論文 / 著書情報
Article / Book Information

題目(和文)	携帯端末への情報提示を指向したテキスト要約に関する研究
Title(English)	
著者(和文)	長谷川隆明
Author(English)	Takaaki Hasegawa
出典(和文)	学位:博士(工学), 学位授与機関:東京工業大学, 報告番号:甲第8074号, 授与年月日:2010年3月26日, 学位の種別:課程博士, 審査員:奥村 学
Citation(English)	Degree:Doctor (Engineering), Conferring organization: Tokyo Institute of Technology, Report number:甲第8074号, Conferred date:2010/3/26, Degree Type:Course doctor, Examiner:
学位種別(和文)	博士論文
Type(English)	Doctoral Thesis

携帯端末への情報提示を指向した
テキスト要約に関する研究

東京工業大学大学院

総合理工学研究科

知能システム科学専攻

指導教員　奥村 学 教授

長谷川 隆明

2010年2月

目 次

第1章　序論

1.1　携帯端末とテキスト要約

インターネットに接続可能な携帯電話や PDA 等の小型で持ち運びが容易な携帯端末の普及により，我々はいつでもどこでもネットワークの上にある情報にアクセスすることが可能になった．例えば，パソコンで受信している自分宛の電子メールを携帯電話に転送することにより，外出先でも携帯電話を使って電子メールの内容を確認することができたり，携帯電話から Web 検索サービスを利用することにより，外出先であっても見知らぬ事物について調べたり欲しい情報を手に入れたりすることができるようになった．電子メールについては，携帯電話の各キャリアはプッシュ型の電子メールサービスを提供しているため，各キャリアのユーザのメールアドレス宛の電子メールは直ちにそのユーザの携帯電話に通知される．これにより，ユーザはパソコンのメールアドレス宛の電子メールを携帯電話に転送することで，パソコンのメールアドレスに届いた電子メールの着信通知として携帯電話を有効に活用できる状況が生まれた．Web 検索については，携帯端末のための Web 検索サービスの進展は目覚しく，携帯端末で閲覧することを前提とした Web ページが増加したことに伴って，今日では携帯端末向け Web ページ専用の検索サービスが実用化されている．さらに，パソコン向けの Web ページであっても携帯端末向けに Web ページを変換して提示する検索サービスが登場するなど，多くのユーザが携帯電話から Web ページを検索する状況に至っている．

しかしながら，携帯端末の画面はパソコンに比べると小さく，一度に表示できる文字数が限られている．このため，携帯端末に電子メールや Web ページの検索

結果をパソコンと同様に表示させると画面のスクロールが頻発するため，パソコンに表示するテキストよりも文字数を少なくすることが必須となる．単純に文字数を少なくするには，電子メールであれば先頭からある一定の長さまでを提示したり，Web 検索であればユーザが入力したキーワードの周辺をある一定の長さだけ抜粋して提示することが考えられる．ところが，このような単純な方法では，重要な情報が盛り込まれる保証がないために内容が十分に伝わらなかったり，テキストを構成する文を途中で切断してしまうことにより読みにくくなるという問題を内包している．このように，テキストの文字数は短くしながら，その一方で内容がわかりやすく読みやすいテキストを提示することは，携帯端末を利用したサービスにとって重要な課題である．

本論文では，携帯端末へ情報を提示するためのテキスト要約技術[1]について論じる．要約とは,「文章などの要点をとりまとめて，短く表現すること，また，そのとりまとめた言葉や文.」と定義されている（岩波書店広辞苑第五版）．この定義から，人間による要約は，文章の内容を理解し，どのような情報が要点なのかを捉え，それを新たな短い文章で記述するという処理が行われていると考えられる．一方，機械による要約は，与えられた文章から何らかのスコアに基づいて重要な文を抽出したり，元の文から何らかのスコアに基づいて不要な部分を削除することにより文を短縮したりする処理が一般的である．機械にとって真に文章を理解することは困難なため，機械による要約では伝統的にこのような処理を行う方針が取られてきた．このアプローチは，テキストからの抜粋に基づくため抽出的な要約であるといえる．

テキスト要約において考慮すべき情報は，下記の通りである．

対象　どのような種類のテキストを要約の対象とするのか．対象とするテキストにはどのような特徴があるのか．

目的　どのような目的・用途に要約を利用するのか．要約の利用目的には大別して

[1] テキスト要約全般について解説した体系的な教科書 [34] が出版されている．

表 1.1: 本論文で扱うテキストと考慮すべき情報

対象テキスト	電子メール（ビジネス利用）	Web ページ
利用目的	要約による携帯端末への着信通知	携帯端末向けの検索用概要文
指示的/報知的	指示的，報知的	指示的
generic/query-biased	generic	query-biased
要約長	500byte[2]	70byte

指示的な要約と報知的な要約がある．指示的な要約は，原文を読むべきかどうかを判断するために原文を参照する前の段階で用いるものである．報知的な要約は，原文は読まずに原文の代わりとして用いるものである．また，利用目的についての別の観点では，要約を利用するユーザの違いによって，誰もが汎用的に利用できる generic な要約なのか，あるいは，特定のユーザが特定の目的に利用するための query-biased な要約なのかにも分類することができる．

要約長（要約率）要約の長さはどの程度なのか．どの程度元のテキストを短くする必要があるのか．

これらは相互に関連し合う情報であり，何を要約の対象とするか，どのような目的で要約するのか，どの程度の長さの要約が必要なのかによって，最適となるテキスト要約の技術も異なるであろう．本論文では，実社会において有用となるアプリケーションのためのテキスト要約の技術を論じる．実社会への導入を考えれば，できるだけ人手によるコストをかけなくても精度を上げることが必要となる．できるだけ人手によるコストをかけないという設定は，正解となるデータを大量に用意するのが困難な状況にも対応することができる．

本論文で扱う要約の対象，目的と要約長について表 1.1 にまとめ，詳細を以下に

[2]必要に応じて原文を参照するために必要な情報を含む．

述べる．本論文で対象とするテキストは電子メールと Web ページである．これらのテキストはユーザ個人が持ち歩く携帯端末と親和性が高い．電子メールも Web ページも一般のユーザによって書かれる事が多く，新聞記事などのメディアによるテキストに比べると，テキストの質は高くないと考えられる．

電子メールについては，記載されている内容やスタイルは多種多様であるため，すべての電子メールを一律に扱うことは難しい．そこで，電子メールを要約する目的に応じて，内容やスタイルを限定することにする．電子メールを要約の対象にしたときの目的は，ユーザが受信した電子メールの要約を携帯端末へ表示したり，合成音声を用いて電子メールの要約だけを読み上げることである．特に，ユーザが受信した電子メールの要約を携帯電話に転送することは，要約による着信通知という新たなコンセプトをも生み出す．このコンセプトに調和する将来の利用シーンとして，例えば外出先のビジネスマンが要約による着信通知を受けて，それが重要な電子メールかどうかを判断し，重要な電子メールのみ原文を参照させることが考えられる．このような利用シーンを見据えてビジネス用途の電子メールを対象とする．多くの電子メールを受信するユーザにとっては，上述の利用シーンのような指示的な要約だけに留まらず，原文を参照することが困難な場合には報知的な要約として用いられるであろう．また，電子メールは事前に受信する内容が想定できない場合が多いため，どのようなユーザに対しても汎用的に有用な generic な要約を目指す．携帯端末の画面の解像度によってスクロール回数に多少の変動はあるが，要約長は最大 500byte で固定とする[3]．受信した電子メールに対して，不要な部分を取り除いたり重要な部分だけを取り出したりすることにより，携帯端末に提示するのに適した要約長でその内容を適切に要約することを目標とする．

