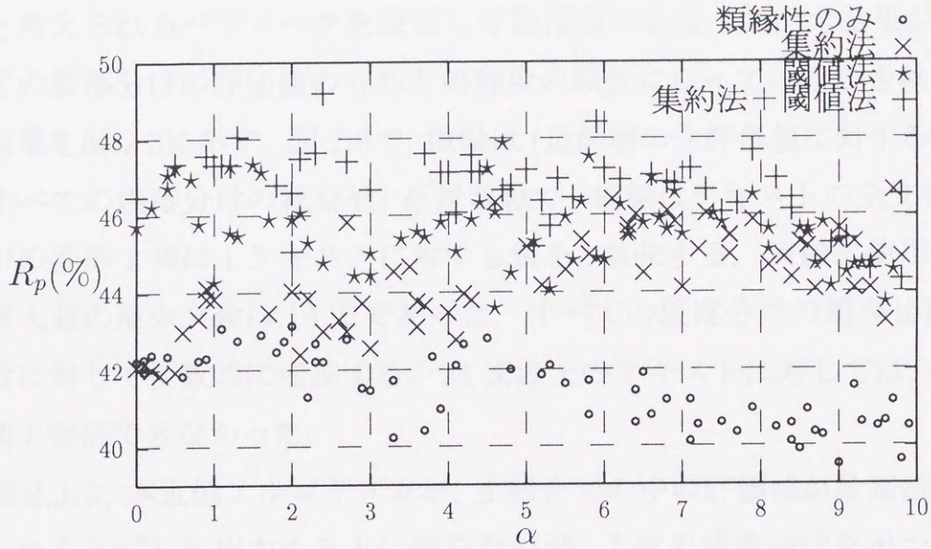


図 2.2: α と適合率との関係

ある．両図によると，集約法，閾値法共に，従来の語の類縁性のみを考慮した方法よりも再現率，適合率共に向上している．これは，前後の 2 文間が語彙的に強い (あるいは弱い) 結束性を持った関係になっていても，手がかり語によってその結束性の強弱を変化させることができることを示している．

本研究で使用している段落分けはヒューリスティックな方法のため，ここでその妥当性を検証する．図 2.1，あるいは図 2.2より，最も妥当であると考えられるパラメータを設定して段落分けを行った出力結果が，すべての段落分けの評価値のうちどの程度の順位になっているかを検証した結果を図 2.3に示す．図 2.3で，横軸は (近似解の全評価値に対する順位／すべての段落分けの組合せ) を対数軸で，縦軸はテキストの全文数を，図中の菱形 1 個は 1 テキストに対する値を，意味する．実験に使用した天声人語の最少文数は 14 文であった．すべての段落分けの組合せ数は，文数に対して指数的に増加する．24 文以上のテキストに対しては，計算時間上評価できなかった．

図 2.3は，本近似アルゴリズムが，全組合せの中の評価値が最高値の段落分けを必ずしも出力するとは限らないが，上位の段落分けを出力しており，近似アルゴリズムの妥当性を示している．

2.4　段落分けの自然さの検証

本節では，段落分けの自然さの評価を，原文の再現性という基準だけで行うのは不十分であることを明らかにするとともに，段落分けのすべて

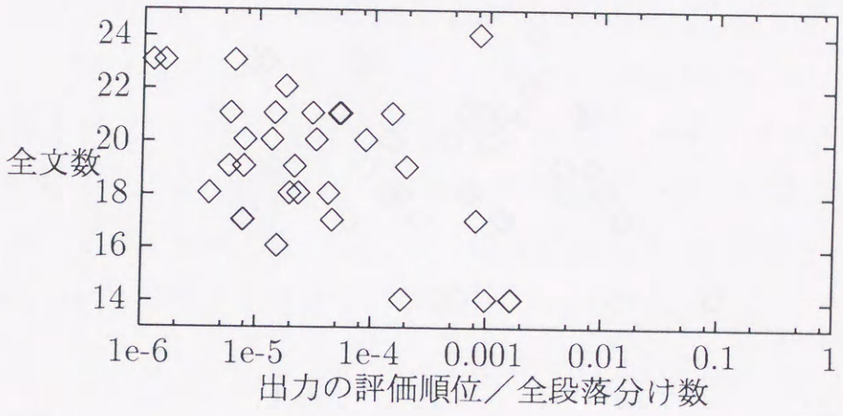


図 2.3: 出力文章の評価値順位

の可能性の中での原文段落分けの評価値順位, アンケート調査による人間の感覚との相関という二基準によって評価関数の妥当性の検証を行う.

2.4.1　原文との比較による評価関数の妥当性

原文の段落分けは自然であると仮定すれば, 段落分けのすべての可能性の中での原文の段落分けの評価値順位を調べることによって, 評価関数の妥当性を検証することができる. そこで, この検証結果を図 2.4に示す. 原文の評価値は比較的上位にあるので, 本研究で定式化した評価関数は, この基準の下で妥当であると考えられる.

2.4.2　アンケート調査

段落分けの評価関数と人間の感覚がどの程度一致するかを検証するため, アンケート調査を行った. アンケートはまず, 文単位に切った文章と,

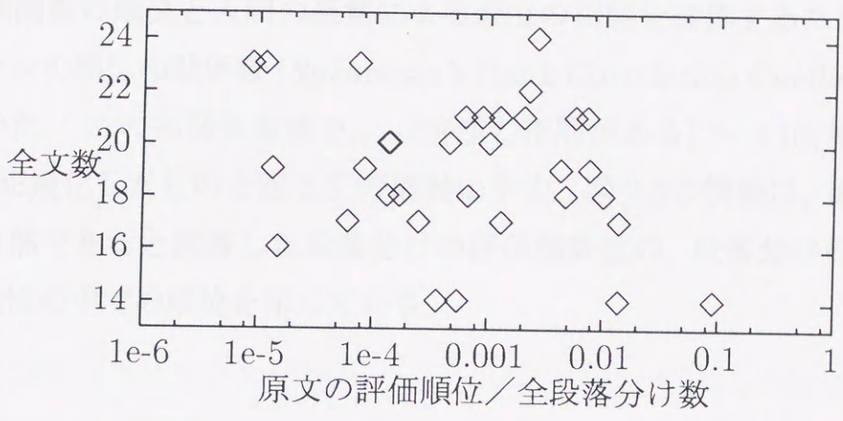


図 2.4: 原文の評価値順位

その文章の段落分けとして尤もらしい候補 (原文の段落分けを含む) を五つ提示する. その上で,

1. 提示した五つの候補を自然な順に並べる
2. 提示した候補の中から原文を選択する
3. (提示した 5 つの候補に関係なく) 最も自然な段落分けを指摘する
4. 絶対に段落に分けない位置を (任意個数) 指摘する

という四つの設問に答えてもらう. 提示文章には, 朝日新聞「天声人語」 3 編を用いた. 被験者は工学系の大学生・大学院生であり, 21 名から回答を得た.

2.4.3　調査結果

評価関数の順位と人間の感覚による順位の相関を評価するために，スピアマンの順位相関係数 (*Spearman's Rank Correlation Coefficient*)[Tak89] を用いた．この相関係数値を，−10(逆の相関がある) ～ +10(相関がある) に正規化したものを図 2.5 の縦軸に示す．図 2.5の横軸は，被験者が最も自然であると回答した段落分けの評価関数値の，段落分けのすべての可能性の中での順位を示している．

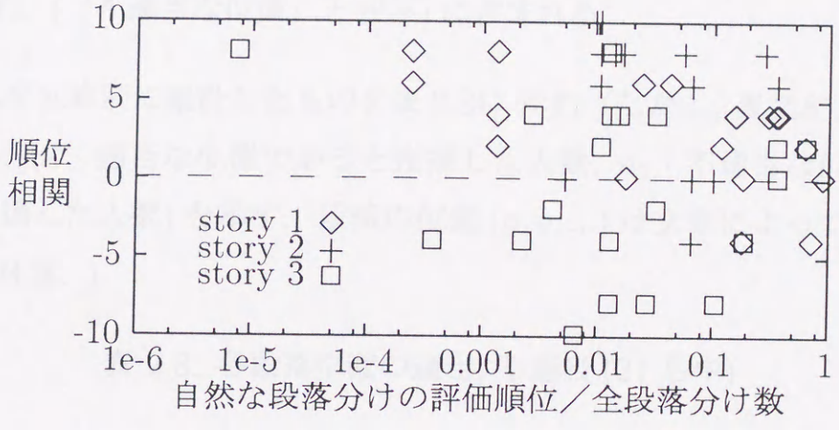


図 2.5: 順位相関係数の分布

各文章ごとの順位相関係数の平均と，原文を正しく指摘した人数を表 2.7に示す．

また，評価関数値の最も良い段落分けの妥当性を評価するために，アンケート調査から，評価関数値の最も良い段落分けの，各段落分けの位置 d に対して，以下の条件を満たすアンケート調査結果を集計した．

- 段落分けの位置 d が「最も自然であると回答した段落分け」（「適

表 2.7: 順位相関係数の平均及び原文指摘者数

	文章 1	文章 2	文章 3
順位相関係数	3.24	5.43	−0.57
原文指摘数 (21 名中)	7	0	1

当な位置」と呼ぶ) に含まれる.

- 段落分けの位置 d が「絶対に段落に分けないと回答した段落分け位置」 (「不適当な位置」と呼ぶ) に含まれる.

これを文章別に集計したものを表 2.8に示す. (ただし, 表 2.8で各欄は (n_1/n_2, n_1：適当な位置であると指摘した人数, n_2：不適当な位置であると指摘した人数) を示す. 段落の位置 $(a, b, ...)$ は文章によって異なることに注意.)

表 2.8: 各段落位置の適当, 不適当 (21 名中)

段落位置	a	b	c	d	e	f
文章 1	9/2	12/2	21/0			
文章 2	0/16	20/0	12/1	17/0	19/0	
文章 3	3/6	16/3	3/4	17/1	8/0	9/2

2.4.4　考察

- 表 2.7は, 原文の段落分けであることの認識率が非常に低いことを

示している．このことは，読者にとって著者による段落分けが最も自然な段落分けであるとは限らないことを意味する．これは，原文の段落分けだけが自然な段落分けではないこと，自然な段落分けには自由度があることを示している．

- 表 2.7より，順位相関係数は，文章によっては必ずしも正の値をとらない．この原因としては，アンケートで提示した 5 つの選択肢が，いずれも尤もらしい候補であったことが考えられる．すべての段落分けの可能性の中から全くランダムに選んだ五つの選択肢の順位づけと，評価関数との間で順位相関をとれば，相関係数値は，調査の最高値 (5.43) 程度，あるいはそれ以上になると予想される．このことから，提示した評価関数と人間の感覚との間の相関度は高いと考えられる．

- 表 2.8より，最も評価値の高かった段落分けの位置の多くは，アンケート調査で「適当な位置」と指摘されていることから，計算機による出力結果の妥当性を支持している．一方，表中の一ヶ所 (文章 $2-a$) は，ほとんどの被験者が「不適当な位置」と回答している段落位置にもかかわらず，原文ではこの位置で段落を分けている．このことから，著者と読者の段落に対する感覚が必ずしも一致しないことがわかる．

2.5　まとめ

本章では, 日本語文章中の結束性を解析することを目的にして, 「手がかり語」, すなわち接続的語句と, 語の類縁性, すなわち出現した語の意味的な類似性の二つの要素に着目し, 計算機を用いて段落分けを行うことによってこれらの要素の影響を考察した. その結果, 次の 2 点が明らかになった.

1. 日本語文章の結束性には手がかり語と語の類縁性が共に影響を与えているが, 特に, 明示的, 意識的な前者の要素よりも後者の方がより大きな影響を与えている.
2. 日本語文章には, 自然と感じる段落分けが存在する. ただし, これは唯一ではなく, また, 個人差などの要素によってある程度変動する.

本章に関連した課題としては次の 2 点をあげることができる.

- 語彙の近接性の考慮

  前述したように, 2.3節の実験では語彙の類縁性のみを考慮して実験を行っており, 語彙の近接性は取り扱っていない. しかし, 語彙的な手段による結束性を考慮するためには語彙の近接性の考慮も必要である. しかし, これを取り扱うためには, 大規模知識ベースなどの情報が必要であり, その開発が待たれる.

- 照応・省略の考慮

  本章では, 結束性表示の要素として, 手がかり語と語彙の類縁性の二つを考慮した. これらの他に結束性を示している要素として, 指示語の使用による照応, あるいは省略がある [Ike83]. これらの要素についても今後, 検討していきたい.

第 3 章

文章内構造を複合的に利用した論説文要約システム *GREEN*

3.1　はじめに

本章では日本語の論説文を対象にした要約文章作成実験システム *GREEN*[1] (以下 *GREEN* と呼ぶ) について述べる.

一般に, 質の良い文章要約を行うためには, 照応, 省略, 接続的語句, 語彙による結束性, 主題・話題, 焦点など多くの談話現象の処理が必要であり, これらの談話現象は互いに複雑に影響しあっているので, これらの談話現象の一部のみの処理を行って要約を試みても, 質の高い要約が得られる可能性は低い. 本研究の目的は, 以上の見地から現状で解析可能な談話要素をできるだけ多く取り込み, 実際に計算機上で動作する実験的な要約作成システムを試作してその効果を検討することである.

文章要約については, 日本語学あるいは日本語教育の分野でも, 現状では定義や手法が確立していない [Sak89]. 本章では, 文章要約とは重要度が相対的に低い部分を削除することであるとみなす. 一般には, 文章中のある部分の「重要度」は文章の種類によって異なるので, 要約の方法は,

[1] *Generator of REcapiturations of Editorials and Notices* の略.

文章の種類によって異なったアプローチを取る必要があると考えられる. 本研究では, 新聞社説などの, 筆者が読者に対して何らかの主張や見解を示す文章 (以下, 論説文章と呼ぶ) を要約の対象にする.

従来, Luhn[Luh58] に始まるといわれるテキストの抄録[2]作成の研究として例えば Hasida *et al.*[Has88], 住田ら [Sum95] などの研究が, あるいは要約[3]生成の研究としては邑本ら [Mur90], 田村ら [Tam92] などの研究がある.　また, 抄録システム[4]としては沖電気工業 (株) の COGITO[Kom87] が有名である. 田村ら [Tam92] は, 文章の構造および話題の連鎖を表現する修辞構造ネットワークおよび話題構造を作成することによる要約方式を提案しているが, 思考実験に留まっており, その実現には, 一般的知識に関するシソーラスの構築や, 修辞構造ネットワークの自動作成手法などの困難な問題が残されている. また間瀬らは, 「重要文に比較的よく出現する表層的特徴を多種類含んでいる文が真の重要文である」という仮定に基づき, 題名語, 高頻度名詞, 主題 (助詞「は」) などのパラメータを総和することによって重要語を決定し, 要約文を選択するという統計的手法に基づく要約法を提案している [Mas89].

本研究では, 要約中で原文章の文をそのまま使用するのではなく, 文内で比較的重要度が低いと考えられる連体修飾要素の削減も行った. 一方, 本手法は文章内の談話構造の利用による文章要約を試みたものであり, [Mas89] などの従来の抄録作成に使用されてきた語の頻度に関する情報は, 使用しなかった. また, 前述の両論文でも使用している文章のタイトル (題名) の情報も, タイトルはそもそも文章の「究極的な要約」で

[2]原文から文単位でその一部分を抽出したテキスト.

[3]抽出した文を加工するか, あるいは命題単位で抽出したテキスト.

[4]製作者は「要約支援システム」と呼んでいる

あるという立場から, 要約処理への使用は循環論的であると考えられるので, 本手法では利用しなかった.

以下, 本章は 3.2節で, *GREEN* のシステム構成を述べる. 3.3節から 3.6節で, 要約文の選択, 一文内で修飾語を削減することによる文長の短縮法, 要約文章の段落分け, のシステム各部の詳細を述べる. 3.7節では, アンケート調査に基づき *GREEN* を評価する. 3.8節では, 大量の要約文生成で明らかになった問題点や, 得られた知見を紹介する. 本章では要約実験対象として日本経済新聞の社説を用いた. 本章での例文, 要約例は, [例文 3-7] ～ [例文 3-8] を除いてすべて 1990 年 9 月と 1990 年 11 月の同社説から引用したものである.

3.2　システム構成

要約システム *GREEN* は Sun SPARC Station I 上で, Perl 言語を使用して作成されている. システムは, 以下の五つの部分からなる. 以下では要約文として使用することを「採用する」と表現する.

形態素解析部 解析方法は, 文節数最小法を基本として, いくつかのヒューリスティックスを取り入れている. これは, 第2章で使用した Lisp 版の形態素解析部を移植して, 改良を加えたものである [Yam91].

辞書は角川類語新辞典 [Oon81] に, 固有名詞及び機能語を追加したものである. 形態素解析を行うと同時に, 各語について同辞典の分類番号を調べておく.

要約文選択部 すでに採用した文の文長の合計と, あらかじめ設定した目標要約率とを比較しながら, 新たに採用する要約文を選択していく. 要約文として採用された文に対して, 手がかり語による結束処理, 及び省略処理を行い, 条件に該当する場合, その前文も要約文として採用する.

文要約解析部 要約文として採用された文に対して, 修飾句を削減することにより文レベルの圧縮を行う.

段落分け解析部 第 2章で述べた方法を用いて, 原文の意味段落にそって段落分けを行い, 要約文章を生成する [Yam91].