一方，Web ページについては，対象は任意の Web ページとし，Web ページの

[3]携帯電話のキャリアがメールサービスを開始した当初の数年は，物理的にも 500byte の長さのメッセージしか送受信できなかったことによる．

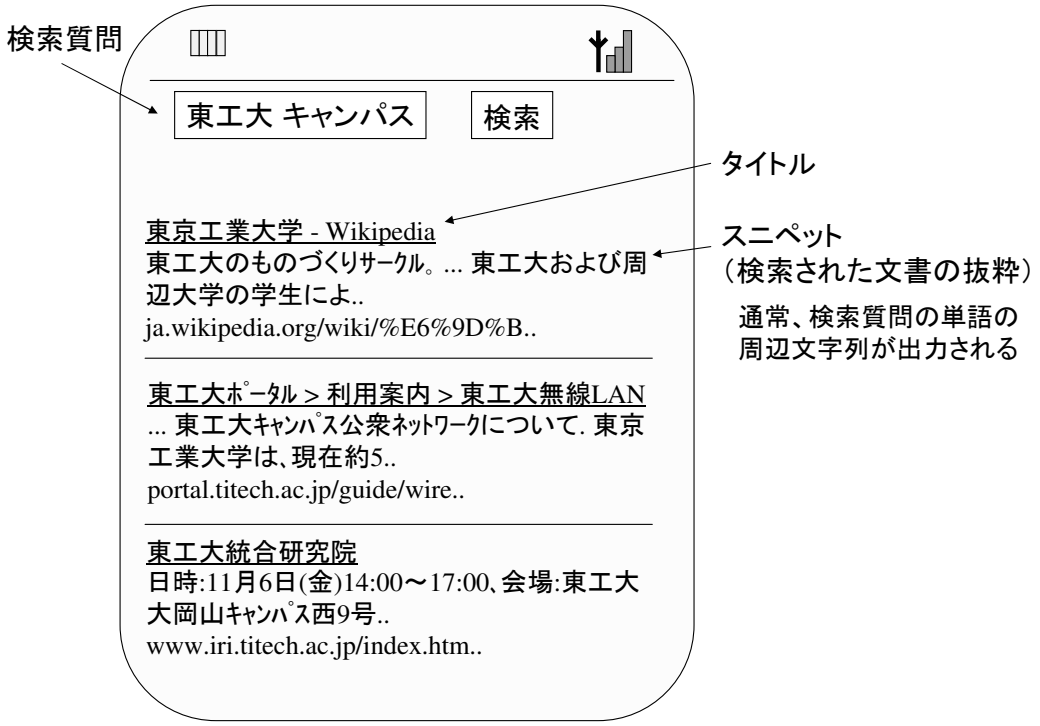

図 1.1: 携帯端末向け検索エンジンにおけるパソコン向けページの情報提示の例

検索結果の提示を目的とする．すなわち，Web ページの概要文（スニペット）の生成を目標とする．Web ページの概要文は，ユーザに提示される情報のひとつであり，指示的な要約として利用されるため，概要文の良し悪しが検索エンジンの使い勝手に大きな影響を与える．多くの検索エンジンでは，図 1.1 に示すように，ユーザが入力した検索キーワードをできるだけ多く含むように，KWIC(keyword-in-context) による方法を用いて Web ページから検索キーワードの周辺の文字列を切り出したものを Web ページの概要文として提示していると推測できる．パソコンで読むことを前提とした概要文では概要文の長さに余裕があるため，KWIC による方法でも検索キーワードの周辺から多くの文脈を得ることができるので比較的実用に耐えると考えられる．しかしながら，携帯端末で読むための短い概要文の場合は，KWIC による方法では主に 2 つの問題が生じると考えられる．ひとつは，概要文の読みやすさが損なわれる問題である．概要文が単語の途中の文字から始まったり単語の途中の文字で終わったりする欠点は，短い概要文では相対的に目に付きやすくなるからである．もうひとつは，概要文は Web ページのタイトルといっしょにユーザに提示されるため，タイトルの内容と重複が生じる場合には，概要文の持つ情報量が低下しユーザに有益な情報を提示できないという問題がある．なぜならば，検索結果の上位にランクされる Web ページにおいては，Web ページのタイトルにも検索キーワードを含むことが多く，タイトルと同等の内容を表す検索キーワードの周辺の文脈を概要文として提示することは十分に起こりえるからである．また，Web ページの概要文は，ユーザの入力した検索キーワードに沿った要約であることから，query-biased な要約であるといえる．要約長においては，携帯端末に表示する検索結果の 1 ページにおいて，多くの Web ページのリストを載せるには概要文はできるだけ短い方がよいという条件と，概要文の読みやすさとわかりやすさを優先すると概要文はできるだけ長い方が望ましいという条件とのトレードオフの関係がある．要約長を決定するために，商用の携帯端末向け検索サービスで使われている長さに合わせて，要約長は最大 70byte で固

定とする．元のテキストの長さにより変動はあるものの，要約率としては極めて低い要約（概要文）が要求される．携帯端末で読むための短い概要文であっても，Web ページの内容がわかりやすく読みやすい概要文を生成することを目標とする．

1.2　テキスト要約と情報抽出

テキスト要約には，先に述べた抽出的な要約のアプローチの他にも，元のテキストの内容を言い換えたり新たに文を合成したりすることで単なる抜粋ではない要約を生成するアプローチがある．これはアブストラクトによる要約のアプローチといえる．このようなアプローチには，テキストから重要な情報を発見し，それらを統合し，生成するという過程が必要となる．テキストから重要な情報を発見するためには，情報抽出という技術が利用できる．情報抽出とは，予め必要な情報をテンプレートとして用意し，テンプレートを埋めるようにテキストから情報を抽出する技術である．このような処理は，任意のテキストを対象とした場合には困難であるが，限られたドメインのテキストだけを対象にすればある程度は可能になってくる．この場合，情報抽出はアブストラクトによる要約のアプローチの過程の一部を担うことになる．

関連する複数のテキストをまとめて要約するような複数文書要約を必要とするアプリケーションにおいては，アブストラクトによる要約のアプローチは非常に効果的であろう．さらに，抽出した情報から新たなテキストを生成することは，不要な情報を含む可能性がないために，携帯端末向けの短い要約を作成する目的には有利である．複数の文書をひとつのテキストに要約する例として，Web 検索サービスにおける検索結果に対して，類似した内容の Web ページを取りまとめ，要点だけを携帯端末に提示するようなタスクを考える．類似した内容の Web ページには同一の事象に言及していることが考えられるが，同一の事象に言及する表現は多様である．これを言い換え（パラフレーズ）と呼ぶ．情報抽出を用いた要約は，

これらの言い換え表現を認識してひとつの表現にまとめることが可能である．例えば，A という会社を検索するときに，B という会社を買収したことに言及する Web ページがたくさん検索され，それぞれが「会社 A は会社 B を買収するために○○円を投じた」「会社 B は会社 A に吸収合併された」「会社 B は会社 A の買収提案を受け入れた」等の異なる表現で記述されている場合を考える．もし，これらの表現が同一の事象に言及していることが認識できなければ，それぞれの Web ページに対して同じような内容の要約が複数個生成されてしまうであろう．このとき，情報抽出により買収企業，被買収企業，買収総額の情報が得られれば,「会社 A が会社 B を総額○○円で買収した」という同一の事象に言及する代表的な一文によって，複数の Web ページの内容をまとめて簡潔に携帯端末に提示することができるだろう．しかしながら，このようなアプローチでは，生成したい要約に対して，抽出すべき情報のテンプレートを事前に決定しておくことが必要になる．

情報抽出の課題は，事前に準備しておくべきテンプレートをいかに低コストで獲得するかである．人手をかけてたくさんの正解事例を準備するのは大変な時間と労力がかかる上に，例え正解事例を準備したとしても，別のドメインではその正解事例がまったく役に立たないことも考えられる．テンプレートの属性値となる情報は，テキストの要点を抽出する上で鍵となる人名や地名などの固有表現や数値情報であることが多く，固有表現を抽出する技術の精度はほぼ実用に達している．テンプレートを獲得するためにさらに必要となる情報のひとつに，固有表現と固有表現の間にある関係がある．例えば，”Barack Obama”（バラク・オバマ）という人名と”United States”（アメリカ合衆国）という組織名の間に”president of”（大統領）という関係がテキスト中に存在することを認識できれば，それを用いて人物に関するテンプレートを作成できる．そして最終的にはそのテンプレートを利用した情報抽出を通して，アブストラクトによる要約が生成可能となる．そこで，本論文では，アブストラクトによるテキスト要約のアプローチに必要な情報抽出を取り上げ，2 つの固有表現の間にある関係の情報を低コストで抽出するため

に，大規模なテキストコーパスを用いた教師なし学習の手法についても論じる．

1.3　本論文の構成

本論文の構成を以下に述べる．

第 2 章では，電子メールに対する要約のアプローチ，Web ページに対する要約のアプローチおよびアブストラクトによる要約に必要な情報抽出のアプローチについて述べる．本論文では，電子メールに対する要約のアプローチとして，電子メールの特徴を利用した重要文抽出を提案する．そこで，まず電子メールの特徴について述べ，次に電子メールに特化した要約の先行研究について概観する．続いて本論文では，Web ページに対する要約のアプローチとして，Web ページからユーザが入力した検索語に対して適合する文を抽出し，さらに抽出した文そのものを短縮する文短縮を行う手法を提案する．検索結果のための概要文生成は，ユーザからの検索要求に沿って Web ページを要約する．このことから，誰でもどんな目的にも汎用的に使える generic な要約に対して，ユーザに適応した動的な要約である query-biased な要約とみなすことができる．そこで，まず query-biased な要約についての先行研究を概観し，次に文短縮についての先行研究も概観する．最後に本論文では，テキスト要約のための情報抽出のアプローチとして，大規模なテキストコーパスを利用して教師なし学習により 2 つの固有表現の間にある関係を抽出する手法を提案する．そこで，まず関係抽出という概念やそれに関するタスクについて説明し，関係を抽出するための教師付き学習を用いた先行研究について述べる．

第 3 章では，電子メールにおける重要文抽出の手法を提案し，携帯電話向け要約システムへの適用について議論する．提案する重要文抽出の手法は，電子メールに特有な引用や署名，箇条書きなどのスタイルの特徴と文分割に基づいて抽出した文から，電子メールに特有な表現の特徴を利用したルールに基づいて重要文

を抽出する方法である．さらに，電子メールの要約のための利用シーンを踏まえ，本手法を適用した携帯電話向け要約システムについて説明する．評価では，人手による正解との比較および着信通知タスクにおける評価について報告する．考察では，実際に携帯電話向け要約システムを利用したユーザからのアンケートの結果から，システムの有効性を議論する．また，人手による重要文抽出の正解を分析した結果に基づいて，重要部分として選択する情報の単位について議論する．

第 4 章では，携帯端末を利用した Web ページの検索において，ユーザの検索要求に沿って Web ページの概要文を生成するために，重要文抽出と文短縮に基づく手法を提案する．提案手法では，まず，Web ページ内の各単語においてユーザが入力する検索キーワードとの相関に基づく単語重要度を計算し，文内の単語重要度の総和が大きくかつ Web ページのタイトルと内容の重複が少ない文を抽出する．次に，抽出された文の文節の係り受け構造に基づく部分木の集合から，各文節内の単語重要度に基づく文節網羅率と文節同士のつながりやすさである文節連接確率に基づいて，最適な部分木を選択することで文を短縮する．評価では，検索のための評価として概要文による適合性判定の結果と，要約としての評価として文法性，非冗長性，内容網羅性の結果について報告する．考察では，Web ページの検索結果に対して議論する．また，Web ページの一種である QA サイトに限定した場合を想定すれば，QA サイトの検索結果を携帯端末に提示するためには，質問と回答の対からなるテキストを短くする必要がある．質問については最初の一文を提示すれば十分であると考えられるが，回答についてはそれでは十分でない場合が多いと考えられるので要約を生成する必要がある．提案手法を質問と回答の対からなる FAQ データに適用し，携帯端末向けの回答の要約における有用性についても議論する．本提案手法の計算量は，概要文の生成のために用いられることが多い KWIC による方法に比べれば多くなるが，事前にオフラインで処理できる計算が多いと考えているため，計算量の問題については論じない．

第 5 章では，アブストラクトによる要約に必要な情報抽出のひとつとして，大

規模なテキストコーパスから教師なし学習により 2 つの固有表現の間に存在する関係を抽出する手法を提案する．提案手法では，固有表現の対とその間に出現する文脈を収集し，収集された文脈同士の類似度を計算し，計算された類似度に基づいて各固有表現の対をクラスタリングする．また，同じクラスタにおいて多くの文脈に共通する単語は，2 つの固有表現の間に存在する関係を表すラベルとみなす．評価では，固有表現の対を対象としたクラスタリングの精度と各クラスタにおける関係のラベル付与の精度について報告する．考察では，評価実験で対象とした関係の特性，収集する文脈の長さ，クラスタリングの方法，低頻度の固有表現の対に関して議論する．

第 6 章では，本論文の結論を述べ，今後の課題について展望する．

第2章　テキストの特徴と要約のアプローチ

2.1　電子メールに対する要約のアプローチ

電子メールの要約は，これまで大きく分けて 2 つの方向で議論されてきた．ひとつは，電子メールの交換により展開される議論を整理して要約する研究 [19, 28, 23] である．これらの応用として，ソフトウェアの開発を支援するためのツール [19] や教育における学習者間のコミュニケーションの支援 [28] がある．もうひとつは，携帯電話等への配信を目的とした要約率の低い要約の研究であり，携帯端末を意識したコンテンツ要約技術 [30] として近年注目されている．本研究の対象も後者であり，以下では電子メールの特徴を踏まえながら従来技術を振り返る．

電子メールは，コミュニケーションの手段であるので，新聞記事のように一方的に何かを伝えるだけでなく，相手への質問や依頼・要求などを含むためインタラクションが起こることが特徴である．また，電子メールのテキストの質はそれほど高くはないことと，改行や空白などを用いてレイアウトが施されていることも特徴的である．携帯端末を前提とした制限文字数がかなり限られた要約を生成する上において，これらの特徴こそ重要な差異化ポイントとなり，これらの特徴を考慮した要約手法が必要となる．言い換えれば，単語の頻度を利用した伝統的な重要文抽出 [25] や構文情報を用いた手法 [16, 26] を適用しても，携帯端末にとって適切な要約を生成することは難しい．さらに，電子メールには私的な個人情報が含まれるため，電子メールを大量に収集し要約としての正解を用意することが困難である．このため，Web ページのヘッドラインを生成するために用いられる

統計的な手法 [5, 3] も簡単には利用できない．また，機械学習により特徴的な単語を抽出する方法 [29] では，要約というよりは単語の羅列を生成してしまう．電子メールに特化した要約として，単語の文字列を同義のより短い文字列に置換する方法 [22, 10] や電子メールの文面から意図の分類を行った後で，その意図を含む表現をより多く含むパラグラフを抽出する方法 [15] が提案されている．前者は要約率を優先するため文字列を記号などにも置換するため可読性が低下するという課題がある．後者は抽出の単位がパラグラフなので，可読性は良いが重要な部分が抽出されるパラグラフから漏れるという問題を含んでいる．

電子メールには，本文以外にもヘッダに情報が存在する．見出し情報を用いる方法 [31] を用いてヘッダのサブジェクトを利用して要約を生成することも考えられる．しかしながら，サブジェクトがそのまま変わらずにリプライが繰り返されると本文の内容が当初のサブジェクトに合致していた内容から離れていくため，サブジェクトが必ずしも適切でなくなるという問題が発生する．

2.2　Web ページに対する要約のアプローチ

本論文で論じる Web ページの要約のスコープは，ユーザからの検索要求に沿って Web ページを要約する技術である．ユーザの検索要求に沿った要約は，ユーザに適応した動的な要約である．これは，誰でもどんな目的にも汎用的に使える generic な要約に対して，query-biased な要約と呼ばれる．Tombros ら [42] は，重要文抽出における単語の重み付けの過程において検索質問に重みを加えることにより，検索結果の出力における query-biased な要約の有効性を示した．また，森 [27] は検索質問そのものではなく，検索質問により検索された文書をクラスタリングする際に得られる尺度である情報利得が大きい単語を用いて重要文を抽出する手法を提案している．テキスト自動要約の評価に関するワークショップ Document Understanding Conference (DUC) においても，generic な要約に加えて query-biased な要約が取り

上げられてきた [11]．これらの方法では，新聞記事を対象としており，Web ページとはテキストの質が異なる．これらの方法を Web ページにそのまま適用した場合には次のような問題が生じると考えている．1) 検索質問に含まれる単語に重みを加える重要文抽出の手法を Web ページに適用した場合，メニューを表すような情報量の少ない文や Web ページの主題とは無関係な広告・宣伝を表す文を抽出してしまう可能性がある．2) リード文を重要と考えてスコアを与えるという位置情報を利用する手法では，タイトルと内容の重複した文が Web ページの先頭に存在する場合には，冗長な情報をユーザに提示するという問題が発生する．3) また，クラスタリングを用いる方法を Web ページに適用しても，複数の主題を含む Web ページから情報利得の高い話題語を適切に抽出することは困難である．