3.3　要約文選択

3.3.1　見解文と現象文

論説文章中の文は, 著者の主張, 意見, 希望などを述べた文と, その主張, 意見, 希望などを述べるために必要な出来事, 事実, 現象を述べた文の二種類に大別することができる[5]. 例えば, 以下の [例文 3-1] は現在の状況を事実として述べた文, [例文 3-2] は筆者の意見を述べた文である.

[例文 3-1] 地球温暖化の防止を目指してジュネーブで開かれていた第二回地球気候会議が終わった。 [13/Nov/1990]

[5] 文献 [Nag86]:p.134 に, 類似した概念の記述がある.

[例文 3-2] 今回の会議は、来年二月から始まる温暖化防止のための条約作りの基礎になるだけに、目標が不明確のままに終わったのは残念である。[13/Nov/1990]

以下では, 著者の主張, 意見, 希望などを述べた文を「見解文」, 出来事, 事実, 現象を述べた文を「現象文」と呼ぶ. 文章中のすべての文は, 見解文か現象文のどちらかに属すると仮定する.

3.3.2　見解文の抽出

日本語の文章から見解文を抽出するためには, 文末表現に注目することが有効である. 例えば, 「〜が必要である」「〜すべきである」などは見解文に特徴的な文末表現である. *GREEN* では, あらかじめ作成した見解文の文末の典型的パターンとのマッチングを行うことにより, 近似的に見解文を抽出する. 表 3.1に, 見解文の文末パターンを示す.

表 3.1: 見解文の文末パターン

文末表現	「〜求められる」「言うまでもない」 「〜と思われる」「〜だろう (か)」 「〜と言える」「〜のではないか」 「〜したい」「ほしい」「〜と考える」 「なければならない」「〜ずにはおかない」 「気になる」
単語	「大切」「必要」「期待」「残念」「はず」 「注目」「べき」「歓迎」「課題」「危険」

3.3.3　冒頭文と最終文

文章の冒頭文は, 前提を全く持っていない読者 (聴者) に, はじめて著者 (話者) の持つ情報を伝達して, 話題に関する情報を共有するという重要な役割を果たす文であり, 文章要約においても原文章の冒頭文は重要な役割を果たすと考えられる. また文章の最終文は, 著者が話を締めくくる必要があるため, 文章の他の部分と異なった, 何らかの特別な意図を持って書かれた文と考えることができる. このため, 冒頭文と同様に重要であり, 要約作成時にも欠かせない重要な文と考えられる.

文献 [Sak89] にも, 「要約文の作成では, 一般的に原文の冒頭文と最終文の重要性が高く, その中間文において思いきった圧縮が行われることが多い」 (p.138) と述べられており, 本研究での考察と一致する. 本章の要約システムでは, 文章の冒頭文および最終文を重要視する.

3.3.4　文章の総括方式と見解文の位置

一般に文章の統括形式は, 以下の 4 つに分類できる [Iti78].

1. 冒頭で統括するもの (頭括式)
2. 結尾で統括するもの (尾括式)
3. 冒頭と結尾で統括するもの (双括式)
4. 中ほどで統括するもの (中括式)

実際の新聞社説の観察では, 中括式や純粋な頭括式はまれであることから, 文章の最終部分では, 文章の主要な結論が述べられていることが多いといえる.

これらの考察に基づき, 冒頭, および最終の見解文を除いた中間の部分に存在する見解文では, 主要な結論が述べられる最終の見解文と距離的に近いところにある, より文章の後半部に出現する見解文の方が相対的に重要であると仮定し, 要約文選択のヒューリスティックスとして採用した.

3.3.5　段落内構造

段落内構造に関しても, 基本的には文章の構造と同じモデルを考える. つまり, 各段落の冒頭文は, 新しい主題に関する前提を持っていない読者に, 著者の持つ情報を初めて伝達する役割を持つので, その段落を代表する文である場合が多いと考えられる.

3.3.6　結束性解析

GREEN では, 第 2章で行った研究 [Yam91] での考察などに基づき, 近似的な結束性の処理を行い, 要約文選択に用いている.

手がかり語による結束性

2.2節では, 文頭に出現する指示語 (これ, その, など) や接続詞 (そして, しかし, など) などの語句を「手がかり語」と定義し, これらの語が文頭に出現した場合に, 前文との強い結束性を示す場合が多いことを示した. このことから, 文頭にこれらの語を持つ文が要約文章に単独で出現した場合, 原文で存在していた強い結束性を要約文に反映させることができず, その結果として要約結果を不自然なものにしてしまう. そこで, 本研究ではこれを回避するために, 原文で保持されていた結束性を要

約結果にも反映させることにした．具体的には，文頭に指示語や接続詞などの語句を持つ文を要約文として選択する際には，同時にこの前文も採用する処理を行う．また，前文の文頭にも「手がかり語」を含んでいる場合には再帰的にさらに前文を要約文として採用する．

例えば，以下に示す [例文 3-3] では選択する文の文頭に接続詞「しかし」が，[例文 3-4] には選択する文の文頭に指示語「その」があるため，それぞれその前文も要約文に含める．

[例文 3-3] 夏場以降は一転して中東情勢に関心が移っている。しかし、構造協議で示された課題はイラク問題の発生で後退したわけではない。[9/Sep/1990]

[例文 3-4] 課徴金の引き上げについては具体化を約し、独禁法改正法案を次期通常国会に提出することになっている。その措置が形式的なものに終らぬように、議論を広めるべきだ。[9/Sep/1990]

省略による結束性

文の要素の省略は，照応よりも扱いが難しい．そこで，*GREEN* では広義の主語 (主格を表す格助詞，またはとりたて詞 (例えば，は，こそ，さえ) が後続している句) が省略されている文は，その文単独で要約文中に採用されても意味の把握が難しいと判断し，前文も採用する[6]．以下の 2 例文では，それぞれ「政府，与野党が」(主格)，「約二十六億円を投じて地方公共団体に電気自動車を普及させようという政策は」(とりたて詞の後

[6]照応の場合と同様に，前文にも主語が存在しない場合は再帰的にさらにその前文を採用するが，実際にはこのような場合はほとんどない．

続する句) が省略されているので, 2 文目が採用される時には, 1 文目も採用される.

[例文 3-5] 放置しておけば大きなツケが残るのは目に見えている。一刻も早い対応を望む。[20/Sep/1990]

[例文 3-6] 日本全国で走っている自動車の数を考えると、大気汚染を防止する直接的な効果は皆無である。この点、割箸と非常に似ている。[8/Sep/1990]

3.3.7　要約文選択アルゴリズム

以上の要因を組み込んだ *GREEN* での要約文選択アルゴリズムを以下に示す. また, 主要語, 要約率は以下のように定義する.

[定義 1] 「主要語」とは, 角川類語新辞典 [Oon81] に掲載されている語のうち, 大分類の番号が {0,5,7,8,9} である語, 及び固有名詞である[7]. 複数の意味分類に含まれる多義性のある語については, その一つが前述の大分類に含まれていれば, 主要語とする.

表 3.2に角川類語新辞典における大分類とその見出しを示す. 本研究では, 記事を内容を表現するための中心的語彙は体言, つまり名詞であると考えた. そのため, 同辞典の掲載語のうち用言の割合が比較的多いと思われる大分類を主要語からはずすため, 前述のような定義を行った.

$$\text{要約率}(\%) = \frac{\text{要約文長}}{\text{原文章の(原文のままの)文字長}} \times 100 \qquad (3.1)$$

[7]実際にはさらに一部の小分類に属する語, 及び一部の語を除いてある.

表 3.2: 角川類語新辞典の大分類

大分類	見出し	大分類	見出し
0	自然	5	人物
1	性状	6	性向
2	変動	7	社会
3	行動	8	学芸
4	心情	9	物品

$$\text{要約文長} = \sum_{\text{採用した文}} \text{単文要約処理後の文字長} \tag{3.2}$$

要約文選択のアルゴリズム：

Step 0. 文章の冒頭文, 及び最終文を採用する.

Step 1.1. 各段落冒頭文に対して手がかり語の検査・省略の検査を行い (詳細は 3.6節を参照), 原文の段落を意味段落に再構成する.

Step 1.2. 第一意味段落の全文と各段落の冒頭文の中で, 文章冒頭文に含まれる主要語が主語になっている文を採用する.

Step 1.3. 最終意味段落の全文と各段落の冒頭文の中で, 文章最終文に含まれる主要語が主語になっている文を採用する.

Step 2. Step 1 が終了した時点で, 要約率が $(\alpha + \delta)\%$ 以上ならば, 冒頭文, 最終文以外の採用した文すべてを未採用にする. ここで, α は要約率の目標, δ は許容範囲パラメータである.

Step 3.1. まだ採用されてない見解文のうち, 文章の最も後ろにある文を採用する.

Step 3.2. 手がかり語による結束性の解析を行い, 要約文の文頭に手がかり語が出現する場合は, その前文を要約文に採用する.

Step 3.3. 要約文に対する省略解析を行い, 広義の主語 (主格を表す格助詞, またはとりたて詞が後接している語) が省略されている要約文に対しては, その前文も要約文として採用する.

Step 4. 要約率が α% 未満ならば, Step3.1 へ. α% 以上ならば, 採用した文を意味段落に沿って出力, 終了. ■

3.4　文要約解析

文中の修飾句を削減することにより, 一文内での要約を行う. 文の中心内容は, 文中の修飾句の削減による影響を受けない. 本研究では, (1) 二重修飾・多重修飾, (2) 固有名詞への修飾, (3) 例示の三通りの場合に, 連体修飾句を削除する.

3.4.1　二重修飾・多重修飾

ある名詞を修飾する連体修飾句には, 表 3.3に示す種類が考えられる [Ter81][8]. そこで, 実際の文によく見られる「二重修飾」及びそれを一般化した「多重修飾」を, 以下のように定義する.

[8]ただし, 文献とは一部の品詞名を変更している.

表 3.3: 連体修飾の種類

文法的性質	例
こそあど詞連体形	この話
連体詞	ある話
形容詞の現在形・過去形	おもしろい話
形容動詞の連体形・過去形	変な話
名詞＋連体助詞「の」	昔の話
名詞＋格助詞＋「の」	昔からの話
名詞＋取り立て詞＋「の」	ここだけの話
副詞＋「の」	突然の話
節 (被修飾名詞が修飾節の格要素)	私が聞いた話
節 (被修飾名詞が修飾節の格要素でない)	子狸が少年と仲良くなる話

[定義 2] 一つの名詞に対して，上の例に示したような要素が二つ以上修飾している状態を「多重修飾」と呼ぶ．特に，二つの要素が修飾している状態を「二重修飾」と呼ぶ．複合名詞 (地域紛争など) は，名詞が名詞を修飾しているとみなす．

本研究では，名詞の二重修飾があった場合，前方の修飾要素を省略しても意味は大きく変化しないことが多いと考え，要約文生成の際には前方の修飾要素を省略する．同様に，多重修飾の場合には，最終の修飾要素を残して，残りを省略する．

例として，「おもしろい昔の話」は「昔の話」に，「突然のうれしい話」は「うれしい話」に，「私が聞いた変なおもしろい話」は「おもしろい話」のように省略を行う．ただし，後半の修飾要素が，名詞 (＋格助詞または取り立て詞) ＋「の」である場合には，意味的なあいまいさが生じる．

[例文 3-7] 私が聞いた作家の話 ...

[例文 3-8] 私がインタビューした作家の話 ...

上の二つの例では，形態上は同一であるが，前者は「私が聞いた」が「話」に，後者は「私がインタビューした」が「作家」に，それぞれかかる．しかし，以上のことを形態情報だけで判断することは不可能である．

ここで，被修飾名詞「話」を基本に考えると，前者の例は「私が聞いた」「作家の」の二つが「話」にかかり，後者の例は，「作家の」だけが「話」にかかって「私がインタビューした」は「作家」にかかる．ここで，両者に共通するのは，「作家の」は「話」にかかる，という点である．

以上より，例えば「私が V した作家の話」（「V した」は任意の動詞）という表現を短縮する場合，

1. 修飾要素をすべて取り除いた「話」だけでは漠然としすぎて一般に意味がつかめない.
2. 「私が V した」は「作家」と「話」のどちらにかかるか不明.
3. 「作家の」は必ず「話」にかかる.

という理由から,「私が V した」を削除し,「作家の話」としている. また一般に, 修飾節のほうが名詞＋「の」よりも文字数が多く, このため修飾節を削除することによる要約の効果が高いことも, 修飾節削除の理由となっている.

実際に出現した例文を以下に示す. ただし, 文中の [···] の部分は計算機が削除可能と判断した修飾要素である.

[例文 3-9] 要綱素案で目につくのは、[海部首相の要請にこたえて去る四月末、選挙制度審議会がまとめた] 衆院選改革の答申と大きく異なることだ。[14/Nov/1990]

[例文 3-10] 統一を実現するうえで最大の難関は、米英仏とともに [[統一問題やベルリンの地位変更に関する] 国際法上の権限を留保している] ソ連の承認をどう取りつけるかだった。[14/Sep/1990]

[例文 3-10] のように, 該当する削除可能候補が複数ある場合は, 削除文字列の長い方を採用する.

3.4.2　固有名詞への修飾

修飾の用法は, 一般に限定修飾と, 非限定的修飾に分類することができる [Ter91]. 個々の修飾語句がこのどちらで使用されているかの判断は, 必ずしも容易ではないが, 本研究では, 非限定に用法が限られる固有名詞への修飾を扱う. ここでいう固有名詞とは, 一般的な意味の固有名詞の

他, 固有物を示す一般名詞 (例えば, わが国) も含めて考える. *GREEN* では, 固有名詞にかかる連体修飾節は限定機能を果たさないので削減可能と判断し, 単文要約処理の際, 削除する.

[例文 3-11] [絵の具に油を混ぜ表面をニスで覆う] 西欧絵画に比べ、[ニカワが下に沈んで絵の具がむき出しになる] 日本の絵は、長期的に照明下に置くことはできないのである。[16/Sep/1990]

[例文 3-12] [消費と設備投資をけん引車とする内需と、堅調な輸出に支えられて順調な拡大を続けてきた] 日本経済に、二つの警戒信号がついた。[3/Sep/1990]

[例文 3-13] [党より人で投票する傾向の強い] 日本の選挙の実情からすると、候補者への投票がそのまま政党への投票とみなされる一票制による小選挙区比例代表制は、[地域密着型の選挙を得意とする] 自民党候補に有利と考えられる。[14/Nov/1990]

[例文 3-11] ～ [例文 3-13] に示す例は, いずれも固有名詞を直接修飾していない. すなわち, 固有名詞の前方にある修飾要素 (例えば, [例文 3-12] では「消費と ··· 続けてきた」) が, 固有名詞 (日本) を修飾しているのではなく, その後方の複合名詞 (日本経済) を修飾している. *GREEN* では, その後方の複合名詞 (日本経済) に固有名詞が含まれていることを利用して, その名詞 (日本経済) も固有名詞に準ずる取り扱いを行い, その前方の修飾要素の削除を行う.

3.4.3　例示

「··· などの」「··· といった」のような例示も, 広い意味の修飾語と考えることができる. 修飾としての例示は, そのほとんどが非限定的修

飾用法と考えられ, 削除しても意味的に変化が生じないと近似的に仮定し, これらの語句が文内に出現した場合に, その例示部分を削除する.

[例文 3-14] 警視庁では、[[最高時には三万七千人という] 空前の警備体制を決めるなど、] 過激派の動きに対応してきた。[3/Nov/1990]

[例文 3-15] 百二十ヶ国の政府代表が閣僚宣言を採択したが、[焦点の二酸化炭素など] 温室効果ガスの排出量を規制する具体的な目標値を設定することはできなかった。[13/Nov/1990]

[例文 3-14] では, 二重修飾の削除対象 (「最高時 ... という」) と, 例示の削除対象 (「最高時 ... など, 」) が重複している. この場合, より広い範囲を対象とした例示部分を削除対象とする.

現在のシステムで, 例示の対象としている語句は「など」「などの」「といった」「のような」「のように」の五つである.

3.5　係り受け解析

前節で述べた修飾句の削減を実現するためには, 修飾句の認定が必要である. しかし修飾, 被修飾の関係を解析する係り受け解析はまだ研究が進められている段階の解析であり, 十分な精度で行える手法が確立しているとは言えない. そこで本手法では, 形態素解析のみを行った後に, 表層的な手法, つまりある特定の語句の出現によって近似的に修飾句の切れ目と認定する方法を行った.

表 3.4に, 今回の手法で修飾句の切れ目と近似的に認定した語句を示す. ただし実際の処理では, 接続助詞「が」は格助詞と区別するため読点が後続している場合に限定したり, 助詞「では」は「... ではない」と

なる場合を除いたりするなど, さらに細かな設定を行っているが, その詳細は省略する.