新聞記事以外を対象とした先行研究では，Berger ら [6] は，質問とその回答からなる FAQ コーパスを利用した統計的な要約手法を提案したが，学習のためのコーパスを必要とする．また，高見ら [41] は，ユーザの検索目的に適した概要文を Web ページから動的に再生成する方法を提案している．この方法では，概要文を生成する際に考慮する Web ページの数という軸と内容の包括要約性を考慮するかという軸を用いて概要文の生成方法を４つのタイプに分類している．そして分類ごとに異なる尺度で重要語を獲得し，それを利用した重要文抽出により検索目的ごとに異なる概要文を生成している．我々は，高見らの分類による 1 つの Web ページに対して検索語を含むという基準で抽出された断片から生成するタイプの概要文を対象とする．しかしながら，その分類で提案されている手法を改善するために，検索質問に対する質問拡張的なアプローチを取ることや Web ページのタイトルとの冗長性を排除することが必要であると考えている．そして，以上の先行研究は，携帯端末への出力を目的とした要約方法ではないため，要約の長さを極端に短くすることは必要なく，文書から検索質問に適した文を選択する重要文抽出に基づいた方法に留まっている[1]．

[1]第 4 章で述べるが，実験に用いたデータにおける最重要文 199 文の長さを測定した結果，特異値を外すために上位 10 文と下位の 10 文を除外したところ，平均値が 135 バイトで目標の要約長

一方，選択した文をさらに短くする文短縮の手法も提案されている．文短縮には構文情報を用いるアプローチと構文情報は用いないアプローチがある．構文情報を用いるものとしては，機械学習を用いた方法 [43]，[33] や Noisy Channel に基づく方法 [4]，コーパスからの統計を用いた [38] が提案されている．しかしながら，提案されている文短縮の方法はいずれも generic な要約を指向したものであり，query-biased な要約を生成するためには，検索要求ごとに別のコーパスが必要となるため，Web ページの概要文を生成するための文短縮にそのまま適用できるかは不明である．また，大森ら [35] は，コーパスを用いずに携帯端末向けに新聞記事を要約する方法を提案している．新聞記事は一般に重要な情報から順に記述されているという特徴を利用して，リード文に対する文短縮を行っている．リード文の係り受け解析木の末端にある文節の中で，tf*idf 値が最も小さい文節から順に削除することで目標の要約の長さまで短縮する．しかしながら，この手法は generic な要約を生成することを目的としたものであり，本論文の目的とは異なる．また，200 文字から 300 文字の新聞記事に対して，目標の要約の長さを 50 文字または 100 文字としているため，要約率としては 17%から 50%の設定である．しかしながら，多くの Web ページの検索サービスでは，検索結果の 1 画面に複数の Web ページの概要文を並べることから，携帯端末に表示できる概要文の長さはこの設定よりもさらに短い要約が求められる[2]．また，構文情報を用いないものとしては，単語を接合する方法 [18, 17]，あるいはルールに基づく方法 [44] が提案されている．これらの方法はロバストではあるため，query-biased な要約に適用できる可能性があるが，対象となる文書の種類は音声認識結果や新聞記事であり，文書のスタイル情報が講演や報道という均質な状況であることを仮定している．このため，様々

である 70 バイトを越える重要文は 69%を占めた．このことから，携帯端末への出力を目的とした場合に，重要文抽出だけでは不十分であり，抽出した文をさらに短縮する必要がある．

[2]第 4 章で述べるが，実際に我々の用いたデータでは，目標の要約の長さを 70 バイト（全角 35 文字）としたが，このときの要約率の平均は 2.56%であった．新聞記事の要約率とは一桁違っていたことから，新聞記事よりもさらに難しいタスクであると言える．

なスタイルが混在している Web ページに適用するには，Web ページそれぞれに特有のスタイル情報を獲得することが必要になるであろう．近年，Web ページの中でもテキストの質が高くないとされるブログに対して，文節の係り先が存在しない自己係りを導入することによる係り受け解析器が提案されている [21]．係り受け解析の精度が向上すれば，Web ページの文短縮においても構文情報を用いることは有用であると考えている．

以上のように，本論文では Web ページを対象とした query-biased な要約を指向し，重要文抽出だけでなく文短縮までもが必要な携帯端末向けの非常に短い概要文を生成することを目的とする．このような目的の先行研究はこれまでに存在しない．

2.3　テキスト要約のための情報抽出のアプローチ

アブストラクトによる要約では，Radev らは特定の話題に関する複数のニュース記事から要約する手法を提案している [36]．この手法では，特定の話題としてテロリストを取り上げて，各記事から事件発生日や事件発生場所，事件のタイプ，犯人，死傷者数を抽出し，抽出された情報を統合し，自然言語生成システムにより要約を生成している．また，Carenini らはレビューのような評価情報を含む複数のテキストに対して，アブストラクトによる要約と抽出的な要約を比較し，意見を扱う要約においてはアブストラクトによる要約が有利であることを主張している [9]．アブストラクトによる要約のアプローチでは，まずテキストから情報抽出によって重要な情報を抽出することが必要であり，情報抽出のためのテンプレートを獲得するコストを下げることが重要な課題である．以下では，テキスト要約のための情報抽出のアプローチに関して，情報抽出のためのテンプレートを構成する情報のひとつとして，テキスト中の人名や地名等の固有表現の間に存在する関係に焦点を当て，テキストから関係を抽出するタスクに関する先行研究につい

て述べる．

関係抽出というタスクは，1995 年に米国 DARPA(Defense Advanced Research Projects Agency) 主導による MUC-6 (6th Message Understanding Conference)[12] において，情報抽出のタスクのひとつである Template Element Task で初めて導入され，MUC-7 にて Template Relation Task が追加された．その後，情報抽出は米国の評価型会議 ACE (Automatic Content Extraction)[32] に引き継がれ，2002 年より RDC (Relation Detection and Characterization) タスクが設定されている．ACE RDC タスクでは，カーネル法のような機械学習によるアプローチ [45] が主流である．このアプローチは人手により関係についての情報が付与されたタグ付きコーパスを必要とする．しかしながら，人手により関係の情報をタグ付けすることは容易ではなく，教師有り学習のアプローチにはタグ付きコーパスの作成に多大な時間とコストがかかるという大きな問題が存在する．さらに，関係の種類は ACE によって定義された種類に限定されるため，様々なテキストから関係を抽出するためには多くの関係に対する定義が必要になる．このように，タグ付きコーパスに基づく教師有り学習は，多くのコストをかけてでも極めて高い精度を追求する場合には有効であるが，低コストで網羅的に関係を抽出するという目的にはそぐわない．

タグ付きコーパスを用いないアプローチとして，半教師有り学習を用いる方法が提案されている．Brin はブートストラッピングを用いて関係を抽出する手法を提案した [7]．Brin の方法は，ある特定の関係にある少量の既知の事例から，その関係を表すパターンとそのパターンにマッチする新たな事例を，ブートストラッピングを用いて交互に獲得していく方法である．Brin は本の題名と著者という関係において少数の事例を用い，それらの事例を含む文脈から共通なパターンを収集し，共通なパターンにマッチする文脈を含む本の題名と著者の新たな事例を獲得した．Agichtein らは Brin の手法を改良し，固有表現抽出器を用いるという制約を導入した [1]．Ravichandran らは質問応答のために類似した手法を提案してい

る [37]．しかしながら，これらのアプローチには，最初に少数のシードを必要とする問題がある．どのようにシードを選択するのか，あるいは，どれくらいの量のシードが必要なのかが明確ではない．また，シードとして与えられる固有表現の対の間に存在する既知の関係に対象が限定される問題も存在する．テキストから幅広く関係を抽出するという目的に対しては，関係の種類を制限しないことが望ましく，ブートストラッピングによる方法ではコーパスに存在する関係を網羅的に発見することは難しい．

同一の関係を示唆している表層的な表現は様々であり，これらの表現はパラフレーズ（言い換え）とみなすことができる．パラフレーズを獲得する先行研究としては，Lin らによる別の半教師有り学習の方法があげられる [24]．Lin らの方法は，動詞句とその構成要素である主語と目的語に注目し，類似した構成要素を持つ 2 つの動詞句はパラフレーズであるとみなすものである．しかしながら，この方法も類似した動詞句を見つけるために，最初に少数の動詞句をシードとして与える必要がある．

第3章　電子メールの要約

3.1　電子メールにおける重要文抽出

本章では，電子メールに特化した重要文抽出の手法について述べる．携帯端末向けの電子メールの要約では，携帯端末に提示するのに適した要約長（制限文字数）において，可読性と内容網羅性のバランスの取れた要約を生成することが課題となる．第 2 章で述べたような電子メールに特化した従来の要約手法でも，この課題を十分に解決しているとは言えない．そこで，可読性と内容網羅性のバランスの取れた要約を生成するために，電子メールの特徴を利用した重要文抽出の手法を提案する．

提案手法と従来手法の差異化ポイントは次の 2 点である．ひとつは，電子メールの重要部分の選択の単位である．従来手法では，電子メールの重要部分の選択の単位はパラグラフまでで，さらに細かい単位に分けることは行っていない．本研究では，意味の取れる範囲で最も細かい文を重要部分の選択の単位とする．特に１つのパラグラフが長い場合には，できるだけ原文の内容をカバーしながら要約率を下げることは難しいので，重要部分の選択の単位を文とすることは有効だと考えられる．電子メールには引用や署名あるいは箇条書の混在などの独特のスタイルがあるため，重要文を抽出する前に電子メールのスタイルの特徴を考慮して文を切り出しておく必要がある．

もうひとつは，電子メールに特有な表現の特徴を利用した重要文抽出である．電子メールの多くは，特定の個人や集団の相手に対する連絡である．連絡には，依頼や質問あるいは通知や報告といったはっきりした目的が存在していることが多

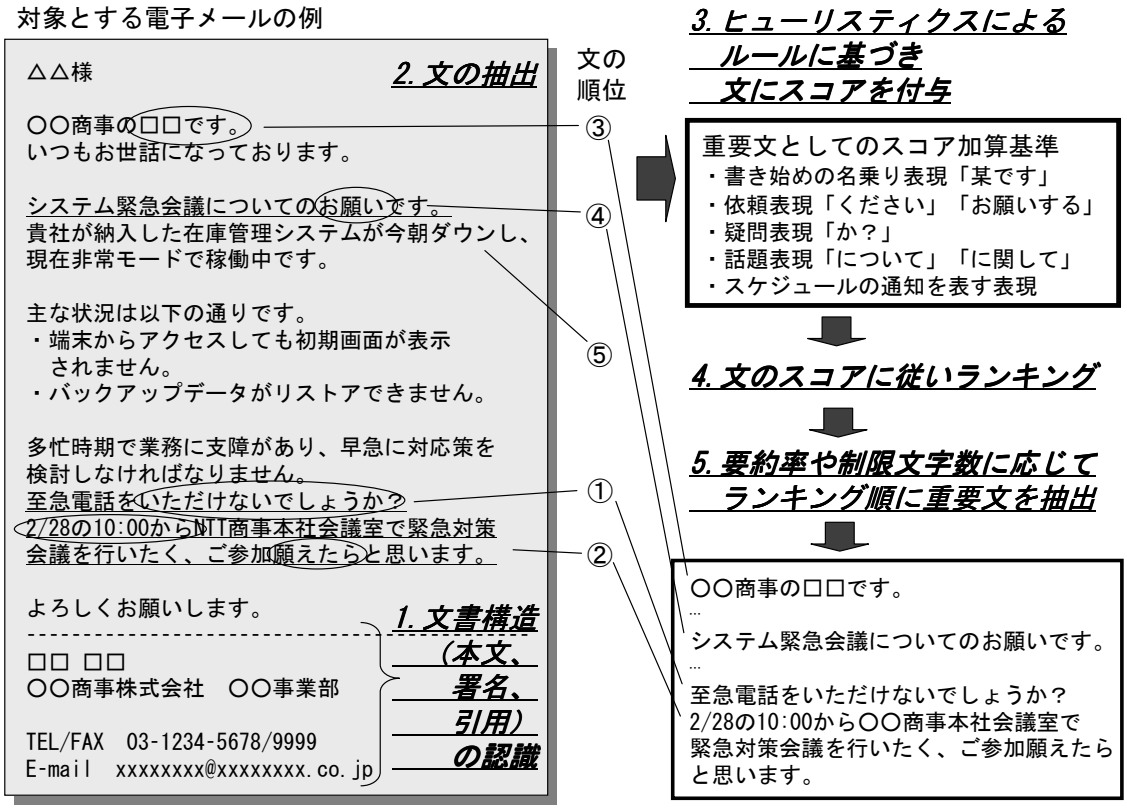

図 3.1: 重要文抽出のステップ

く，本研究ではその範囲の電子メールを対象とする．このような目的を持った電子メールには，目的を表すための特有の表現が存在する．また，電子メールの多くには，送信者を特定するための名乗りの表現や時候の挨拶が存在するなど独特の表現が見られる．本研究では，電子メールに特有な表現の特徴に着目し，これをルールとして記述した上でルールにマッチした文を抽出することにより，重要文抽出を実現する．反対に，時候の挨拶や決まり文句など不要な表現もルールとして記述すれば，これらも不要な箇所としての手がかりとして利用でき，重要文抽出の精度を高めることができる．

本研究で提案する重要文抽出のステップを図 3.1 に示し，以下にまとめる．

ステップ 1 ヘッダや添付ファイルおよび署名や引用を除いて送信者が記述した本

文のみを重要文抽出の対象とするため，電子メールの文書構造を認識し本文と署名や引用を行ごとに区別する．

ステップ 2 行ごとに区別された文書構造に加え，本文に混在する箇条書の認定や行間の接続を行い，重要文抽出の単位となる文の抽出を行う．

ステップ 3 電子メールの目的に着目した重要文抽出を行うために，これらの表現を含む文に対してルールに基づきスコアを付与する．具体的には，抽出された各文に対する形態素解析 [14] の後，形態素単位に分割された文に対して，形態素パターンとそれに対応するスコアからなるヒューリスティクスによるルールを適用し，形態素パターンにマッチする文に対して設定されたスコアを付与する．

ステップ 4 スコアの高い重要文を選択するために文のスコアにしたがって文をランキングする．

ステップ 5 要約率や制限文字数に応じてランキング順に文を選択し，選択されない脱落文を記号「...」で置換して出力する．

以下ではそれぞれの処理について詳説する．第 1 節では，重要文抽出の前処理となるステップ 1 とステップ 2 について述べる．ヘッダや添付ファイルを除いた本文の持つスタイルの特徴を整理し，これらの特徴に基づいて抽出の単位である文をどのように切り出すのかについて述べる．第 2 節では，ステップ 3 について述べる．特に電子メールの目的に着目した重要文抽出を実現するためのルールについて説明する．第 3 節では，後処理となるステップ 4 とステップ 5 について述べる．ステップ 3 によるルールの適用後に制限文字数の中でどのように重要文を選択するかについて述べる．

3.1.1　スタイルの特徴と文分割

電子メールテキストのスタイルの特徴として以下があげられる．

- 引用や署名などの構造がある
- 文の途中であっても改行が挿入されることがある
- 文章と箇条書きが混在することがある

以上の特徴を持つ電子メールを対象とする場合には，重要な箇所を抽出する，すなわち重要文抽出を行う前に，文の単位で本文を切り出しておかなければならない．これを実現するために，次のようなステップを実行する．

1. 文書構造の解析による署名の削除

   署名は通常テキストの最後部に付いているので，末尾の行から上の方へ一定の行数だけ順に空行や記号のみの行あるいは行末が文末表現である行を検出する．検出された行の次の行から最後部までを署名と見なし [2]，これを削除する．

2. 引用記号に基づく引用行の削除

   各行に対して，行頭に英数字を含む引用記号が存在するかを調べることによって引用行を特定し，引用行を後述の処理対象から除外する．英数字のみの行やあらかじめ用意されたメールソフトにより付与される行も同様に除外する．ただし，コメントを表す記号を先頭に持つコメント行は除外しない．

3. 英数字や記号または空白とその位置に基づく箇条書き行の特定

   各行に対して，行頭に英数字やコロン等の記号が存在する箇条書きラベルを持つ箇条書き行を特定する．あるいは箇条書きラベルがない場合には，空白がある一定の行頭からの位置以内に存在し，空白より後の文字列がある一定の長さ以内である行を箇条書きラベルなしの箇条書き行として特定する．