表 3.4: 修飾句の切れ目と認定する語句

要素	備考
句点, 読点	読点は, 動詞または「が」に接続している場合
係助詞	は, こそ, さえの 3 語
接続的語句	接続詞, 接続助詞

3.6　段落分け解析

GREEN では, 原文の段落をそのまま意味的な段落, つまり「一つの主題を持つ文の集合」と考えるのではなく, 前述した結束性に関する処理を行うことによって意味段落を再構成している. 意味段落は, 以下の手順によって原文章の一つ以上の段落の集まりで構成される. すなわち, 原文の段落の冒頭文について, 手がかり語の検査 (冒頭に「手がかり語」を含むかどうか) 及び省略の検査 (主語が省略されていないかどうか) を行い, 少なくとも一方が該当する場合, その段落はその前の段落とつながっている (一つの意味段落を構成している) とする. この処理を第 2 段落以降の全段落の冒頭について行い, 最終的にできた意味段落を対象にして, その後の様々な処理を行っている. 以下の例では, 原文中の手がかり語 (「こう (した)」) を検知し, 原文にある直前の改行を削除して, 要約文章を作成する.

[例文 3-16] ··· 一般市民にも危険が迫ったことをうかがわせる。(改行) こうした過激派のゲリラ活動は、かねて十分予想されていたことであった。 ··· [3/Nov/1990]

3.7　評価

要約システム *GREEN* の有効性を評価するために, 被験者 18 名に対して要約文章の評価に関するアンケート調査を行った.

調査はそれぞれの被験者に対し, 各 5 編の社説と *GREEN* による要約結果 (ただし, 目標要約率 $\alpha = 25\%$, および $\delta = 5\%$ に設定して出力した. 調査対象要約文章とその原文を付録 A に示す.) を提示し, 以下に示す 3 項目について, それぞれ 0 ～ 5 までの数字 (整数に限らず, 小数を含んだ数も許す) で回答してもらう, という形式で行った.

1. (社説の原文を読まずに) 各要約文章のみを独立した文章として読んだ時に, 自然かどうか
2. 社説本文, 及びそのタイトルと比較した時に, 原文で重要と考えられる部分を抽出しているか
3. 文内で修飾語を省略している部分について, それが適切な省略かどうか

以下では, この 3 つの質問とその回答結果について述べる.

3.7.1　要約文の自然さ

要約文章は文章全体としてまとまりのあるものでなければならない. そこで最初の質問では, 要約文章を独立した文章と考えた時の自然さを被験者に判断してもらった. なお, 被験者が評価する際の判断基準を以下のように設定し, 被験者に提示した.

5 点 ほぼ自然である. つまり, このような文章を書く人間もいると考えられる.

0 点 非常に不自然である. つまり, 文章全体としてのまとまりがなく, 文章として何を意味するのかほとんど理解できない.

表 3.5に調査結果の平均と, 満点 (5 点) の評価をつけた人数を示す. 全体の平均点は, 3.78 という比較的良好な評価が得られている. 特に文章 A について高い評価点が得られた.

質問項目とは別に被験者に感想を求めたところ, 今回の要約率が 25 % となっていることに関連して, 要約文の中に多くの内容を盛り込み過ぎて, その結果として内容にまとまりを欠いている点を指摘する声があった. また, 要約された文章の段落間, 及び文間に接続詞がないために読みやすさに欠けるという意見が多かった. この事実から, 日本語においては文章内の結束性の維持に接続詞が重要な役割をしていることが再確認される. また, 要約文章は相対的に文章の情報量が減少しているため, その質を高めるためには, 接続詞などを加えることで文章全体の結束性を補う必要があることを示唆しているものと考えられる.

表 3.5: 要約文章の自然さの評価

文章	A	B	C	D	E	平均
評価	4.05	3.45	3.96	3.81	3.62	3.78
満点の人数	5	0	2	3	0	

3.7.2　要約内容の適切さ

要約文章は, それ自身にまとまりがなければならないと同時に, 原文で重要と考えられる情報を適切に抽出しなければならない. そこで, 被験者を対象に, 原文で重要と考えられる部分を *GREEN* で抽出しているかどうかを尋ねた. 評価基準及び調査結果を以下に示す.

5 点 このような抽出を行う人間もいると考えられる.

0 点 全く *at random* に抽出したものとそう大きな差はない.

表 3.6: 要約内容の適切さの評価

文章	A	B	C	D	E	平均
評価	3.81	3.39	3.89	3.61	3.78	3.70
満点の人数	2	0	2	2	1	

この調査では, 他と比較して文章 B の評価値が低かったが, 全文書の平均では評価値 3.70 という比較的良好な値が得られた. *GREEN* の行った文書 B の要約では, その第 3 段落で日本の行動計画を全文引用しているが, これについて, 要約に全文引用することの必要性に疑問を持った被験者がいた. また, 要約では第 2 段落に米ソの話題を採用している

が, これよりも, 目標値の設定がなぜ重要なのかという理由を採用すべきであるという指摘など, 目標値に関する話題を採用すべきだという意見が多かった (要約例参照).

また, *GREEN* の要約では原文の後半に要約の重点が置かれ過ぎている傾向があるという意見もいくつか聞かれた. このことから, *GREEN* では見解文を必要以上に抽出した場合に, 近似的に文章の後半部分の見解文を採用するようにしているが, この近似方法をさらに検討する必要性が明らかになった.

3.7.3　修飾句省略

GREEN では, 単文単位での修飾句の省略も試みている. この点についても, 被験者にその適切さを判断してもらった. 判断基準は以下の通りとした.

5 点 ほぼ適当である. つまり, 省略された部分は, 原文の中では重要性の低い部分であり, このような省略を行うことは妥当だと考えられる.

0 点 ほぼ不適当である. つまり, 全く at random に省略したものと大差はない.

表 3.7: 修飾語省略の適切さの評価

文章	B	C	D	E	平均
評価	3.28	4.02	3.02	3.72	3.64
満点の人数	0	4	1	1	

被験者の判断した 4 文章[9]のうち, 特に文章 C で高い評価が得られた. ただ, *GREEN* で使った修飾語省略のヒューリスティックスの適切さに関しては, 今後の課題として検討すべきであるとの意見がいくつか出された.

文中で, 構文解析を行わずに修飾・非修飾の関係, あるいは主述関係を完全に特定することは不可能である. 現在のシステムはこの処理を近似的に行っているため, 以下の例のような連体修飾節の認定に問題が生じる場合がある.

[例文 3-17] だが、各国の意見が割れたまま会議を開くのは [危険だとする] スペインの主張はうなずける。 [11/Sep/1990]

また, 以下に示す例のように, 文法的には正しいが連体修飾節が限定用法を示す場合, 削除すると意味的におかしくなる.

[例文 3-18] インドシナ難民の受け入れが十二万人を超し米、加に次ぎ、[人口比でみた] 日本語学習者数が韓国に次いで多いという事実はそうした方向の反映だろう。 [21/Sep/1990]

3.7.4　主語のない文

文章では, さまざまな談話メカニズムにより結束性が保たれている. このような状況下では主語などの省略は自然に行われているが, [例文 3-19] のような主語のない文を要約文章として抽出してしまうと, 要約文章の結束性は原文のそれよりも弱いために人間は主語のない文にとまどいを感じ, また場合によっては文の意味が理解できなくなってしまう. 今回のアンケート調査でも, 主語のない文に不自然さを感じた被験者が何

[9]文章 A では修飾句省略の処理が行われなかったため, 評価の対象から外している.

人かいた. 要約文章作成では, 接続詞などを補うことと同様に, 主語などの省略語を補うことも検討する必要がある.

[例文 3-19] 経済のメカニズムを働かせる試みで、うまく機能するかどうか注目したい。[16/Nov/1990]

3.8　議論

ここでは, 機械処理した大量の要約結果の考察から得られた知見や明らかになった問題点などを述べる.

- 本研究では, 原文の結束性のまとまりをくずすことなく, そのまま要約文にも反映させることで, 要約文の読みやすさの維持に努めた. その結果, 文を芋蔓式に採用してしまい, 文数の短縮にならない例が見られた.
- 例えば文章の冒頭文に, より抽象的で, その結果文章中に多用される語 (例えば「経済」「事件」) が含まれている場合, 有効な要約文抽出が出来ないことがあった. またこのことは, 高頻度の語が (文章の分野特定には有効でも) 必ずしも要約で重要な語にはならないことも示している.
- 文章の冒頭文が例えば「文化の日. ([3/Nov/1990])」のように, 極端に短い場合, あるいは, 文章冒頭文が極端に長い場合, 何らかの比喩から文章が始まっている場合などは, 本手法が有効に機能しない.
- 比較的短い文で構成されている文章に対して, 本手法が特に有効であることが観察された. この理由としては, 文が短いと, 要約率の微調整が容易になること, 重要文の抽出の精度が上がることが考えられ

る．このことより，要約の前処理として，文の短文への分割，あるいは文のパラフレーズを行うことが有効であると考えられる．

- 論説文中にある現象文は，以降の記述内容の特定のために周知の事実を述べる場合と，未知の事実そのものを紹介するために事実を述べる場合とに分けられる．このうち，前者の要約文としての重要性は低いが後者のそれは高い．これの判別は，構文や修辞関係の情報を使用しただけでは不可能なことから，現実世界の一般的な知識が必要と考えられる．

次に，要約率が要約結果に与える影響について検討する．一般に，文章の要約率は原文の文長や要約の目的などによって変化すると考えられるが，多くの場合にその要約率は 10% から 50% 程度と考えられる．本章の手法がどの程度の要約率に対して有効に機能するかの例として，前節の評価実験に使用した文章 E を例にして，要約率を 5% から 50% まで，5% 刻みで変動させた場合の要約結果を付録 B に示す．

付録 B の要約例が示すように，要約率が低い例では文章の冒頭と最終の文が選択される結果になった．これは文長が同じ程度であれば，本手法においてはどのような文章に対しても同じ結果になる．このため，これらの要約率に対しては本手法は有効に機能しないが，20 文程度で構成される文章を 2, 3 文にまとめる作業は一般的ではないと考えられる．

また，付録 B の要約例では，目標要約率が 35%, 40%, 45%, 50% の要約結果が一致している．これは，目標要約率が 35% の時点ですでに 50% 以上の文を採用してしまったために同じ要約結果となったものである．このように，本実験の対象である文長がほぼ 20 文程度，あるいはそれ以下の文章では，要約率の目標と結果との差が大きくなりやすい傾向がある．

3.9　まとめ

本章では日本語の論説文を対象にした要約文章作成実験システム *GREEN* を紹介した．*GREEN* は現在論説文だけを対象としているが，見解文抽出に関連する処理以外の処理は，他の種類の文章の要約にも十分に利用できるものである．

GREEN の要約文の品質の評価をアンケート調査により行ったが，アンケート結果の中には，出力された文章に最小限の後編集をすることによって，人間が行った要約と変わらない程度の質の高い文章となる要約文が多い，という意見があった．このことから，*GREEN* では論説文からの重要な情報の抽出は比較的うまくいっているが，より「まとまりのある文章」，つまり，結束性と首尾一貫性をより強く持った文章にするための編集機能を強化することが *GREEN* の今後の課題であると考えられる．ただしこのために必要な処理である，接続詞や主語の補完の実現には，より高度な文章の解析が必要であり，困難な問題が多く存在する．

第 4 章

関連テキストを利用した重複表現削減による要約

4.1　はじめに

本章では, 今後需要が高まってくることが予想される複数テキスト, 特に複数の新聞記事に対する要約について述べる.

最近の計算機とネットワークの進歩により, 情報検索がより一般的に, より身近なものになりつつある. 例えばある長期間に及ぶ事件の記事や一連の政治的問題について述べられている記事を検索対象にして検索を行うと, 一般には検索結果として複数の記事が該当する. それらの記事の内容には当然ながらお互いに関連があり, 記事間で重複している記述も多い. このような中で, 検索者がこれらの検索結果すべてに目を通すのは効率的とはいえず, その結果, 複数記事から重複部分を除去した記事の需要が発生する. 今後このような状況はさらに増加することが予想され, 複数記事の要約処理が情報検索関連の重要な処理の一つとなると考えられる.

図 4.1に, 本章で行う要約の全体像を示す. 複数記事において, 重複部分をどう把握するかという問題は, 単一記事の要約とは異なる固有の問

題である．また，一般的には要約記事の記述順序なども問題となるが，本章では要約記事の記述順序は元記事の時間順とし，所与のものと仮定する．また本研究では，複数記事の中で後続記事の重複部分の除去による要約をその研究目的とし，冒頭記事の要約はここでは考えない．また，対象言語を日本語とし，以下では関連した二記事に対する後続記事の要約に問題を限定する．

第3章で述べたように，単独の文章を対象にした文章の短縮化処理は特に抄録という形で多く研究が行われてきている．また本論文の第3章では，論説文を対象にして日本語での様々な表層的特徴を取り入れ，生成テキストの結束性も考慮した要約を試みた [Yam95a]．しかしながら，これらはいずれも単独のテキストを対象にしたものであり，本章で対象とする，複数テキストに対する要約を試みたものではない．複数文章を対象にした文章の短縮化処理の試みは，従来ほとんど知られていない．

図 4.1に示すように，本章では連体修飾語，類似節，名詞句の言い替えの３点に着目し，これらを利用した要約手法を提案する．アプローチとして，現時点での自然言語処理の諸技術の限界をふまえて，第3章で述べた *GREEN*[Yam95a] などと同様に形態素解析のみの表層的な情報から要約テキストを生成することを目指す．また，本手法のプロトタイプ版を計算機上に実現し，実際の新聞記事に対して実験を行った結果について述べる．

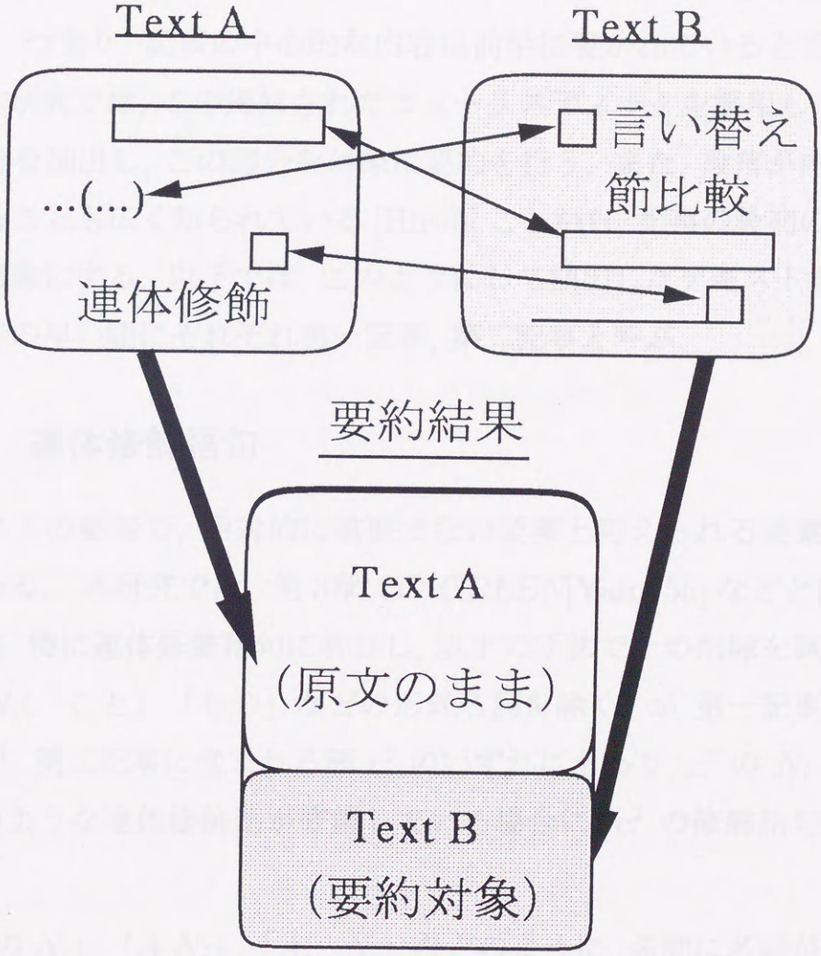


図 4.1: 複数テキストの要約処理

4.2　要約手法

4.2.1　新聞記事とその要約

一般に，新聞記事はトップダウンに記事が書かれていると言われる [Hir90]．つまり，記事の中心的な内容は前半に書かれていると言われている．本研究では，この周知されたヒューリスティックを採用し，記事の前半部分を抽出し，この部分を対象に要約を行う．また，段落が内容の単位となることも広く知られている [Hir90] ことから，記事の最初の段落を処理の対象にする．以下では，このようにして抽出したテキストを，その掲載日時の早い順にそれぞれ第一記事，第二記事と呼ぶ．

4.2.2　連体修飾語句

テキストの要素で，相対的に重要でない要素と考えられる要素は修飾語句である．本研究では，第 3章での *GREEN*[Yam95a] などと同様に，修飾語句，特に連体修飾語句に着目し，以下の手法でその削除を試みる．

名詞 N_i(「こと」「もの」などの形式名詞を除く) が 第一記事に含まれる節 c^1，第二記事に含まれる節 c^2 のいずれにもあり，c^2 の N_i に対して以下のような連体修飾語が修飾している場合に，c^2 の修飾語句を削除する．

- 「A の N_i」「A N_i」「A、N_i」などのように，名詞に名詞が修飾している形式．ただし，A は名詞とする．
- 「A である N_i」「A となっている N_i」「A になっていた N_i」など，名詞に判定詞，およびそれに準ずる語句が修飾している形式．ただし，これらの連用形および条件形は，明らかに N_i を修飾していな

いので, 除外する.

- 「V する N_i」 (ここで, 「V する」は動詞. サ変動詞を含む) のように, 名詞に (連用形および条件形を除く) 動詞が修飾している形式.