4. 同じタイプの行の接続と文の切り出し

   箇条書き行と連続する箇条書き行の行頭位置を比較し後続行の方が大きければ箇条書き行を接続する．また，箇条書き以外の行と連続する箇条書き以外の行を接続する．文末表現および文末表現がなくても行末に一定の行頭からの位置以内の特定の助詞や助動詞相当の文字列があればそこで文を分割する．コメント行については，コメント行が連続する場合のみ，後続する行の先頭にあるコメントを表す記号を除いて接続する．後述する処理のため，1 行 1 文の形に整形しておく．

3.1.2　表現の特徴とルールに基づくスコア付与

不要な表現を削除しながら電子メールの連絡の目的を表している重要な表現を抽出するためのルールについて詳説する．重要な表現や不要な表現は，位置情報にも密接に関係している．例えば，時候の挨拶は本文の最初に存在する．前処理により抽出された各文は形態素解析により，形態素単位に分割され，以下で述べるルールが適用される．重要文抽出において位置情報を考慮するため，引用や署名を除いて 1 行 1 文の形式で抽出された本文は，ルールが適用された時に動的に先頭文と冒頭文，通常文の位置情報に区分される．

図 3.2 にルールの適用時に区分される電子メールの位置情報を示す．ルールは区分される位置情報に基づき以下の 3 種類に分類され，重要あるいは不要な表現を表す形態素パターンとそれが持つスコアから構成される．

先頭文ルール 本文の先頭から適用され，あらかじめ用意された時候の挨拶（「いつもお世話になっております。」等）や宛名表現（「△△様」等）にマッチする場合に適用され，その文のスコアを下げる．名乗り表現（「○○商事の□□です。」等）にマッチする場合はその文のスコアを上げる．

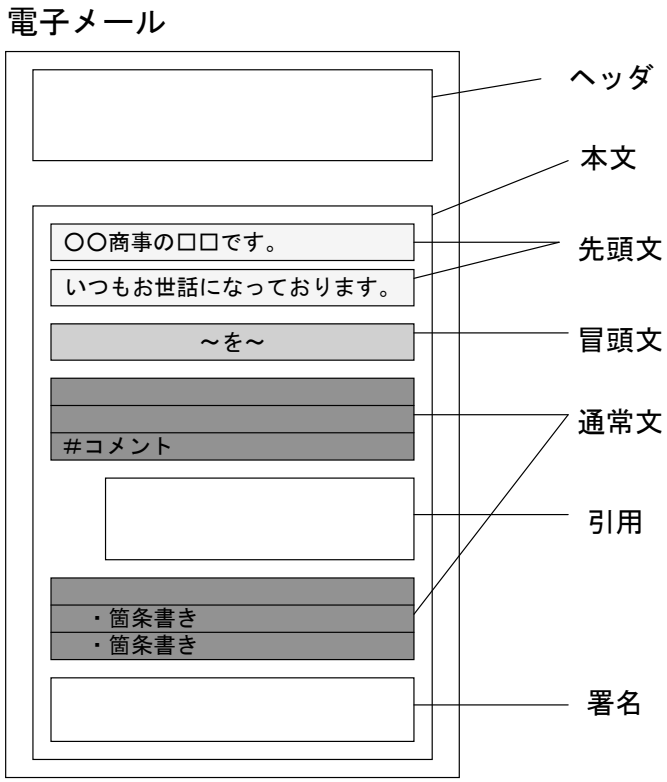

図 3.2: 重要文抽出ルールの適用時に割り当てられる文の位置情報

冒頭文ルール 先頭文ルールがマッチした次の文のみに対して，格助詞「を」格または「が」格を含む文があれば適用され，話題を含んでいるとしてその文のスコアを上げる．

通常文ルール 先頭文ルールと冒頭文ルールのどちらも適用されないすべての文に対して適用され，あらかじめ指定した重要表現の形態素パターンにマッチした場合にその文のスコアを上げる．

以上のルールは入れ子構造が許されている．1 つのルールの中には，入れ子として記述されるすべてのルールを満たさなければならないものと，どれかを満たせばよいものの指定が可能である．また，入れ子として例外を記述するルールも用意されている．例えば，逆説の接続助詞「が」を含む文を重要だとしてスコアを付与するルールを定義する際に，例えば「申し訳ありませんが」にマッチした文を重要としたくない場合には，これを例外として入れ子に記述することができる．

次に，表現や位置情報を用いた詳細なマッチングを実現するための形態素パターンについて述べる．形態素パターンの記述は，形態素の満たすべき位置の制約と形態素自身を指定するための情報という 2 つの軸がある．形態素の位置の制約には，形態素が以降の形態素と連続しなければならないもの，形態素が以降の形態素とは連続しないもの，形態素が文頭になければならないもの，形態素が文末になければならないものの 4 種類が記述できる．形態素自身は，表記，読み，品詞，一般名詞カテゴリ[1]が一致するかどうかにより指定される．これらの指定情報は単独でも組み合わせても用いることができる．例えば，表記が「が」でかつ品詞が格助詞であるという具合に指定できる．また，表記と読みは部分一致も指定できる．例えば，読みを「*ガツ」と指定すれば「1 月」から「12 月」まですべてを効率よく指定することができる．

ルールは形態素パターンとそれがマッチした文に加算あるいは減算するスコアからなる．ひとつの文に対して各々の形態素パターンがマッチするたびに指定さ

[1] 日本語語彙体系 [20] に基づく意味属性である．

れたスコアが加算あるいは減算される．どのような表現がどのくらい重要と見なすかは人によってそれぞれ異なるため，一意に決定することは極めて難しい問題である．このため，簡潔でわかりやすいルールの記述は，チューニングを容易にしておく上で不可欠である．以下では，ルールを作成する上で，具体的にどのような表現をどのくらい重要とみなすかについて述べる．なお，本研究では，ルールに与えられるスコアの大きさは，経験的に決定している．

本研究では，コミュニケーションにおいて重要と考えられる表現，すなわち依頼や要求，質問，名乗り，スケジュールの通知表現に焦点を当て，これらを重要とみなすルールを用意しておく．ここでは，ルールの一例として，名乗り表現を表すパターンとそのスコアを図 3.3 にあげる．最初のタグ<SENTOU>は先頭文パターンを表し，<RULEG nanori>という形態素パターンにマッチすればスコアが 35 だけ付与されることを記述している．<RULEG nanori>は<RULE daresoredesu>と<RULE at_mark>のどちらかがマッチすれば全体がマッチしたと見なされる．これらの中で用いられている HYOU や ICAT は表記と一般名詞カテゴリ番号を表し，NEXT=NEAR と NEXT=FAR はそれぞれ次の形態素が隣接か否かを表している．

表記は，論理和の形で記述されている．ちなみに一般名詞カテゴリ番号 363 は「機関」を 5 は「人間」を表している．例えば,「○○商事の□□です。」のような文に対して,「□□」の一般名詞カテゴリが「機関」や「人間」であれば，このルールはマッチし，名乗りを表現する文という位置付けでスコア 35 が与えられる．スコアは数値自身に意味はなく，相対的な重要度を表現する．我々の作成したルールの中では，35 というスコアは，ほとんどの場合，文のランキングにおいて上位に残る程度の重要性があることを意味する．

3.1.3　スコアに基づく文選択

指定される要約率や制限文字数の範囲内で，電子メール本文から重要文を抽出する方法について述べる．基本的には，１つまたは複数のルールによってスコア

```
<SENTOU>
RULEG=nanori VAL=35
</SENTOU>

<RULEG nanori>
RULE=daresoredesu
RULE=at_mark
</RULEG>

<RULE daresoredesu>
ICAT=[363|5] NEXT=NEAR
HYOU=[です|と] NEXT=NEAR
HYOU=[。|申|もう]
</RULE>

<RULE at_mark>
ICAT=[363|5] NEXT=NEAR
HYOU=@ NEXT=FAR
HYOU=[です|申|もう]
</RULE>
```

図 3.3: 形態素パターンとスコアから構成されるルールの例

が付与された各文に対して，スコアの大きい順にソートする．スコアの正規化およびソートの方法には様々なバリエーションがあるが，本研究では経験的に，1) 文の文字数により正規化されたスコアの総和，2) スコアの総和，3) スコアの最大値，の順でソートを行う．スコアの総和は複数のルールが付与したスコアの総和であり，スコアの最大値はその中で最も大きいスコアである．

要約率が指定される場合には，メールテキストの大きさと要約率から制限文字数を計算する．スコアに基づいてソートされた各文の順序に従って，指定された制限文字数を超えるまで，文を選択していく．このとき，引用行等も含めて出現順に付けられた文番号に基づいて選択された文同士が連続しない場合には，文が脱落していることを明示するために記号「...」を挿入する．これは特に箇条書きの項目がとびとびに選択される場合などにおいて，内容についての誤解を与えないようにするのに役立つ．これらの記号を含めて選択した文が制限文字数を超える場合は，記号とその文の選択を断念し，文の選択を終了する．なお，要約を 1 行 1 文の形式で出力するために，各文と記号は改行コードを付けて出力し，それぞれの文末の改行コードの文字数も制限文字数の考慮に入れている．

3.2　携帯電話向け要約システムへの適用

本研究で提案した重要文抽出手法を携帯電話向け要約システムへ適用する．本手法を適用した重要文抽出エンジンを搭載した SummaryBIFF システムの構成を図 3.4 に示す．インターネットサービスプロバイダ（ISP）で受信したユーザの電子メールを SummaryBIFF システムへ転送することにより，電子メールの要約（重要文）に原文へのリンク（全文参照 URL）が付与された新たな電子メールが作成される．ユーザ ID と携帯電話のメールアドレスが格納されている DB サーバとメールサーバを経由してユーザの携帯電話に即時に通知される（Push）．ユーザは重要文に目を通してそれが重要な電子メールであると判断すれば，暗号化され

たユーザ ID を含む全文参照 URL から Web サーバとユーザ ID が格納されている DB サーバを経由することにより，その場で原文へアクセスできる（Pull）．このように，SummaryBIFF システムでは電子メールを受信するとすぐに要約による着信通知が届き，ユーザ自身が要約に基づいてすべての電子メールを取捨選択できるため，大量に電子メールを受信するユーザでも重要な電子メールだけに効率よくアクセスすることができる．

携帯電話の送信可能文字数が 500byte と仮定すると，全文参照リンクの文字数を考慮すれば，重要文には 360byte から 390byte 程度の文字数しか割くことができない．このため，制限文字数を以上のように設定して重要文抽出エンジンを駆動することが妥当である．以下では，携帯電話向け要約システムを前提とした本手法の評価および考察を行う．

3.3　実験

本研究で提案した重要文抽出手法の評価を行った．以下では，実験の設定，人手による正解との比較，着信通知タスクにおける評価について述べる．

3.3.1　実験の設定

重要文抽出のためのルールは，先頭文ルールが 3 個，冒頭文ルールが 2 個，通常文ルールが 18 個である．通常文ルールの内訳は，設定されているスコアの高い順に，スケジュール通知表現（2 個），名乗り表現[2]，依頼表現（3 個），意思を表す表現，疑問表現，期限を表す表現，話題表現，数値表現，主張を表す表現である．その他に，これらのルールにマッチしない場合でも順位が付けられるように，助詞「を」を含む表現，助詞「は」を含む表現，表記自身が重要さを表す文字列，

[2]名乗り表現は通常先頭文ルールだけで十分であるが，電子メールが転送部分を含むことも考慮して転送元を抽出するために設定している．

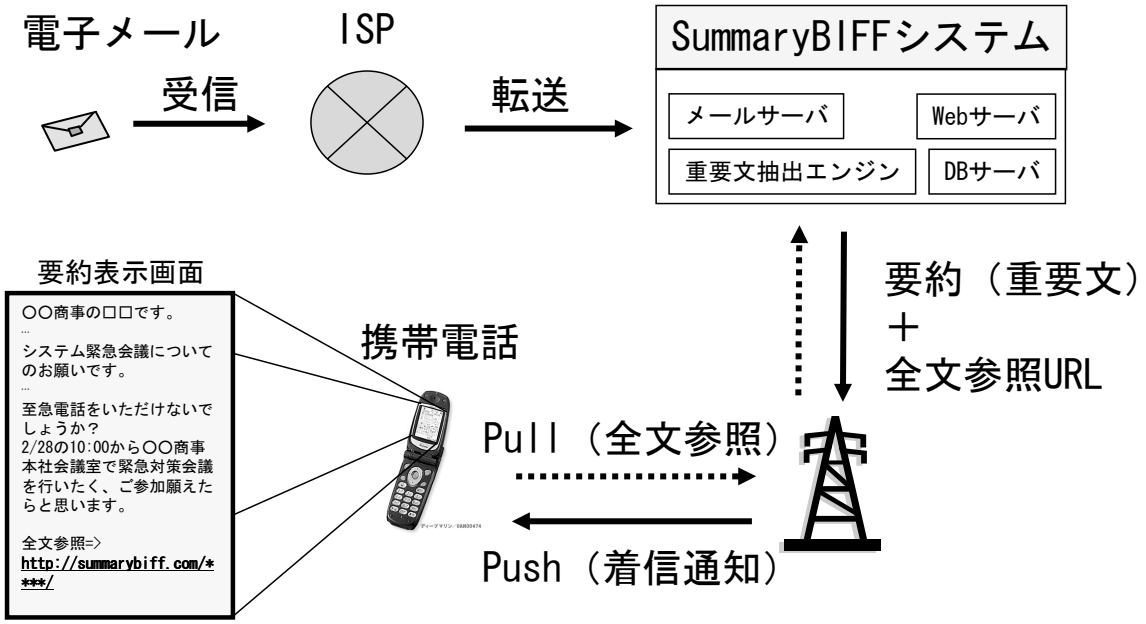

図 3.4: 携帯電話向けメール要約システムの構成

否定表現，接続助詞を含む表現，断定表現を表すルールも用意した．ルールは図 3.3 全体の単位を 1 つと数えている．

評価者 3 名により評価者自身が受信した電子メール 139 通をテストセットとして準備した．テストセットは，メールニュースなどは含まず，特定の個人や集団を相手とした目的のはっきりした電子メールである．携帯電話向け要約システムを想定して，短すぎて要約する意味がないメッセージをテストセットから除外しておく必要がある．我々は評価者による重要文の選択を単純にすることも考慮して，メッセージから重要な 5 文を選択するというタスクを設定したことにより，テストセットは 6 文以上のメッセージとした．このため，本手法では対象外としている署名や引用行や英数字のみの行を除外して，重要文抽出の評価のために文の単位をあらかじめ揃えておく必要から，システムによる文分割を実行した．この結果，無視できない誤分割[3]が存在した 22 通を除外し，117 通をあらためてテストセットとした．つまり，テストセットにおける文分割の精度は 84%であった．1 文あたりの平均バイト数は 52.8byte[4] (117 通）であったので，メッセージから 5 文の選択は平均で 264byte となる．これは，携帯電話向け要約システムを想定したときの制限文字数を超えない適度な長さであると考えられる．

3.3.2　人手による正解との比較

はじめに，テストセットに対する重要文抽出の難易度を調べるために，テストセットがどのような長さ（サイズ）のメッセージから構成されているかを，バイト数を基準にして調べた．また，人手による重要文抽出のコンセンサスの度合いを見るために，評価者 3 名がそれぞれ重要だと思われる 5 文を選択することにより重要文 5 文の正解を作成し，5 文のうち何文が一致するのかについて，その重な

[3]例えば，文の中に箇条書きが割り込むような形で存在するような場合等が該当する．

[4]テストセットの中にはレイアウトのための空白文字が含まれるが，どの単位を文とするかにより不要な文間の空白文字となったり必要な文内の空白文字となったりするため，区別が難しい．そのため，単純に空白文字のバイト数は数えていない．