表 4.1に, 上の 3 種類のパターンについてそれぞれ例を示す. この処理は, 第一記事に含まれている名詞はすでに読者に周知されている名詞であり, 第二記事で修飾語句と共に使用された時は冗長になりやすい, というヒューリスティックを使用している. また, 「広い海」などの形容詞による連体修飾は, 前述した三つのパターンの修飾よりも名詞に対する限定の要素が強いと考え, 削除の対象から外した.

表 4.1: 対象とした連体修飾語の例

分類	例
名詞修飾	輸入の拡大, 環境問題, 「歌手、マドンナ」
判定詞修飾	ユーザーである業界, 焦点となっている問題
動詞修飾	落ち込む可能性, 勉強する学生

第 3章では, 多重修飾, 固有名詞への修飾, 例示による修飾の 3 種類の連体修飾に対してその削減を試みた [Yam95a]. 今回の複数テキスト間の要約では, 重複部分の削減という視点での要約を目的としているため, これらの 3 要素による修飾句の削除は取りいれていない.

4.2.3　節照合処理

本章で提案する節[1]照合処理は, 同一の動詞を持つ節 c^1 と c^2 が, 同じ助詞を含む文節に異なる内容を含んでいないかどうかを判定する. もし, 節 c^1 と c^2 がほぼ同一の内容, あるいは類似した内容であるならば, それらの節は冗長であり, 要約の際にはどちらかの節 (普通は c^2) を削除すべきである. 本研究での照合処理では, 第一次近似として動詞や助詞のマッチング処理のみを行い, 構文解析などは行わなかった.

第一記事中の節 c^1 と第二記事中の節 c^2 について, 以下の処理を行う.

1. 節 c^1 と 節 c^2 で類似した動詞 v_i^1 と v_j^2 を探す. ここで, 「類似した」とは, 角川類語新辞典 [Oon81] において, v_i^1 と v_j^2 が同一の末端分類に含まれていることを意味する. また, 動詞には「する」を除くサ変動詞 (「勉強する」など) 及びサ変名詞 (「勉強」) などを含むが, 名詞が後続するサ変名詞は動詞の要素が弱いと考え, これに含めない. 以下同様.
2. 動詞 v_i^1 の一つ前の動詞 v_{i-1}^1 の次の語から v_i^1 までの語句を動詞 v_i^1 にかかる要素とする. ここで, v_{i-1}^1 の直後の語が助詞, または助詞＋「、」であった場合, それらの次の語句から v_i^1 までを動詞 v_i^1 にかかる要素とする. v_j^2 についても同様.
3. v_{i-1}^1 から v_i^1 までを助詞別に分類する. ただし, 助詞「の」は分類の対象に入れない. また, 助詞「に」と「へ」は同一視する. 例えば, 「小学生の太郎が新しい学校に行く」ならば, 助詞「が」に「小学生の太郎」が対応, 「に」に「新しい学校」が対応する要素とする. これは格関係を近似的に把握するために用いる. 「は」「も」などの

[1] ここでは動詞から動詞までの文字列を「節」と近似的に呼んでいる. 実際には節と命題の中間的な単位と考えられる.

副助詞は, 分類はするが, 以下の処理では使用しない. v_j^2 についても同様.

4. v_i^1 と v_j^2 のそれぞれの同一の格要素が同一かどうかを判断する. ここで, 同一かどうかは (接尾辞を除く) 最後の語で判断する. 例えば, 「小学生の太郎」と「その隣の太郎たち」は最後の名詞「太郎」が同一であるので同一とみなす.
5. v_j^2 にかかるすべての格要素について, v_i^1 でも同一であるかまたは対応する格要素が v_i^1 にない場合, c^2 は省略可能と判断し, v_{j-1}^2 の直後から v_j^2 までを削除する.

ここで, 助詞を持たない動詞は対象外とする. この理由は, 補助動詞 (例えば「食べに行く」などの「行く」や「参加する」などの「する」) を削除処理から除外するためであり, また冗長ではないため, 対応する節と比較できないためでもある.

以上の手順の適用例を図 4.2に示す.

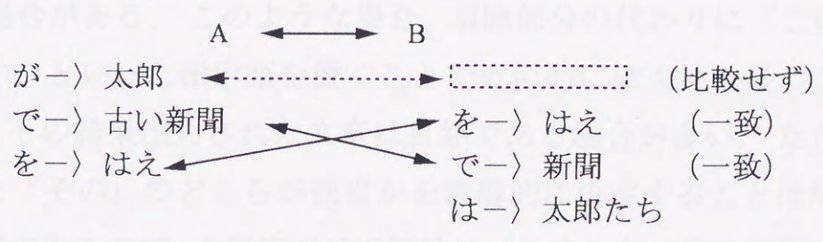
4.2.4　用語の言い替え

新聞記事においては, 用語の言い替え, あるいは別名の使用などが行われる. これは, ある語句が例えば「国連カンボジア暫定統治機構」などのように長い場合には頻繁に起こり, 「UNTAC」などのように短い別名で呼ばれる. 前後 2 記事の両方にこのような語が使用された場合, 両記事に言い替えを行っている箇所が存在するはずであり, このような場合, 二度目以降は不要である. 本研究ではこの現象, 特に括弧を伴った言い替えに着目する. もし第一記事で名詞 (またはその連続) が括弧を伴っている場合, 例えば「$NS_1(NS_2)$」 (ここで NS は名詞列) のような場合

Step 1: 類似した（同一の）動詞を探す。
A: 花子がはえを見つけ、太郎が古い新聞ではえをたたいた。
B: 太郎たちははえを新聞でたたいて殺した。

Step 2: 動詞にかかる語句を対応づける。
A: 見つけ (見つける): 花子がはえを
たたいた(たたく): 太郎が古い新聞ではえを
B: たたいた(たたく): 太郎たちははえを新聞で
殺した (殺す):

Step 3: 格によって語句を分類する。
Step 4: 対応する文で比較する。



「は」は格助詞ではないので考慮対象外とする。）

図 4.2: 節照合処理の例

は, NS_2 が NS_1 の言い替えであると判断し, 第二記事に NS_1 があった場合にそれをすべて NS_2 に置き換える. また, 置き換えた語の多くには「(NS_2)」が後接しているが (つまりここで言い替えが行われている) これらは当然取り除く.

4.2.5　削除後の処理

連体修飾語や節の削除後の文章が自然さを保つために, 以下の処理のいずれかを行う.

後続する数語を削除 もし削除した語句に助詞「が」などが後続している場合 (例えば「...したが」) には, 残った文章は文法的に正しくない文章となってしまうので, これらの語句を削除する.

指示語を付加 「...した問題で」における「問題」のように, 固有名詞以外の名詞に対する修飾語を削除したときに普通名詞のみが単独で残る場合がある. このような場合, 削除部分の代わりに「この」や「その」といった指示語を置くことで近似的に結束性を保つことができ, その結果出力された文章は自然である場合が多い. ただ「この」と「その」のどちらが適当かを表層的に決定することは現状では困難であるので, 本研究では近似的に「この」を一律に使用する.

何もしない 削除した修飾語が固有名詞にかかっている場合には, ただ修飾語を削除するのみでそれ以上何もしない. 固有名詞を指示語で修飾すること (例えば「このマドンナ」) は一般に不必要な場合が多いためである.

4.2.6　処理手順

対象としたのは日本経済新聞の記事である (コラムを除く). 処理手順の概要は以下の通りである.

1. 各記事について, 冒頭の段落を抽出し, 形態素解析を行う. ただし, 冒頭の段落が短い場合は, 冒頭二段落を抽出する.
2. 言い替えの判定, 節の照合処理, 連体修飾語に対する処理を順に行い, 該当する場合には削除を行う.
3. 削除後の処理として, 必要ならばさらに数語削除し, あるいは指示語を付加する.
4. 要約結果として前述の処理を終えたテキストを出力する.

4.3　実験

本章で述べた手法を計算機上で実現し, 本手法の有効性を確認するために, 簡単な実験を行った. なお, 実験を行った計算機は Sun SPARC Station I, 作成言語は Perl である. また本研究では, 京都大学／奈良先端科学技術大学院大学で開発された形態素解析システム JUMAN 2.0[Mat93] を使用した.

実験の入力記事として, 日本経済新聞を使用した. 以下の記事は同新聞の 1992 年 10 月 28 日と同年 12 月 11 日の記事である. 付録に, これらの全記事及び要約結果を示す.

記事例では, 第二記事に出現する「共同開発中」での「開発」が最初に節照合処理の対象に該当する. 第一記事では「と」と「を」の助詞が

「開発」にかかり，第二記事では助詞「により」と「で」がかかる．この例では同一の助詞に対して異なる要素を持っていないので削除の対象になり，「この M&A(企業の合併・買収) により日米で」が削除される．言い替えの処理では，「GD」と「FSX」が該当し，第二記事ではこれらの語に統一される．

連体修飾語の処理では，「GD」にかかる「共同開発の米側の担当企業だった」がまず削除され，次いでその直後の「GD から計画を引き継ぐ」が削除される．この結果名詞句「共同開発の米側の担当企業だった GD から計画を引き継ぐロッキード」は単に「ロッキード」となる．

本手法で削除の効果を評価するために，簡単な調査を行った．対象記事は日本経済新聞の 1990 年，および 1992 年の記事であり，経済，政治，国際など様々な分野の内容を含む．実験は，まず筆者の判断で類似していると考えられる記事対 15 組 (合計 30 記事，付録の要約例もこれに含む) を選び，それぞれの第 2 記事を計算機によって要約した．

表 4.2に，本手法でどの程度の削減効果があったかを，対象記事における文の位置と共に示す．なお，表では実験した 15 記事の合計の文字数と削減率 (1－要約率) を示している．表 4.2に示すように，全体で約 27% の削減効果が見られた．また，文の位置との関係では，冒頭に近い文ほど削除の割合が高いことがわかり，冒頭文においてその効果は最大となった．この結果は，対象記事中において冒頭に近い文ほどより重複した部分が多いことを示しており，これは筆者らが事前に予想した結果と一致する．

様々な例に対して実験を行った結果，形態素解析段階での誤りは存在するが，それらの部分以外ではほぼ良好な結果が得られた．また，係り受け解析を行わないことによる誤りは比較的少ない．日本語は比較的語順が緩やかであり，また既知の情報，あるいは補完可能な情報に対しては省

表 4.2: 文の位置と削除率

文の位置	第 1 文	第 2 文	第 3 文	第 4 文	合計
原文章 (文字)	1483	1070	609	160	3322
要約文章 (文字)	907	821	556	143	2427
削除率 (%)	38.8	23.3	8.7	10.6	26.9

略も頻繁に行われるが, 本研究の目的であるこれらの現象への対応については, ほぼ達成されている. また, 本システムはまだプロトタイプの段階にはあるが, 実用性を考え, あらゆる新聞記事に対して対応する.

本システムがうまく動作しない場合には, 二つの記事の扱う話題の間の関連性が低いことが原因であった. 現在は人間の判断によって関連記事を選び, それをシステムの入力としている. 関連記事の自動抽出は今後の課題である.

4.4　まとめ

本章では, 関連する話題を扱った複数のテキストを一つのテキストに要約するための基礎として, 複数のテキストに重複した部分を削除することによって要約する手法を提案した. 本章で提案した手法はまだ第一次的なものに過ぎないが, 同一の話題に対する連続した記事に対しては有効に機能する.

本研究では対象とする文章として互いに内容の関連性の深い新聞記事を採用したが, これは情報検索の出力結果などを入力とすることを想定

しているためである．この点は，第3章で述べた *GREEN* が論説文章を対象としていたのとは目的が異なるので，直ちに両手法を融合することはできない．しかしながら，*GREEN* における重要部分の抽出処理以外は複数文章の要約に対しても利用可能であり，今後の課題として残されている．また，要約の効果のさらに正確な評価，および3編以上のテキストに対する要約処理の手法についても検討を行う必要がある．

第 5 章

分類体系相互の関係を利用したテキストの自動分類

5.1 はじめに

近年, 計算機の普及と共に日本語テキストの機械可読化が進んでいる. 機械可読化されたこれらのテキストは, 文献検索等の用途のために少数の専門家によって適切に分類 (インデクス付け等) が行われ, また整理されてきた. しかし非常に大量のテキストが計算機上で利用可能となりつつある現在, 専門家による手作業による分類だけではその作業量に限界がある. このような背景の下, テキストを自動的に分類することの必要性が高まってきている.

テキストの自動分類手法には, あらかじめ分類すべき分野 (以下カテゴリーと呼ぶ) を設定し, 各テキストにこのうちのいずれかの分野を割り当てる手法 (判別分析 discriminant analysis, または教師ありの分類 supervised recognition)[Kaw92, Kes93, Yua93] と, カテゴリーをあらかじめ設定しない手法 (クラスタリング cluster analysis, または教師なしの分類 unsupervised recognition)[Tsu94] の二つに分けられる. 本章では, 判別分析による分類を対象とし, 以下ではこれを単に分類と呼ぶ.

自動分類の手法は現在までに様々な手法が提案されてきており，最近でも活発に研究が行われている (例えば [Kaw92, Gut94, Kar94]). それらの手法の中で，ベクトルモデルに分類される *term-weighting* と呼ばれる手法は，テキストに出現した単語の頻度にある重みづけを行うことによって頻度を補正する手法で，この手法は最も重要なアプローチの一つである. 重みづけの方法には，有名な *idf* (inverse document frequency)(例えば [Spa72]) や重みづけした *idf* [Tok94] などが提案されており，語間の意味的関係を用いた手法も提案されている (例えば [Sal88]).

これらの手法の多くは語の言い替えによる影響を考慮していない. しかし，既出の語を別の語で言い替える現象は任意の言語のテキストで起こり得る一般的な現象であると考えられる. そこで本章では，語の言い替えを考慮した *term-weighting* の手法を提案する. この手法は，テキストに出現した語をシソーラスで分類し，その頻度情報をカテゴリー相互の関係によって加工した特徴ベクトルを用いてテキストの分類を行う. 本手法は，シソーラスと統計情報のみを使用し，処理は簡単であり，またカテゴリーに依存した情報は統計情報から自動作成できるので，汎用性が高い.

芥子らは文脈ベクトルを用いた連想検索手法を提案している [Kes93]. この手法は，数百の出現頻度の高い語について手作業で文脈ベクトルを作成して，重要単語の文脈ベクトルを機械学習するというものである. この手法では，人手を介することで労力がかかり，文脈ベクトル作成の際に個人差による揺れが生じる可能性がある. また，ベクトルの要素となる語の選択という問題が生じる.

統計的な情報を用いた分類としては，漢字を単位にした方法 [Wat94],

名詞の共起関係を使用した方法 [Yua93], キーワードの χ^2 値を利用した方法 [Tam88] などが提案されている．また, シソーラスを用いた日本語テキストの自動分類については [Kaw92] の手法がある．河合は, 各カテゴリーごとに偏って出現する意味属性をあらかじめ自動学習し, その結果を用いてテキストの自動分類を行っている．本章で提案する手法はシソーラスを用いる点では河合の手法と共通するが, 以下に示す相違点を持つ．

- カテゴリー相互の相対的な関係によって特徴ベクトルを作成する．
- シソーラスの分類項目別に集計した情報だけを用いて単語別の統計情報は使用しないため, 記憶容量の消費が少ない[1]．
- シソーラスが階層型である必要がないため, ネットワーク型などその他の形状であっても対応でき, 汎用性が高く, 拡張性にも優れている．この性質により, シソーラスに専門用語, 固有名詞などの分類項目を自由に追加できる．

以下では, 5.2節で本手法の内容について述べる．5.3節では本章で行った実験の内容, 及び結果を示す．5.4節では実験で正しく分類できなかったテキストについて, 5.5節では本手法を定性的に, それぞれ考察を行う．最後に 5.6節で本章のまとめを行う．

5.2　分類手法

ここでは, 本研究で提案する手法について説明する．本手法は大きく以下の三つのプロセスに分かれる．

[1] [Kaw92] では, 単語別に集計した情報も併用して用いたほうが高い精度が得られている．

1. カテゴリーの特徴ベクトルの作成
2. テキストの特徴ベクトルの作成
3. テキストと各カテゴリーとの類似度計算

図 5.1に, 本手法の大まかな処理の流れを示す.

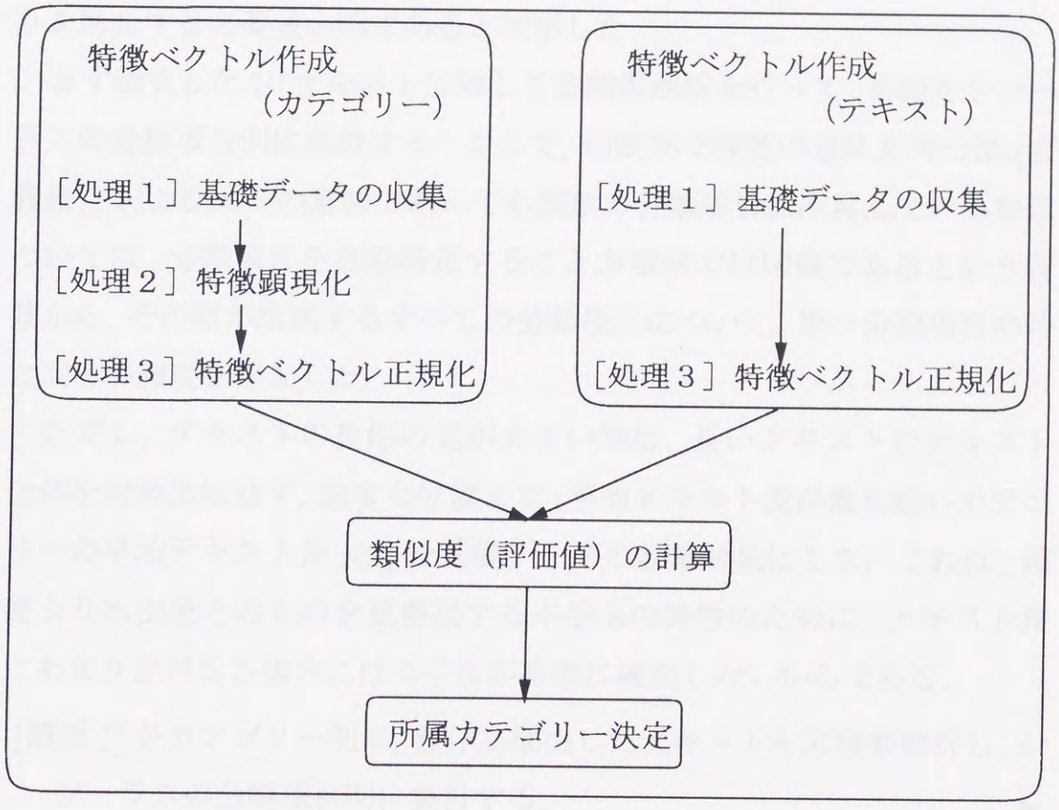


図 5.1: 本手法全体の処理の流れ

5.2.1　カテゴリーの特徴ベクトルの作成

基礎データの収集　まずカテゴリーごとにそのカテゴリーの基礎データを収集する. 基礎データの対象となるテキストは, そのカテゴリーの特

徴を最も典型的に表しているテキスト 1 記事を選ぶ方法も考えられるが, 典型的な 1 記事を選ぶことは容易ではない. そこで, 本研究ではカテゴリー対象のテキスト中から無作為に 10 記事を抽出し, それらの平均を基礎データとする方法を採用した. 一般に, 抽出する記事数が多いほどデータとしての信頼性は高くなるが, 本研究では 10 記事程度が記事内容の偏りを防止する必要最小限であると判断した.

まず選択した 10 テキストに対して形態素解析を行って, 単語をシソーラスの分類項目別に集計する. ここで, 同表記で複数の意味を持つ語 (多義語), および同一の意味であっても複数の分類項目に所属している語については, 分類項目を自動特定することが現状では困難であるという背景から, その語が所属するすべての分類項目について, 単一分類項目の時と同一の頻度を追加した.

ただし, テキストの長短の差が大きい時は, 長いテキストはテキスト全体を対象にはせず, 適度な位置まで (平均テキスト長が最も短いカテゴリーの平均テキスト長 ×2) を基礎データの収集対象にした. これは, 頻度よりも出現そのものを重要視する本手法の特性のために, テキスト長にあまり差がある場合には本手法が有効に機能しないからである.

[**処理** 1] 各カテゴリー別に, 無作為抽出したテキストを形態素解析し, シソーラスの分類項目別に集計する. ■

以下では, 予め定められたカテゴリーを $\mathcal{C}_1, \mathcal{C}_2, \cdots, \mathcal{C}_M$ (M: 分類するカテゴリー数), カテゴリー $\mathcal{C}_i$ ($0 \leq i \leq M$) に属するテキストに含まれている全単語をシソーラスによって分類, 集計した結果を $\mathcal{C}_i = (e_{i1}, e_{i2}, \cdots, e_{iN})$ と表記する (e_{ij}: 出現頻度, N: シソーラスの分類項目数).

特徴顕現化　テキスト中に出現する語の中には, 分類に有効な語とそうでない語が含まれている. テキストを有効に分類するためには, これらの語のうち, 分類に有効な語, あるいは分類項目をより目立たせる必要がある. 本処理では, 分類体系相互の関係を使用して特徴の顕現化処理を施す.

ここでは, 前処理と同様に分類項目 i の出現したカテゴリー数に着目する. 例えば, ある分類項目 i が半分以上のカテゴリーに出現した場合と, 分類項目 i がある一部のカテゴリーにしか出現しなかった場合を比較すると, 後者の場合はテキストを分類する際の非常に重要な情報となり得る. このように, 分類項目 i の出現したカテゴリー数が少なくなるほど, その分類項目は重要な役割を果たし, 唯一のカテゴリーの場合にその重要性が最大になる. ここでは, このような効果を与えるために以下の処理を行う.

[**処理** 2] 分類項目 j について, $e_{ij} > 0$ であるカテゴリー数を $m(0 \leq m \leq M)$[2]としたとき, すべての $\mathcal{C}_i(1 \leq i \leq M)$ のすべての要素 $e_{ij}(1 \leq j \leq N)$ について, 以下の計算を行う.

$$e_{ij} \leftarrow e_{ij}(M - m) \tag{5.1}$$

以上の処理をすべての分類項目 $j(1 \leq j \leq N)$ に対して行う. ■

[**処理** 2] のうち, $m = M$ の特別な場合は以下の [**処理** 2-1] に該当する.

[**処理** 2-1] 分類項目 j について, もしすべての $\mathcal{C}_i(1 \leq i \leq M)$ の要素で $e_{ij} > 0$ であるならば, すべての $\mathcal{C}_i(1 \leq i \leq M)$ の要素 e_{ij} について, 以下の処理を行う.

$$e_{ij} \leftarrow 0 \tag{5.2}$$

[2]形式的に, $m = 0$ の場合 (すべてのカテゴリーで $e_{ij} = 0$ の場合) も含めて [**処理** 2] とした.

以上の処理をすべての分類項目 $j(1 \leq j \leq N)$ に対して行う. ■

この処理は, 全カテゴリーに共通して出現が見られる分類項目に対してその値を 0 にする処理であり, カテゴリーの弁別効果を持たない分類項目の, 特徴ベクトル全体への影響を排除する効果を持っている.

全カテゴリーに出現の見られる分類項目に対して, どのカテゴリーにどれだけ出現したかという頻度情報は特徴ベクトル作成の有力な情報とはなり得ず, むしろこれらの分類項目は一般に高頻度であるために悪影響を与える可能性が強いと考えられる. このため, これらの情報は無視する方が適当と考えられ, 本研究ではこれらの語の情報は特徴ベクトルの作成には反映させなかった.

特徴ベクトルの正規化　最後に, 以上のようにして得られた特徴ベクトルの長さが同一になるように, 以下の [**処理** 3] に示す正規化を行う.

[**処理** 3] すべての $\mathcal{C}_i(1 \leq i \leq M)$ のすべての要素 $e_{ij}(1 \leq j \leq N)$ について, 以下の計算を行う.

$$e_{ij} \leftarrow \frac{e_{ij}}{l_i} \tag{5.3}$$

ただし, $l_i = \sqrt{\sum_{k=1}^{N} e_{ik}^2}$ とする. ■

5.2.2　テキストの特徴ベクトルの作成

これから分類すべきテキストについて, カテゴリーと同様に特徴ベクトルを作成する. ただし, 図 1 に示すように, カテゴリーの特徴ベクトル作成における [**処理** 2](カテゴリー間比較のための処理) は必要ないので,

[**処理** 1] と [**処理** 3] を行う[3].

5.2.3　類似度の計算

前節に示す処理で得られたカテゴリーの特徴ベクトルとテキストの特徴ベクトルを比較し, その類似度 (評価値) を見ることで, テキストが属するカテゴリーを決定する. 本研究で使用する類似度を, 以下で定義する.

[**定義** 1] カテゴリー $\mathcal{C}_i$ とテキスト $\mathcal{T}$ の特徴ベクトルがそれぞれ, $\mathcal{C}_i = (e_{i1}, e_{i2}, \cdots, e_{iN})$, $\mathcal{T} = (t_1, t_2, \cdots, t_N)$ であるとする. カテゴリー $\mathcal{C}_i$ とテキスト $\mathcal{T}$ の類似度 $eval(\mathcal{C}_i, \mathcal{T})$ を以下の式 (内積) で定義する.

$$eval(\mathcal{C}_i, \mathcal{T}) = \sum_{k=1}^{N} e_{ik} \cdot t_k \tag{5.4}$$

■

5.2.4　未知語の取扱い

本研究では, 解析対象をシソーラスに掲載されている語のみに限定した. シソーラスを使用することで, シソーラスにない未知語, すなわち新語, 特殊な専門用語, 多くの固有名詞などを取り扱うことが不可能になる.

[3]テキストの自動分類を行うだけなら [**処理** 3] も必要ないが, 異なるテキストで評価値の比較を行うために, 本研究の実験では [**処理** 3] も行った.

一般的に，専門用語，固有名詞などはしばしばテキストの特徴を反映し，そのテキストの分野を特定する上で重要度の高い要素である可能性がある．しかしこれらの語は非常に多くあるため，汎用性を持ったテキストの自動分類を行うためには，これらの語に依存しない手法が望ましいと考えた．また，固有名詞は分類の有力な手がかりにはならないという指摘 [Kaw92] もあるため，本研究ではあえてこれらの語を使用せずに分類を行うことを試みた．

5.3　評価実験

本手法の有効性を確認するために，テキストのカテゴリーへの分類を自動的に行う実験を行った．ここでは，この実験について述べる．

5.3.1　実験内容

実験に使用したテキストはいずれも日本経済新聞の 1990 年，および 1992 年に掲載されたコラムである．実験に使用したコラムとその記事数を表 5.1に示す．ただし表の「記事数」には，カテゴリーの特徴ベクトル作成に使用した 10 記事を含む．

実験では，コラム名に応じて表 5.1に示す 10 個のカテゴリーを設定した．コラムによって記事数が異なるのは，前述した年度に掲載された同タイトルのコラムをすべて使用しているためである．ただし，コラム「ニューフェース」は 1 記事のテキスト長が極端に短く記事数が多いので，1 日分をまとめて 1 記事とし，テキスト長の長い上位 200 記事を使用した．

表 5.1: 評価実験に使用したコラム

カテゴリー	コラム名	記事数
つり	「つり」	59
証券	「まちかど」	93
医学	「医フロンティア」「くすり百科」	94
政治	「92 選挙駆ける」	98
税金	「税金相談」	103
家庭	「家族はいま」	116
音楽	「音楽」	154
グルメ	「味」「味力」	195
新製品	「ニューフェース」	200
経済学	「やさしい経済学」	248

シソーラスには角川類語新辞典 [Oon81] を使用した．同辞典は日本語の語彙を 10 種類の大分類に分類し，さらに各大分類を 10 種類の中分類，各中分類を 10 種類の小分類へと階層的に分類している．本実験では，この辞典の 1000(10^3) の小分類を「分類項目」として設定した．よって，各カテゴリー，及びテキストの特徴ベクトルは 1000 次元のベクトルとなる．ただし，実際にはどのカテゴリーにも出現しない分類項目は除いているので，実際の次元は 1000 未満となる．

また形態素解析は，独自に作成した第 3章でのシステム [Yam95a] をそのまま使用した．

5.3.2　実験結果

トレーニングデータとして無作為抽出した 100 記事のコラム (各カテゴリー 10 記事 ×10 カテゴリー) について，5.2.1節に示した処理を行った結果，実際に使用した (語の出現があった) 分類項目数は 380 となった．これは各特徴ベクトルが 380 次元のベクトルであったことを意味する．

表 5.2に 5.2.1節 [**処理** 2-1] の対象となった分類項目を示す．表 5.2に列挙されている分類項目はいずれも全カテゴリーに対して出現が予想される分類項目ばかりであり，またここに挙げられていない分類項目でも，多くのカテゴリーに共通する分類項目は出現したカテゴリー数が多かった (つまり，[**処理** 2] で m が大きかった) ことから，カテゴリーの弁別効果を持たない分類項目の除去には成功していることが確認できる．

表 5.3には，[**処理** 1] の結果出現したカテゴリーが唯一だった分類項目のうち，カテゴリー別に出現頻度の高かった上位 5 項目をそれぞれ列挙する．表 5.3には，そのカテゴリーに非常に関係の深い分類項目が必ずしも列挙されているわけではなく，相対的に見てそのカテゴリーのみの出

表 5.2: 全カテゴリーに出現した分類項目名 (分類番号)

こそあど (101)	内外 (103)	数 (120)	多少 (126)
時機 (151)	先後 (156)	今昔 (158)	同一 (188)
等級 (192)	限度 (194)	大変 (195)	こんな (199)
思考 (411)	世界 (709)	番号 (823)	単位 (828)
助数詞 (829)	接辞 (834)		

現が予想される分類項目が多い．このことから，本手法はこのような複数の分類項目によってカテゴリーの特徴が捉えられていることがわかり，カテゴリー間の相対的な関係を抽出する本手法が有効に機能していることがわかる．

表 5.3: 単独のカテゴリーに出現した主な分類項目

カテゴリー	分類項目名 (分類番号)				
つり	欺瞞 (456)	狩猟 (398)	川 (036)	海 (034)	魚介 (062)
証券	札 (974)	行為 (360)	証明 (418)	地位 (682)	文書 (845)
医学	薬剤 (910)	内臓 (067)	光学器械 (993)	病気 (608)	針金 (987)
政治	推挙 (778)	党派 (715)	選択 (378)	賛否 (445)	国民 (538)
税金	取捨 (373)	従業 (364)	用地 (042)	鉱物 (088)	飾り物 (979)
家庭	家庭 (717)	親族 (528)	学校 (722)	出会い (781)	干支 (827)
音楽	楽曲 (874)	音楽 (870)	演奏 (871)	音 (096)	芸術家 (574)
グルメ	風味 (144)	野菜 (927)	料理 (923)	炊事 (355)	調味料 (925)
新製品	機械 (990)	球技用語 (899)	写真 (864)	容器 (953)	電器機具 (992)
経済学	村落 (707)	奉仕 (795)	応対 (787)	騰落 (743)	国家 (719)

次に実験結果を表 5.4に示す．ただし表中で，「正解」とは「正しいカ

表 5.4: 実験結果

カテゴリー	正解	2	3	4	5	6	7	8	9	10	r_i	p_i
新製品	190	0	0	0	0	0	0	0	0	0	100.00%	94.06%
税金	93	0	0	0	0	0	0	0	0	0	100.00%	76.23%
家庭	105	1	0	0	0	0	0	0	0	0	99.06%	87.50%
政治	87	1	0	0	0	0	0	0	0	0	98.86%	95.60%
つり	48	1	0	0	0	0	0	0	0	0	97.96%	97.96%
証券	80	2	1	0	0	0	0	0	0	0	96.39%	98.77%
音楽	138	4	0	0	1	1	0	0	0	0	95.83%	100.00%
医学	77	5	0	0	1	1	0	0	0	0	91.67%	98.72%
グルメ	169	6	4	0	2	1	2	0	1	0	91.35%	100.00%
経済学	204	25	4	3	1	1	0	0	0	0	85.71%	97.14%
1260 記事	1191	45	9	3	5	4	2	0	1	0	94.52%	94.52%

テゴリーに分類した (評価値で 1 番目になった) 記事数」, 「$n(2 \leq n \leq 10)$」は「評価値で n 番目になった記事数」を示す. また, 各カテゴリーの再現率と適合率を以下の式で定義する.

[**定義** 2] カテゴリー $\mathcal{C}_i$ の再現率 r_i, 適合率 p_i を以下の式で定義する.

$$r_i = \frac{correct_i}{original_i}, \quad p_i = \frac{correct_i}{result_i} \tag{5.5}$$

ただし, $correct_i$: カテゴリー $\mathcal{C}_i$ のテキストの内, カテゴリー $\mathcal{C}_i$ に分類された記事数, $original_i$: カテゴリー $\mathcal{C}_i$ の原記事数, $result_i$: 実験でカテゴリー $\mathcal{C}_i$ に分類された記事数とする. ■

また, 本実験では最も高い評価値となったテキストの場合のみを正解としているので, 全カテゴリーの再現率と適合率は一致する. 以下では, この全カテゴリーの再現率 (=全カテゴリーの適合率) を, 正解率と呼ぶ.

本実験では, 表 5.4に示す通り約 95% の正解率となった. 個々のカテゴリーについても, 9 カテゴリーの再現率, 8 カテゴリーの適合率が 90% 以上という結果となった. また, 正解とならなかった 69 記事のうちの 45 記事 (65%) は評価値 2 位であるので, 仮に上位 2 位までの出力を正解とすると全体で 98.1% の再現率であることを意味する. 以上の実験結果は, 他の文献とは実験環境 (対象とするテキスト, 分類するカテゴリー数など) が異なるため単純に比較することはできないが, 非常に精度が高く, 実用面でも十分耐え得るものであるといえる.

5.4　考察

ここでは, 本研究の実験で正しく分類されなかったテキストの計 69 記事について検討する. 表 5.5に, 1260 記事がどのカテゴリーに判断されたかを示す. 例えば, 表 5.5で 18(下線部分) とあるのは, カテゴリー「経済学」を「税金」と判断したものが 18 記事であったことを示す.

表 5.5: 判断した記事のカテゴリーによる分類

カテゴリー	新製品	税金	家庭	政治	つり	証券	音楽	医学	グルメ	経済学	計
新製品	190	0	0	0	0	0	0	0	0	0	190
税金	0	93	0	0	0	0	0	0	0	0	93
家庭	0	1	105	0	0	0	0	0	0	0	106
政治	0	1	0	87	0	0	0	0	0	0	88
つり	0	1	0	0	48	0	0	0	0	0	49
証券	0	3	0	0	0	80	0	0	0	0	83
音楽	1	1	4	0	0	0	138	0	0	0	144
医学	2	0	2	0	0	0	0	77	0	3	84
グルメ	7	4	1	0	1	0	0	0	169	3	185
経済学	2	18	8	4	0	1	0	1	0	204	238
計	202	122	120	91	49	81	138	78	169	210	1260

例えば, 誤りの最も多かった「経済学」のカテゴリーの個々のテキストをみると,

- 「政治」と誤った「選挙と経済（3）一橋大学助教授伊藤隆敏氏」[1990 年 5 月 9 日][4]
- 「新製品」と誤った「大ヒットの理論分析（1）一橋大学教授榊原

[4]タイトルは日本経済新聞 CD-ROM 版に付されていたタイトル, 日付は原テキストの新聞掲載日を示す. 以下同様.

清則氏」[1990 年 7 月 13 日]

- 「証券」と誤った「金融業の賃金水準（1）京都大学教授橘木俊詔氏」[1990 年 5 月 28 日]

などのように, テキストの内容そのものが, 正しいカテゴリーと誤って判断したカテゴリーの両者共に近い内容のものが多く, 誤って判断されたテキストの多くはこのことが原因と考えられる.

一方で, これとは異なる原因で誤ったテキストも存在する. 例えば「経済学」を「医学」と誤って特定したテキストは「道路混雑の政治経済学（3）創価大学教授岡野行秀氏」[1990 年 9 月 19 日] である. このテキストは道路の交通量について数学的に述べられたテキストであり, 「医学」とは関係のないテキストである. またこのテキストには (交通, 密度, 速度) などの単語が特に多く使用されていた. このテキストが「医学」と判断された原因は, テキストに多用された前述の語のうち「交通」という語は特徴ベクトル作成の際のどの学習テキストにも出現せず, 「密度」と「速度」が属する分類項目 (分類番号 122) は, 「医学」のカテゴリーの基礎データのみに出現していたためである. このため, カテゴリー「医学」の類似度が高くなったと考えられる.

次に, 本研究の実験で評価値順位の最も低かった (9 位) テキストを検証する (表 5.4下線部分). このテキスト「川波, 大阪」[1990 年 7 月 18 日] はカテゴリー「グルメ」のテキストであり, 内容はウナギの店の紹介である. このテキストが「グルメ」で高い評価値が得られなかったのは, シソーラスには漢字表記「鰻」しかない, という点と, テキスト中に「グルメ」に主に影響する語が (ウナギ以外に) 少なかった点の二つが主な理由と考えられる. またこのテキストは, 他テキストよりも全カテゴリーで低い評価値が与えられており, いわば「特徴のないテキスト」であっ

たと位置付けることができる．このため，このテキストに頻出した「～円 (特に新製品に影響，以下同様)」「～時 (税金，医学)」「～分 (新製品)」「横綱・大関・関脇・小結[5](証券)」などの語が他のカテゴリーに影響し，結果として「グルメ」のカテゴリーの評価値が相対的に低いものとなったことが原因と考えられる．

5.5　議論

文献 [Kaw92] での手法において，正しいカテゴリーに分類できない要因として，河合は以下の 3 点を指摘している．

1. 使用したシソーラスの分類項目では異なる分類項目に分散して出現したことにより，傾向が捉えられなかったため
2. 多義語に対して意味属性を絞り込んでいないため
3. テキスト中の比喩，慣用表現の使用のため

以下では，これらの要素について順に検討する．

まず最初の要素であるが，[Kaw92] の例では，(出土品，発掘，文化財) という単語に対してそれぞれ (物品，採取，財産) の分類項目となるため，これらの単語から分野 (考古学) を導き出すのは困難であるとしている．一方本手法では，他のカテゴリーで「採取」や「財産」などの分類項目が出現していなければ，これらの分類項目はカテゴリー特定の有力な判断材料となる．また，仮に出現していてもその頻度が他のカテゴリーでまれである場合も判断材料となる．このように，複数の分類項目に分散

[5]テキストでは，いずれも店のメニューとして使用されている．また，これらの語が属する分類項目は「地位 (682)」である．

して出現することは，本手法においてはカテゴリーの特徴が複数存在することに相当する．従って，仮にこのうちいくつかの分類項目で他カテゴリーでの出現が見られその分類項目が特徴でなくなったとしても，他の特徴が存在するため，全体としてカテゴリーを的確にとらえている可能性が高い．

このように，複数の分類項目に分散出現することは，むしろ本手法に対して有効に働く．一般のテキストは，必ずしも出現する分類項目に偏りがあるとは限らないので，本手法は有効であると考えられる．

次に語の多義性の問題であるが，多義語の意味特定は困難であるので，本手法においても意味の特定は行っていない．しかし，カテゴリーの特徴ベクトル作成時に例えば「スポーツ」のカテゴリーで「アンカー」という語が使用された場合，もう一つの意味である「いかり」の属する分類項目に他のカテゴリーで出現がなければ，カテゴリー特定の妨げにはならない．このことから，多義性の問題は本手法においても依然として存在するが，悪影響を及ぼす可能性は比較的低い．

また第三の要因である慣用表現については，その表現が一般的に多用されるものであればそれがどのカテゴリーの特徴ともならないため問題なく，まれに出現するものであればその語が全体に及ぼす影響は低いと考えられるためこれも問題ない．また，ある表現がある特定のカテゴリーのみで頻出した場合は，その表現がどのような表現であってもカテゴリー特定の判断材料となり得る．一方で，テキスト中で多種多様な慣用表現を使用している場合には本手法が有効に機能しない可能性があるが，そのようなテキストは特殊なものであると考えられる．以上より，慣用表現についても悪影響を及ぼす可能性は比較的低い．

5.6　まとめ

本章では, 分類体系相互の関係を利用した日本語テキストの自動分類手法を提案した. この手法を用いて合計 1260 記事の新聞コラムを対象にした 10 カテゴリーへの分類を行った結果, 全体の平均で約 95% が正しく分類できた. なお, 実験は日本語のテキストを対象にして行ったが, 本手法は対象とする言語のシソーラスさえあれば任意の言語に適用可能であり, 本手法の一般性は高い.

あるテキストを特徴づけることは他テキストとの比較によってはじめて可能となる. 本手法はこの「相対性」の精神に基づいた手法であると位置付けできる. キーワードの自動抽出や文章の抄録・要約の作業も, 文章分類と同様にこの相対性の要素を持っていることから, 今後は本手法のこれらへの適応が課題である.

第 6 章

結　論

6.1　まとめ

本論文では今後ますます重要になると考えられる自然言語処理の一分野である談話処理, その中でも, 従来ほとんど研究が行われていないが機械可読文書の有効利用のための基礎となる文章の段落分け, 実在の文章を対象にした要約, 類義語を考慮した文章の自動分類などの処理を取り上げ, 検討を行った. これらの問題は互いに独立した問題ではあるが, それぞれの処理を行うための談話解析は多くの部分が共通している. また, 対象言語としては日本語を取り扱っている.

論文の第 2章では, 日本語文章中の結束構造の解明を目的として日本語文章の段落分けについて論じた. ここでは, 日本語の結束構造に影響を与える要素として, 「手がかり語」と呼ぶ接続的語句と単語間の類縁性, すなわちシソーラスで把握することのできる単語間の関係の二つを取り上げ, 計算機による段落分け実験を行うことによって, 個別の要素を使用した場合と両者を併用した場合の比較を行った. その結果, 両者共に文章中の結束性保持に重要な役割を果していることが確認された. また, 出力結果の自然さを評価するために, アンケート調査を実施し, 作成した

手法の有効性を支持する結果を得た.

第3章では日本語の論説文章を対象とした文章の要約作成について考察し, 一手法を提案した. ここでは, 従来多く行われている抄録, すなわち文章からのある評価基準に基づく文の抽出にとどまらず, 抽出した文の修飾語の短縮も試みた. また, 要約された出力結果にある程度の読み易さを備えることを目指し, 原文から最低限度の結束性を保持したまま文を抽出する手法を提案した. また, ここでも被験者による評価を行い, 本提案手法に基づく実験システムで生成した文章の品質を確認した.

一方第4章では, 複数の文章に対する要約について述べた. 最近の計算機とネットワークの進歩により, 情報検索がより身近なものになり, 複数の文章となる検索結果の概要の把握が今後, より重要な技術となることが予想される. ところが, 従来は単独の文章を対象にした抄録または要約に関する研究がほとんどであり, その中での本研究の存在は貴重である. 第4章では特に, 複数文章中に共通して出現する記述部分を削除することによる複数文章の要約を試みた. ここでは削除する3要素を指摘し, これらを削除することによる要約手法を提案した.

また, 第5章では情報検索の一分野である文章の自動分類について論じた. ここでは, シソーラスの分類項目を語群に設定し, テキストの統計情報を用いた手法で大量の新聞コラムを対象にした10カテゴリーへの分類実験を行った結果, 再現率, 適合率ともに高い値を得ることができた.

現代の社会においては, 計算機の性能が急速に向上, また価格が低下してきたことによって, 過去に印刷媒体で記録されていた情報が機械可読な形で保存されることが多くなってきた. また同時に, 情報ネットワークが急速に整備されてきたことに伴い, それら膨大な情報が各所に分散

して蓄積されるようになってきた．このような状況下で，膨大な情報の中から必要な情報だけを抽出，そして要約など必要な形に加工する技術の確立が急務な課題となりつつある．

画像情報，テキスト情報，音声情報に大別することのできる情報の中で，最も中心的な情報はテキスト情報である．そのテキスト情報を取り扱うためには自然言語処理が，また高度な処理を行うためには特に談話処理の技術が非常に重要である．このように，社会の要請に応えるためには談話処理の進展が不可欠であるが，談話処理またはこれに関連した研究で，特に本研究で行ったような，例えば実際の新聞記事などを対象にするなどといった実働を視野に入れた研究は従来多く行われてこなかった．

これまで述べてきたように，本論文では今後の談話処理研究を進めるにあたってのいくつかの示唆や，実用的なシステムを構築する際に不可欠なヒューリスティックスの提案を行ったが，本研究で得られた知見は談話処理を組み込んだ高度の情報処理の実現に必要ないくつかの技術の一端を示したにすぎない．現代の，さらには近未来の高度情報化社会の実現に不可欠な「必要な時に必要な情報を必要な形で利用できる」技術の確立のために，今後談話処理研究のより一層の進展が望まれる．

6.2　今後の課題

本研究ではそれぞれの課題に対して一応の成果を得ることができたが, その一方, 今後の課題として解決しなければならない問題も多い. 以下では, そうした課題のいくつかを指摘する.

段落分けや要約処理は, 文章生成処理の一部とも考えられる. 文章生成は what-to-say と how-to-say の二つの処理に分解することができると考えられており [Tok91], 本論文では要約作成が what-to-say 部の, 段落分けは how-to-say 部の, それぞれ部分的な処理と考えることができる. このうち, how-to-say に関しては本論文では文章内の構成を変更するまで到っておらず, これからの課題として残された部分は多い. what-to-say に関しては, 重要性の指標が言語の表層的特徴に比較的表出しやすく, これを利用すれば本論文で提案した手法のような第一次の近似処理は可能であるが, より詳細な解析手法による出力文章のさらなる品質向上が望まれる.

一方, 文章の自動分類, 情報検索に関しては, 本研究での手法を含めて現在提案されている多くの手法が, 対象文章中の単語出現頻度などの統計情報を利用している. 統計的情報は分類のための最も重要な情報で, これを利用した処理は汎用性が高く, 処理を工夫すればある程度高い分類精度が得られるが, 例えば「日本の対米自動車輸出」と「米国車の輸入」(=「米国の対日自動車輸出」) という二つの主題を区別する場合などはほとんどの統計的手法が有効に作用しないと予想される. このような場合は談話処理の分野で使用されている技術を利用, 応用すれば対応できると考えられる.

最後に, 研究環境面での課題について触れて, 本論文の結語とする.

一般に, 自然言語処理研究, 特に実働を視野に入れた研究を行うためには計算機を必要とする. 幸い, 最近の性能向上と低価格化によって多くの研究者が比較的高性能な計算機を利用することが可能になった. ハードウェア面での課題が少なくなった一方で, 研究者, 特に談話処理研究者は研究の内容によって以下に示すいくつかの「言語資源」を用意しなければならない.

- 処理ツール：形態素解析, 構文解析などを行うツール
- 言語データ：テキストデータベース, コーパス
- 知識データ：国語辞書, シソーラスなどの各種辞書

実際には, 研究利用のためにこれらの言語資源を入手, 使用することは今まで容易ではなく, 何らかの制限を受けることが多い. そもそも, 高度の意味情報を含んだ「知識ベース」と呼ばれるもののように, 現在依然として研究段階で入手可能なものが存在しない言語資源もあり, 研究者の希望通りに全てが準備できることは一般に稀である. また, 前述の言語資源の使用, 公表に関して, 著作権をはじめとする権利の問題も重要である.

幸い本研究の遂行に際しては, 処理ツールとしては京都大学／奈良先端科学技術大学院大学のグループによる JUMAN[Mat93] が, 言語データとしては日本経済新聞 CD–ROM が, 知識データとして角川類語新辞典 [Oon81] を利用することができた. このように, 最近になってこれら共通の言語資源が徐々に整備されつつあることは非常に歓迎すべき傾向であるが, まだまだ十分な段階にあるとはいえない. これらを整備することは一般に長時間を要するため地道ではあるがまた必要不可欠でもある. これからも継続的に自然言語処理の研究者全体で努力を続けなければならない.

謝　辞

本研究で, シソーラスに使用した「角川類語新辞典」[Oon81] を機械可読辞書の形で提供いただき, その使用許可をいただいた (株) 角川書店に深謝する. また, 日本経済新聞の一部記事について, 機械可読テキストとしての使用許可、及び本論文への引用許可をいただいた (株) 日本経済新聞社にも深謝する.

本研究の開始当初から, 懇切丁寧に指導, 激励していただいた, 豊橋技術科学大学知識情報工学系助教授で筆者の指導教官である増山繁先生, 及び本研究全般に, また詳細にわたり熱心に助言をいただいた NTT 基礎研究所, 現在 NTT ソフトウェア研究所の内藤昭三氏に対して心から深謝の意を表する. 両先生の援助なくしては本論文は有り得なかった. また, 本論文作成に際し数々の有益なコメントをいただいた豊橋技術科学大学の磯田定宏教授, 斉藤制海教授, 中川聖一教授, 河合和久助教授の各先生に対しても, 深く感謝する.

本研究の討論に参加していただき, また快適な計算機環境に整備するなど, 執筆作業を陰で支えてくれた増山研究室の皆さんにも, お礼を申し上げたい. また, 豊橋技術科学大学の渋澤博幸氏 (現在同大学助手) と同大学増山研究室の中山慎一氏は同期生の友人であり, 現在に至るまで様々な面での助言と協力をいただいた. 両氏に感謝の意を表する.

最後に, 絶えまない協力と励ましを贈ってくれた母に, 心から感謝の言葉を贈りたい.

参考文献

[Fuk91] 福本淳一, 安原宏：日本語文章の構造化解析, 研究会資料 NL85-11, 情報処理学会 (1991).

[Gut94] GUTHRIE, L. and WALKER, E.: Document Classification by Machine: Theory and Practice, *Proceedings of COLING 94*, pp. 1059–1063 (1994).

[Has88] HASIDA, K., ISIZAKI, S., and ISAHARA, H.: An Approach to Abstract Generation, *Bulletin of the Electrotechnical Laboratory*, Vol. 52, No. 4, pp. 1–14 (1988).

[Hir90] 平井昌夫：何でもわかる文章の百科事典, 三省堂 (1990).

[Ike83] 池上嘉彦：テクストとテクストの構造, 談話の研究と教育 I, pp. 7–42, 国立国語研究所 (1983).

[Iti78] 市川孝：国語教育のための文章論概説, 教育出版 (1978).

[Joh95] 情報処理学会（編）：新版 情報処理ハンドブック, オーム社 (1995).

[Kar94] KARLGREN, J. and CUTTING, D.: Recognizing Text Genres with Simple Metrics Using Discriminant Analysis, *Proceedings of COLING 94*, pp. 1071–1075 (1994).

[Kaw92] 河合敦夫：意味属性の学習結果にもとづく文書自動分類方式,

情報処理学会論文誌, Vol. 33, No. 9, pp. 1114–1122 (1992).

[Kes93] 芥子育雄, 乾隆夫, 石鞍謙一郎：大規模文書データベースからの連想検索, 技術研究報告 AI92-99, 電子情報通信学会 (1993).

[Kom87] 小松英二, 加藤安彦, 安原宏, 椎野努：要約支援システム COGITO - 文章の構造解析 -, 研究会資料 NL64-11, 情報処理学会 (1987).

[Luh58] Luhn, H. P.: The Automatic Creation of Literature Abstracts, *IBM J.Res. and Dev.*, Vol. 2, pp. 159–168 (1958).

[Mas89] 間瀬久雄, 大西昇, 杉江昇：説明文の抄録作成について, 技術研究報告 NLC89-40, 電子情報通信学会 (1989).

[Mat93] 松本裕治, 黒橋禎夫, 宇津呂武仁, 妙木裕, 長尾真：日本語形態素解析システム JUMAN version 2.0, 京都大学工学部長尾研究室／奈良先端科学技術大学院大学松本研究室 (1993).

[Mor91] Morris, J. and Hirst, G.: Lexical Cohesion Computed by Thesaural Relations as an Indicator of the Structure of Text, *Computational Linguistics*, Vol. 17, No. 1 (1991).

[Mot95] 望月源, 本田岳夫, 奥村学：複数の知識の組合せを用いたテキストセグメンテーション, 研究会資料 NL109-7, 情報処理学会 (1995).

[Mur90] 邑本俊亮, 阿部純一：物語文章の要約化処理について, 研究会資料 NL78-10, 情報処理学会 (1990).

[Nag86] 永野賢：文章論総説, 朝倉書店 (1986).

[Nak91] 中澤俊哉, 重永実：エピソードネットワークを用いた物語のあらすじ生成, 情報処理学会論文誌, Vol. 32, No. 10, pp.

1215–1224 (1991).

[Nom94] 野本忠司：日本語テキストの統計的構造化について，技術研究報告 NLC93-63, 電子情報通信学会 (1994).

[Nom96] 野村浩郷, 井佐原均, 徳永健伸, 中村貞吾：情報ハイウェイ時代のテキスト情報への知的アクセス, 情報処理, Vol. 37, No. 1, pp. 1–9 (1996).

[Oon81] 大野晋, 浜西正人：角川類語新辞典, 角川書店 (1981).

[Sak89] 佐久間まゆみ（編）：文章構造と要約文の諸相，日本語研究叢書, No. 4, くろしお出版 (1989).

[Sal88] Salton, G. and Buckley, C.: Term-Weighting Approaches in Automatic Text Retrieval, *Information Processing & Management*, Vol. 24, No. 5, pp. 513–523 (1988).

[Sal94] Salton, G., Allan, J., Buckley, C., and Singhal, A.: Automatic Analysis, Theme Generation, and Summarization of Machine-Readable Texts, *Science*, Vol. 264, pp. 1421–1426 (1994).

[Spa72] Sparck Jones, K.: A Statistical Interpretation of Term Specificity and its Application in Retrieval, *Journal of Documentation*, Vol. 28, No. 1, pp. 11–21 (1972).

[Sum95] 住田一男, 知野哲朗, 小野顕司, 三池誠司：文書構造解析に基づく自動抄録生成と検索提示機能としての評価，電子情報通信学会論文誌, Vol. J78-D-II, No. 3, pp. 511–519 (1995).

[Suz88] 鈴木章, 小橋史彦, 深谷健一：ビジネス通知文書の自動分類，技術研究報告 OS87-42, 電子情報通信学会 (1988).

[Tak89] 竹内啓（編）：統計学辞典, 東洋経済新報社 (1989).

[Tam88] 田村淳, 渡辺道枝, 原良憲, 笠原裕：統計的手法による文書自動分類, 全国大会論文集 36-6U5, 情報処理学会 (1988).

[Tam92] 田村俊哉, 田村直良：文章の表現形式に基づいた要約文章の生成について, 研究会資料 NL92-1, 情報処理学会 (1992).

[Ter81] 寺村秀夫：日本語の文法 (下), 日本語教育指導参考書, No. 5, 国立国語研究所 (1981).

[Ter91] 寺村秀夫：日本語のシンタクスと意味 III, くろしお出版 (1991).

[Tok91] 徳永健伸, 乾健太郎： 1980 年代の自然言語生成 － 2 －, 人工知能学会誌, Vol. 6, No. 4, pp. 510–519 (1991).

[Tok94] Tokunaga, T. and Iwayama, M.: Text Categorization based on Weighted Inverse Document Frequency, 研究会資料 NL100-5, 情報処理学会 (1994).

[Tsu94] 津田宏治, 仙田修司, 美濃導彦, 池田克夫：共起関係の固有ベクトルを用いる単語クラスタリング法, 研究会資料 NL103-6, 情報処理学会 (1994).

[Wat94] 渡辺靖彦, 竹内雅人, 村田真樹, 長尾眞： χ^2 法を用いた重要漢字の自動抽出と文献の自動分類, 技術研究報告 NLC94-25, 電子情報通信学会 (1994).

[Yam83] 山崎誠：文章の話題の展開を計る尺度, 計量国語学, Vol. 13, No. 8, pp. 346–360 (1983).

[Yam91] 山本和英, 増山繁, 内藤昭三：手がかり語を用いた日本語文章の段落分けに関する実証的考察, 研究会資料 NL84-9, 情報処理学会 (1991).

[Yam94] 山本和英, 増山繁, 内藤昭三：段落分けを用いた日本語文章における結束構造の検討, 情報処理学会論文誌, Vol. 35, No. 10, pp. 2029–2037 (1994).

[Yam95a] 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).

[Yam95b] Yamamoto, K., Masuyama, S., and Naito, S.: An Empirical Study on Summarizing Multiple Texts of Japanese Newspaper Articles, *Proc. of Third Natural Language Processing Pacific-Rim Symposium*, pp. 461–466 (1995).

[Yam95c] Yamamoto, K., Masuyama, S., and Naito, S.: Automatic Text Classification Method with Simple Class-Weighting Approach, *Proc. of Third Natural Language Processing Pacific-Rim Symposium*, pp. 498–503 (1995).

[Yua93] 湯浅夏樹, 上田徹, 外川文雄：大量の文書データから自動抽出した名詞間共起関係による文書の自動分類, 研究会資料 NL98-11, 情報処理学会 (1993).

付録 A:*GREEN* による要約結果の例とその原文

ここでは, 第 3章で被験者にアンケート調査を行った際の *GREEN* による要約記事, およびその原文を示す. 要約目標は 25% とした. なお, 本研究で計算機実験, およびアンケート調査に使用した日本経済新聞の社説について, 本論文への引用許可をいただいた (株) 日本経済新聞社に深謝する.

文章 A ([8/Nov/1990])

原文 (タイトル「再検討が必要な政令恩赦」)

政府は天皇陛下の即位の礼に伴い、恩赦（政令恩赦と特別恩赦）を実施する方針だ。対象範囲を定めて一律に救済する政令恩赦は復権だけにとどめるようだが、これによって大量の選挙違反者の公民権が回復される。これでは選挙違反者救済が目的ととられてもやむを得ない。国民の不信を招くような恩赦は極力避けるのがのぞましく、再審や個別恩赦などの法制が整備されている現在、政令恩赦それ自体、必要なのかどうかも再検討する時期に来ている。

天皇のご逝去や新天皇の即位など国家的な慶弔事があった時に恩赦を実施するというのが戦前からの慣行で、戦後も法制化されている。罪を犯した人もともに悲しみ、

喜ぶという趣旨である。戦後だけでも、終戦時に始まって、国連加盟、沖縄返還などに伴い九回の恩赦が行われた。

こうした際には政令恩赦とともに、政令恩赦からもれた人を申請に基づいて救う特別恩赦を実施する。救済の対象はおおむね狭まっており、戦前のように殺人犯まで釈放してしまうようなことはなくなった。その代わりに復権という形で救済措置がとられるようになり、その結果として選挙違反者の公民権回復が目立つようになった。

昨年二月には昭和天皇のご逝去に伴う恩赦を実施、やはり選挙違反者の権利回復を認めた。それからわずか一年半余り後の今回、また大量の選挙違反者を救済するという。これでは来春の統一地方選挙をにらみ、昨年の参議院選挙、今年の衆議院総選挙後に公民権を停止された数千人を救うのが目的、とみられてもやむを得ない。道路交通法違反者の免許取得などの権利を回復する件数が数百万に上るようだが、だからといって選挙違反者を含めていいという説明にはならない。こうしたことが繰り返されるようでは、政治への信頼は損なわれるばかりだ。

司法権に基づく決定と執行を、行政権が事後に変更する恩赦は本来、法の画一性から生じる妥当性を欠く判決を矯正したり、誤判からの救済などを目的としている。戦前はともかく、現在は再審制度など法制が整備され、仮釈放制度や中央更生保護審査会が決める常時恩赦も用意されている。裁判結果を変更する機能は十分に働いているのである。

政令恩赦は再審や常時恩赦を知らない人をも救える側面はあるが、それにしても不定期に行われるために、運良く救われる人とそうでない人が出るという不公平を生じる。政府部内には「ずっと続いてきたものを、今やめることも難しい」との意見もある。しかし、国民に影響を与えることを惰性で続けるのはおかしい。広く意見を聞きながら、恩赦を改めて考え直すべき時期に来ている。

要約結果 (要約率： 30.6 %)

政府は天皇陛下の即位の礼に伴い、恩赦（政令恩赦と特別恩赦）を実施する方針だ。

救済の対象はおおむね狭まっており、戦前のように殺人犯まで釈放してしまうようなことはなくなった。その代わりに復権という形で救済措置が取られるようになり、その結果として選挙違反者の公民権回復が目立つようになった。

道路交通法違反者の免許取得などの権利を回復する件数が数百万に上るようだが、

だからといって選挙違反者を含めていいという説明にはならない。こうしたことが繰り返されるようでは、政治への信頼は損なわれるばかりだ。

政府部内には「ずっと続いてきたものを、今やめることも難しい」との意見もある。しかし、国民に影響を与えることを惰性で続けるのはおかしい。広く意見を聞きながら、恩赦を改めて考え直すべき時期に来ている。

文章 B ([13/Nov/1990])

原文 (タイトル「環境保全は目標値が必要」)

地球温暖化の防止を目指してジュネーブで開かれていた第二回世界気候会議が終わった。百二十カ国の政府代表が閣僚宣言を採択したが、焦点の二酸化炭素など温室効果ガスの排出量を規制する具体的な目標値を設定することはできなかった。今回の会議は、来年二月から始まる温暖化防止のための条約作りの基礎になるだけに、目標が不明確のままに終わったのは残念である。

閣僚宣言では、「これ以上遅らせることなく気候変動への世界的対策を決定し、実施すべきである」「科学的な不確実さがあるからといって、対策の実施を延期すべきでない」「気候に悪影響を与えない水準で温室効果ガスの濃度を安定化する」など、原則的な考え方は盛り込まれている。今回の会議で、これらの原則が政府レベルで合意できたことは、大きな意味がある。

しかし、具体論では「すべての先進国が排出抑制に効果のある目標や計画を設定するよう促す」として各国の自主的な判断を待つ形になっている。わが国をはじめ、ＥＣ諸国、オーストラリアなどは、すでに目標や計画を打ち出している。しかし、二酸化炭素の排出量が特に多い米国とソ連は、必ずしも積極的ではない。今回の会議でも、具体的な目標値を設定できなかったのは米国の抵抗が強かったためである。

環境問題では、具体的な目標がきわめて大切である。環境保全の努力は、経済のメカニズムにゆだねて進むものではないし、人間の善意に期待するだけではもっと進むはずもない。目標値が明確であってこそ、それぞれの努力に対する評価もできるし、

企業などに対する圧力にもなる。目標値がなければ、ただの精神論に終わってしまい、地球を温暖化から守ることはできない。

米ソの緊張緩和が進み、世界に対する米ソ両国の影響力が減少したといわれる。しかし、湾岸危機で示されたように、世界の安全に対する両国の役割はまだまだ大きい。地球環境問題も、まさに人類共通の安全の問題である。両国共に厳しい国内事情があるとはいえ、地球環境問題でもその指導的役割を自覚して欲しいものである。

わが国が先に決めた「国民一人当たりの二酸化炭素排出量を二〇〇〇年に一九九〇年の水準で安定化する、日本の総排出量を二〇〇〇年以降一九九〇年の水準で安定化するよう努力する」という地球温暖化防止行動計画は、会議の席上でも高く評価され、日本の指導力にも期待が寄せられている。日本は、この行動計画に向けて最大限の努力を払うと共に、条約交渉に当たって米ソ両国に対し目標を設定すべく強く働きかける必要がある。

要約結果 (要約率： 27.9 %)

地球温暖化の防止を目指してジュネーブで開かれていた第二回世界気候会議が終わった。

米ソ両国の影響力が減少したといわれる。両国共に厳しい国内事情があるとはいえ、地球環境問題でもその指導的役割を自覚して欲しいものである。

わが国が先に決めた「国民一人あたりの二酸化炭素排出量を二〇〇〇年に一九九〇年の水準で安定化する、日本の総排出量を二〇〇〇年以降一九九〇年の水準で安定化するよう努力する」という地球温暖化防止行動計画は、会議の席上でも高く評価され、日本の指導力にも期待が寄せられている。日本は、この行動計画に向けて最大限の努力を払うと共に、条約交渉に当たって米ソ両国に対し目標を設定すべく強く働きかける必要がある。

文章 C ([17/Nov/1990])

原文 (タイトル「全欧安保会議に期待する」)

東西冷戦後の欧州新秩序を討議する全欧安保協力会議（ＣＳＣＥ）の首脳会議が十九日、パリで開幕する。前回七五年八月のヘルシンキ会議から十五年、ソ連・東欧の改革、ドイツ統一、欧州共同体（ＥＣ）の市場統合など、欧州は内部から歴史的ともいえる変革を遂げた。ＣＳＣＥは来るべき統合欧州の「屋根」としての役割を担い、世界各地域の安全保障体制のあり方にも示唆を与える。

欧州三十二カ国に米国とカナダが加わる三日間の会議では、まず北大西洋条約機構（ＮＡＴＯ）、ワルシャワ条件機構双方の全加盟国二十二カ国による欧州通常戦力（ＣＦＥ）条約調印、両軍事同盟間の不可侵宣言などがあり、ＣＳＣＥの常設機構化や新秩序の方向を盛り込んだ「パリ宣言」が採択される。

前回の首脳会議は七〇年代の緊張緩和（デタント）を象徴する出来事だった。採択された最終合意文書「ヘルシンキ宣言」には安全保障、経済協力、人権の各項目が並び、欧州社会の将来を律する精神的支柱とも見えたが、第二次大戦後の姿、つまり政治的に東西に分断された現状をありのまま認め、共存していこうとする現実主義が根底にあった。いわば、次善の策だった。

その点、ヘルシンキとパリには本質的な違いがある。昨年の革命で東欧の社会主義政権が倒れ、民主主義体制に変わった。ソ連もゴルバチョフ政権のペレストロイカ（改革）により、政治改革を経て市場経済への移行を進めている。東側が西側の価値観を受け入れ、協調しながら欧州統合を目指そうとしているのがいまの局面と見ていいだろう。

軍事衝突の可能性が減り、安全保障の主役は軍事力から経済力へと入れ替わりつつある。ＥＣの果たす役割は一段と大きくなるが、その中軸である統一ドイツとソ連の接近で、欧州の重心は東の方に移動するだろう。ドイツのコール首相は独ソ善隣友好協力条約調印後の記者会見で、「ＥＣの場でソ連の利益を代弁する」と明言した。一方、フランスもソ連と友好協力条約を結び、欧州統合の主導権を握ろうとしている。

欧州にとって当面の最大の課題は、どうすればソ連・東欧の市場経済化に成功するか、にある。経済改革には十年単位の時間が必要だ。しかし、人々は気が短い。政治が不安定になり、経済にはね返る――こんな悪循環に陥る危険もある。逆に、うまく

いけば資源やエネルギーを含め自給自足も可能な一体化した欧州経済圏が誕生する。

そして、長期的には米国と欧州との関係がどう変わるか。欧州の統合が進めば、米国のプレゼンスは確実に小さくなっていく。米側の対欧州戦略にも注目したい。

要約結果 (要約率： 27.0 %)

全欧安保協力会議（ＣＳＣＥ）の首脳会議が十九日、パリで開幕する。ＣＳＣＥは来るべき統合欧州の「屋根」としての役割を担い、世界各地域の安全保障体制のあり方にも示唆を与える。

東側が西側の価値観を受け入れ、協調しながら欧州統合を目指そうとしているのが今の局面と見ていいだろう。

ＥＣの果たす役割は一段と大きくなるが、その中軸である統一ドイツとソ連の接近で、欧州の重心は東の方に移動するだろう。

欧州にとって当面の最大の課題は、どうすればソ連・東欧の市場経済化に成功するか、にある。経済改革には十年単位の時間が必要だ。欧州の統合が進めば、米国のプレゼンスは確実に小さくなっていく。米側の対欧州戦略にも注目したい。

文章 D ([23/Nov/1990])

原文 (「突然終わった『サッチャーの時代』」)

サッチャー英首相は二十二日辞任を表明、一九七九年五月以来の長期政権に終止符を打った。先の保守党党首選挙で当選を決められなかったためで、サッチャー首相の退陣は内外に大きな波紋を投げかけよう。

サッチャー首相の長期政権が残した足跡は大きい。国内では、サッチャー革命と呼ばれる経済改革を強力に推進、対外的にはフォークランド紛争の解決、ゴルバチョフ氏との対ソ対話の積み上げ、対米協調路線など、国際政治に多くの功績を残した。

特に、国内のサッチャー革命は、それまで「英国病」と言われてきた英国経済の病根に鋭いメスを入れ、英経済を自由主義経済の伝統に戻す上で大きな貢献をなした。米国ではレーガン革命と言われた自由市場尊重の経済政策も、元をただせば、サッチャー首相の自由主義経済への回帰がお手本だった。

具体的には、労働党政権下で国有化された主要産業の多くを民営化して民間活力の回復を図り、労働組合に対しても強い態度で臨み、国民の各階層に自助努力と勤労の必要性を説くものだった。

この結果、英経済は立ち直り、「リトル・ブリテン」ながらもポンドは安定し経済基盤の健全化に成功した、と言っていいだろう。ときに大恐慌以来といわれる失業者を出しながらも、国民がサッチャー首相を支持したのも、その政策の正しさにあったといえる。

サッチャー首相の対外政策は、米国との「特別な関係」を誇示しながら欧州大陸と一歩置くことで外交の主導権を維持しようとする、伝統的な政策だった。レーガン前大統領との親密な関係とこれを背景にした欧州外交の展開には見るべきものがあった。

しかし、長期政権に人心が倦（う）むと同時に、サッチャー首相の強引な政策展開に批判が強まっていたことも事実である。サッチャー首相は、閣内の統一に厳しく、批判閣僚を容赦なく解任した。だが、政権誕生以来首相の股肱（ここう）の臣だったハウ副首相の辞任は、さすがに首相にこたえたようである。ハウ氏は辞任のあいさつで「過去への感傷にとらわれてはならない」と批判した。

国内政策では、新地方税（人頭税）が金持ち優遇として悪評を買い、英経済もインフレなど悪化の兆しを強めている。さらに保守党内部での強い批判のタネは、サッチャー首相の欧州共同体（ＥＣ）政策である。欧州通貨統合を目指すＥＣの基本路線に対してサッチャー首相は、国家主権の尊重を理由に慎重論を展開していた。

サッチャー退陣は、英国の国内政策ばかりでなく、対外政策に大きな影響を及ぼそう。特に、サッチャー首相はブッシュ米大統領の強力な支持者であっただけに、湾岸危機への西側の対応で変化が生じかねない。また、ゴルバチョフ大統領は良き理解者を失うことになり、ソ連の欧州政策にも微妙な影響が出よう。英国の対ソ政策、対ＥＣ政策など、グローバルな影響の分析に日本も鋭意注意を払わねばなるまい。

要約結果 (要約率： 25.9 %)

サッチャー英首相は二十二日辞意を表明、一九七九年五月以来の長期政権に終止符を打った。

欧州外交の展開には見るべきものがあった。サッチャー首相は、閣内の統一に厳しく、批判閣僚を容赦なく解任した。だが、政権誕生以来首相の股肱の臣だったハウ副首相の辞任は、さすがに首相にこたえたようである。ハウ氏は辞任のあいさつで「過去への感傷にとらわれてはならない」と批判した。

サッチャー退陣は、英国の国内政策ばかりでなく、対外政策に大きな影響を及ぼそう。特に、サッチャー首相はブッシュ米大統領の強力な支持者であっただけに、湾岸危機への西側の対応で変化が生じかねない。また、ゴルバチョフ大統領は良き理解者を失うことになり、ソ連の欧州政策にも微妙な影響が出よう。

文章 E ([25/Nov/1990])

原文 (タイトル「国連の武力行使決議案と日本の選択」)

国連安全保障理事会はイラクのクウェート撤退を求めた国連諸決議の実効をあげるため、対イラク武力行使を認める新たな決議案の協議に入る。国際紛争の解決を国連の活動に期待するわが国の国連主義外交にとっても、重大な局面である。

湾岸危機をめぐる国際政治には、先週複雑かつ重要な変化があった。全欧安保会議舞台裏での米国とソ連およびフランスとの協議、ブッシュ米大統領の湾岸諸国を中心とするアラブ首脳陣との会談、そして対イラク強硬策を主張し続けてきたサッチャー英国首相の辞意表明など、いずれも湾岸紛争の行方に影響を与える動きである。

これら一連の外交活動を通じて、いくつかの点が明らかになった。第一は、米国およびサウジアラビア、エジプト、シリアなど対イラク強硬派が武力行使の必要性について合意、国連安保理の新決議成立を急ぐことを確認したこと。第二はソ連、フランス、中国の安保理常任理事国が米英の武力行使決議提案に対し、紛争解決のため軍事

力オプションの必要性を認めつつも、なお平和的解決を追求する立場に立つという中立的姿勢をとっていることである。

武力行使が必要という考えに傾いている米国、サウジアラビアなどの政府当局者からみれば、この国連決議を通すことは、坂道で重い荷車を引くに等しい困難が伴うことだろうが、国際紛争の解決に国連の機能を使うためには、そうした忍耐は決して避けて通れないものである。

米国や英国政府内部には、国連憲章で集団的自衛権が認められており、安保理決議がなくともクウェート、サウジなど紛争関係国の要請と合意にもとづく対イラク武力行使はできるという見解がある。しかし、現実には国連での合意形成を図る方針をとろうとしている。湾岸紛争をめぐる国際政治や軍事情勢がそうした選択を必要としているのであろうが、国連外交を重視するわが国の立場からもこれは望ましい方向だ。

国連の重要メンバーであることを自認してはいるが、いま常任理事国でも理事国でもないわが国として、どうこの国連外交の重大局面に対処するのか。湾岸危機の実態認識と解決策について、どう発言するのかが問われている。ここでの傍観は決して「沈黙は金」を意味しない。

軍事力行使による解決策に反対するのならば、代替案を明確に表明しなければならないし、そのための外交努力を一段と強化する必要がある。

平和的解決には時間と大きな経済的、政治的コストがかかる。わが国が平和的解決論に立つためには、そのことをはっきりさせることが肝要だ。たとえば現実問題として、イラクに撤退決意を促すためにも、米軍など多国籍軍の増強が必要とみられるが、戦わず存在し続けるだけでも意義のある「国際警察軍」の維持費をだれが、どう負担すべきかについて明確にしなければならない。

そうした対応を欠けば、湾岸危機は欧米のニッポンただ乗り論の火に油を注ぐ結果になりかねない。

要約結果 (要約率： 25.8 %)

国連安全保障理事会は国連諸決議の実効をあげるため、対イラク武力行使を認める新たな決議案の協議に入る。

米国や英国政府内部には、紛争関係国の要請と合意に基づく対イラク武力行使はできるという見解がある。

国連の重要メンバーであることを自認してはいるが、わが国として、どうこの国連外交の重大局面に対処するのか。

軍事力行使による解決策に反対するのならば、代替案を明確に表明しなければならないし、そのための外交努力を一段と強化する必要がある。

例えば多国籍軍の増強が必要とみられるが、「国際警察軍」の維持費をだれが、どう負担すべきかについて明確にしなければならない。そうした対応を欠けば、湾岸危機は欧米のニッポンただ乗り論の火に油を注ぐ結果になりかねない。

付録 B: 要約率を変化させた場合の *GREEN* の要約結果の例

ここでは, 要約システム *GREEN* において, 同一の文章に対して目標要約率を変化させた場合の要約結果を示す. 要約対象文章は付録 A の文章 E(タイトル「国連の武力行使決議案と日本の選択」 [25/Nov/1990]) とした. なお, この文章の原文と目標要約率 25 % の要約例は付録 A を参照のこと.

目標要約率: 5 %

出力要約率: 8.9 %

国連安全保障理事会は国連諸決議の実効をあげるため、対イラク武力行使を認める新たな決議案の協議に入る。

例えば多国籍軍の増強が必要とみられるが、「国際警察軍」の維持費をだれが、どう負担すべきかについて明確にしなければならない。

目標要約率: 10 %

出力要約率: 14.2 %

国連安全保障理事会は国連諸決議の実効をあげるため、対イラク武力行使を認める

新たな決議案の協議に入る。

軍事力行使による解決策に反対するのならば、代替案を明確に表明しなければならないし、そのための外交努力を一段と強化する必要がある。

例えば多国籍軍の増強が必要とみられるが、「国際警察軍」の維持費をだれが、どう負担すべきかについて明確にしなければならない。

目標要約率: 15 %

出力要約率: 18.5 %

国連安全保障理事会は国連諸決議の実効をあげるため、対イラク武力行使を認める新たな決議案の協議に入る。

国連の重要メンバーであることを自認してはいるが、わが国として、どうこの国連外交の重大局面に対処するのか。

軍事力行使による解決策に反対するのならば、代替案を明確に表明しなければならないし、そのための外交努力を一段と強化する必要がある。

例えば多国籍軍の増強が必要とみられるが、「国際警察軍」の維持費をだれが、どう負担すべきかについて明確にしなければならない。

目標要約率: 20 %

出力要約率: 21.9 %

国連安全保障理事会は国連諸決議の実効をあげるため、対イラク武力行使を認める新たな決議案の協議に入る。

米国や英国政府内部には、紛争関係国の要請と合意にもとづく対イラク武力行使はできるという見解がある。

軍事力行使による解決策に反対するのならば、代替案を明確に表明しなければならないし、そのための外交努力を一段と強化する必要がある。

例えば多国籍軍の増強が必要とみられるが、「国際警察軍」の維持費をだれが、ど

う負担すべきかについて明確にしなければならない。そうした対応を欠けば、湾岸危機は欧米のニッポンただ乗り論の火に油を注ぐ結果になりかねない。

目標要約率: 25 %
出力要約率: 25.8 %

(付録 A を参照)

目標要約率: 30 %
出力要約率: 33.6 %

国連安全保障理事会は国連諸決議の実効をあげるため、対イラク武力行使を認める新たな決議案の協議に入る。

米国や英国政府内部には、紛争関係国の要請と合意にもとづく対イラク武力行使はできるという見解がある。しかし、現実には国連での合意形成を図る方針をとろうとしている。湾岸紛争をめぐる国際政治や軍事情勢がそうした選択を必要としているのであろうが、わが国の立場からもこれは望ましい方向だ。

国連の重要メンバーであることを自認してはいるが、わが国として、どうこの国連外交の重大局面に対処するのか。

軍事力行使による解決策に反対するのならば、代替案を明確に表明しなければならないし、そのための外交努力を一段と強化する必要がある。

例えば多国籍軍の増強が必要とみられるが、「国際警察軍」の維持費をだれが、どう負担すべきかについて明確にしなければならない。そうした対応を欠けば、湾岸危機は欧米のニッポンただ乗り論の火に油を注ぐ結果になりかねない。

目標要約率: 35 %, 40 %, 45 %, 50 % (**同一結果**)

出力要約率: 52.3 %

国連安全保障理事会は国連諸決議の実効をあげるため、対イラク武力行使を認める新たな決議案の協議に入る。

第二はソ連、フランス、中国の安保理常任理事国が米英の武力行使決議提案に対し、紛争解決のため軍事力オプションの必要性を認めつつも、なお平和的解決を追求する立場に立つという中立的姿勢をとっていることである。武力行使が必要という考えに傾いている米国、サウジアラビアなどの政府当局者からみれば、この国連決議を通すことは、坂道で重い荷車を引くに等しい困難が伴うことだろうが、国際紛争の解決に国連の機能を使うためには、そうした忍耐は決して避けて通れないものである。

米国や英国政府内部には、紛争関係国の要請と合意にもとづく対イラク武力行使はできるという見解がある。しかし、現実には国連での合意形成を図る方針をとろうとしている。湾岸紛争をめぐる国際政治や軍事情勢がそうした選択を必要としているのであろうが、わが国の立場からもこれは望ましい方向だ。

国連の重要メンバーであることを自認してはいるが、わが国として、どうこの国連外交の重大局面に対処するのか。

軍事力行使による解決策に反対するのならば、代替案を明確に表明しなければならないし、そのための外交努力を一段と強化する必要がある。

例えば多国籍軍の増強が必要とみられるが、「国際警察軍」の維持費をだれが、どう負担すべきかについて明確にしなければならない。そうした対応を欠けば、湾岸危機は欧米のニッポンただ乗り論の火に油を注ぐ結果になりかねない。

付録 C: 複数記事の要約実験に使用した記事とその要約結果の例

ここでは, 論文の第 4章で述べた手法による要約の例を示す.
まず、以下に第一記事を示す.

二十七日付の米ウォールストリート・ジャーナルは米国防産業大手のゼネラル・ダイナミックス (GD) がジェット戦闘機部門をロッキード社に売却する計画だ、と報じた。 GD は報道について「うわさにはコメントできない」としているが、業界関係者の間では GD はボーイングとも交渉しているとの説が有力になっている。 GD の同部門は三菱重工業と次期支援戦闘機「FSX」を共同開発中でプロジェクトへの影響も予想される。

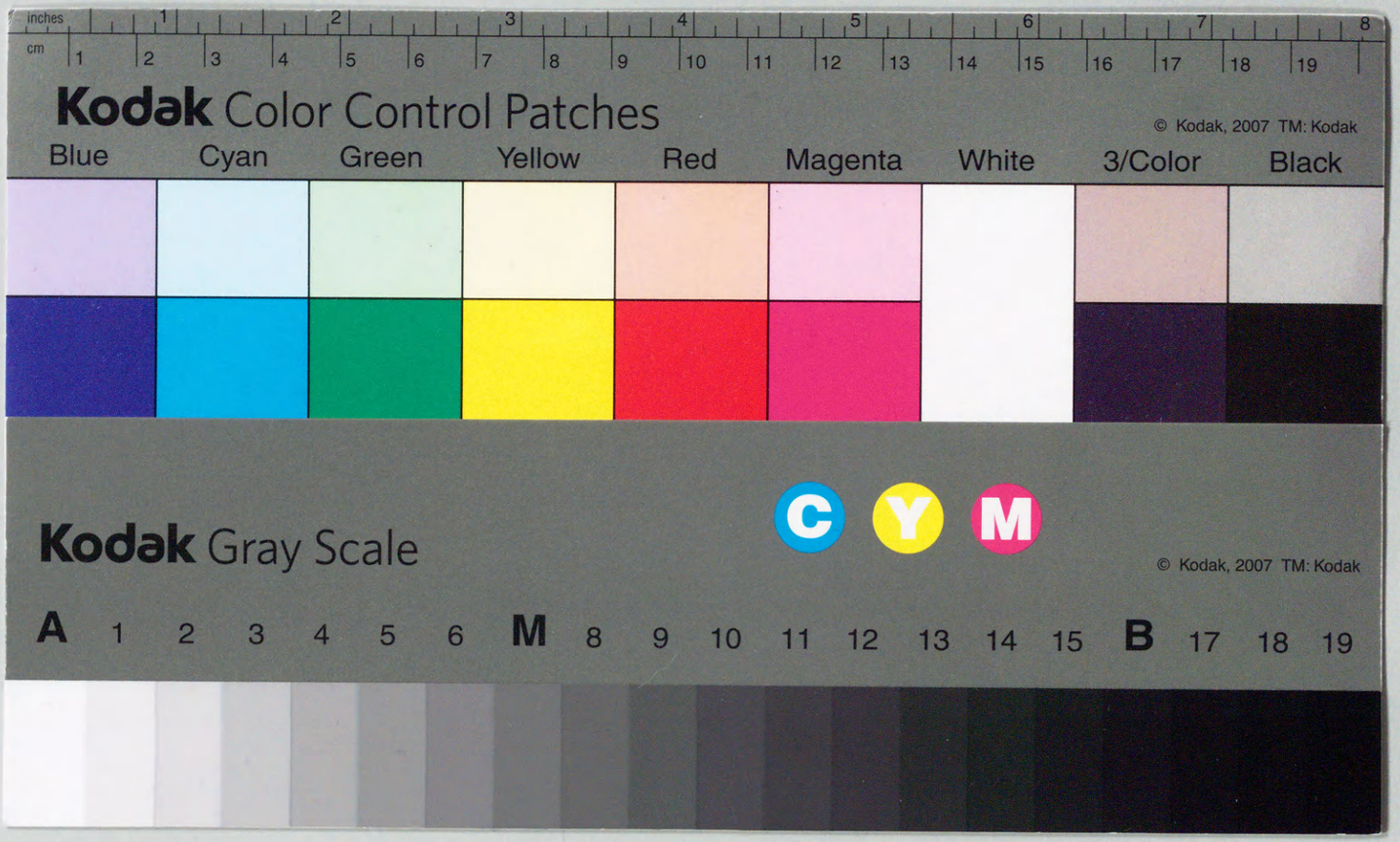
以下に要約の対象となる第二記事を示す.

米防衛大手のロッキードが九日、ゼネラル・ダイナミックス (GD) の戦闘機部門を買収すると発表したが、この M&A(企業の合併・買収) により日米で共同開発中の次期支援戦闘機 (FSX) の計画に影響が及ぶ可

能性が出てきた。共同開発の米側の担当企業だったGDから計画を引き継ぐロッキードの今後の方針次第では、工場の統廃合や日米間の技術移転などで新たな展開も予想される。

これに対して, 第二記事の要約結果を示す.

米防衛大手のロッキードが九日、GDの戦闘機部門を買収すると発表したが、共同開発中のFSXの計画に影響が及ぶ可能性が出てきた。ロッキードの今後の方針次第では、工場の統廃合や日米間の技術移転などで新たな展開も予想される。


inches
cm
Kodak Color Control Patches
© Kodak, 2007 TM: Kodak
Blue
Cyan
Green
Yellow
Red
Magenta
White
3/Color
Black
Kodak Gray Scale
C
Y
M
© Kodak, 2007 TM: Kodak
A 1 2 3 4 5 6 M 8 9 10 11 12 13 14 15 B 17 18 19