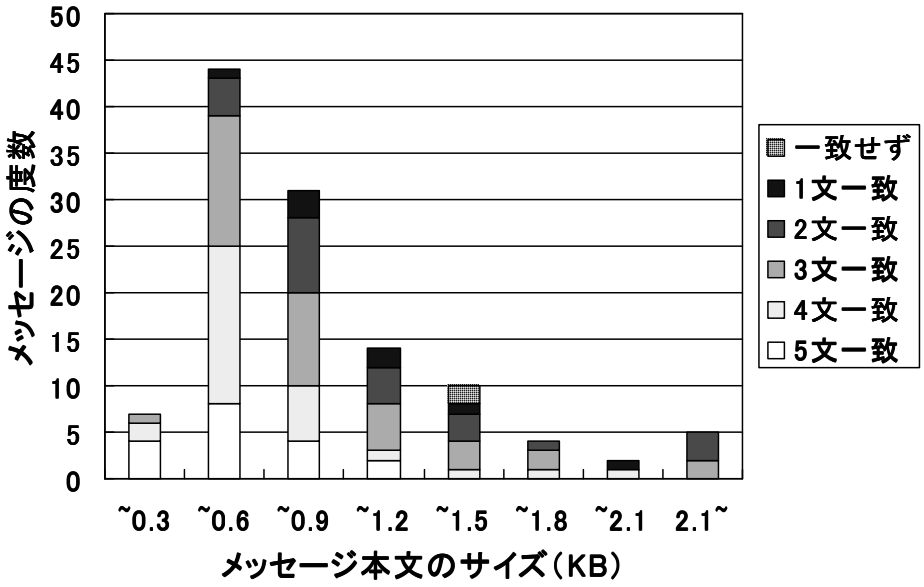

図 3.5: メッセージの長さの違いにおける複数の人手による選択文の重なり度合の分布

りの度合いでテストセットを分類した．3名による選択文の重なりの度合いがメッセージの長さの違いによってどう変化するのかを見るため，図 3.5 にメッセージのバイト数と3名の選択文の重なりの度合いの関係を示す．例えば，表中の「5文一致」とは，評価者3名の選択した5文がすべて一致することを表し,「1文一致」とは評価者3名の選択したそれぞれ5文の中で一致する文が1文であることを表す．テストセットは 1.5KB 以内のメッセージが多数を占め，コンセンサスが取れるメッセージはバイト数が小さいメッセージに偏り，コンセンサスが取れないメッセージはバイト数が大きいメッセージに分布していることがわかる．これは要約率の大小と関係するためで，メッセージの長さが長くなれば要約率は小さくなり，メッセージの長さが短くなれば要約率は大きくなるからであると考えられる．なお，評価者3名の選択文が1文も一致せず，まったくコンセンサスが取れないデータは2通であった．

次に，本手法により出力される重要文の評価を行った．本手法による重要度順

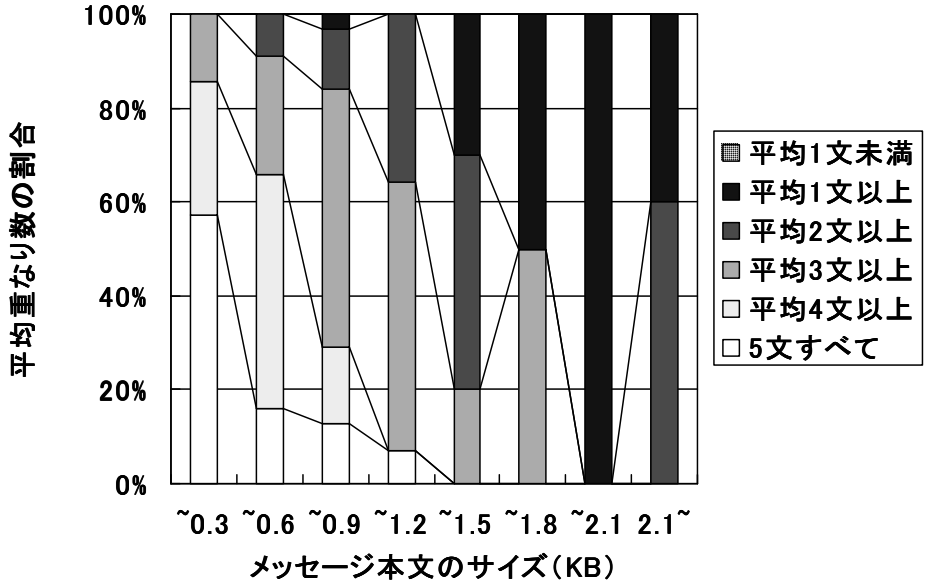

図 3.6: 本手法による選択文5文と人手による選択文5文の平均重なり数の分布

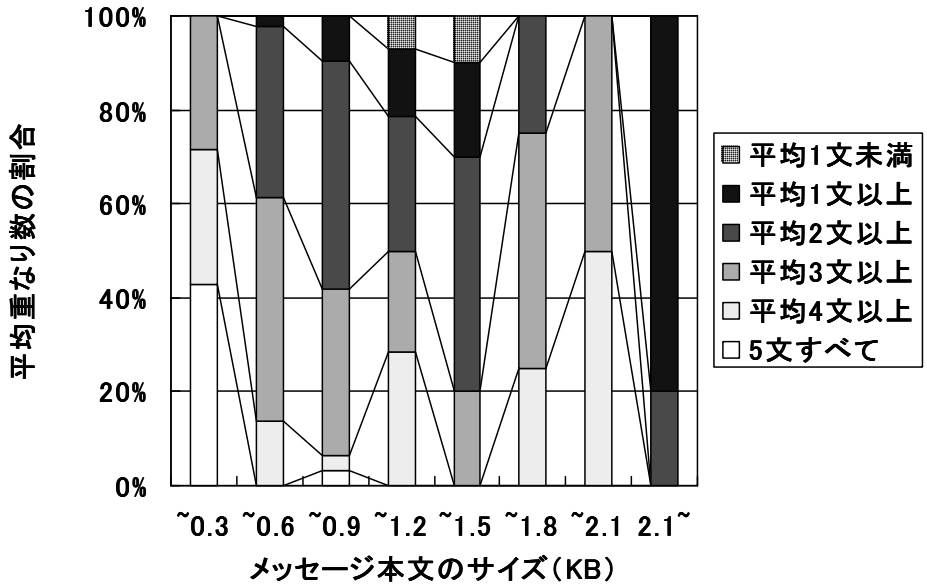

図 3.7: リード手法による5文と人手による選択文5文の平均重なり数の分布

上位 5 文と各評価者の選択した 5 文の重なりを平均し，平均重なり数の割合の分布をメッセージの大きさごとに図 3.6 に示す．表中の「5 文すべて」は評価者 3 名のいずれもが選択した 5 文とシステムが抽出した 5 文が完全に一致することを表し，「平均 4 文以上」は評価者 3 名の選択したそれぞれ 5 文とシステムが抽出した 5 文の重なりが平均で 4 文以上 5 文未満であることを示す.「平均 3 文以上」,「平均 2 文以上」,「平均 1 文以上」についても同様で,「平均 1 文未満」は評価者 3 名のそれぞれが選択した 5 文とシステムが抽出した 5 文の重なりが平均で 1 文未満であることを示す．ベースラインは文の数を揃えるために先頭から 5 文を選択したリード手法とする．ベースラインについても同様に平均重なり数の割合の分布を図 3.7 に示す．これらより，テストセットの多数を占める，メッセージの大きさが 1.5kb 以内までは，本手法の方がベースラインを大きく上回ることがわかる．特に，図 3.5 より，117 通のうち 5 文すべてが一致するのは 1.2KB 以内に集中する 18 通 (15%) であったが，本手法ではこのうちの 16 通 (89%) が正解に完全一致するが，ベースライン手法では，4 通 (22%) しか完全一致しない．これは，コンセンサスが得られたメッセージ，つまり，誰もがこのデータに対してはこのように要約されると良いと考えるメッセージに対しては，リード手法では不十分であり，本手法は非常に高い精度で重要文を抽出することができる．逆に，1.5kb を超えるメッセージは母数が少ないものの，ベースラインの方が良い．これは，評価者は先頭から多く重要文を選択したことを意味しており，長いメッセージになると先頭に全体の内容を表す要約が存在するためと考えられる．このことから，極度に長いメッセージを除けば，本手法は人手による選択文をより多く含む適切な重要文を抽出することができると言える．

3.3.3　着信通知タスクにおける評価

本手法を携帯電話向け要約システムに適用した場合のタスクに基づく評価を行うために，要約による着信通知として評価を行うときの観点を整理する．要約さ

表 3.1: 着信通知タスクにおける重要文による電子メールの取捨選択の精度

	評価者 A	評価者 B	評価者 C	合計
正解データ数	83	81	71	235
抽出データ数	75	80	84	239
正解抽出数	74	72	71	217
再現率	89%	89%	100%	92%
適合率	99%	90%	85%	91%

れたテキストを見て，本当に重要なメールが着信したことがどれくらいわかるのかがポイントだと考えられる．そこで，テストセットの 117 通を対象に評価者 3 名が本手法により電子メールから抽出された重要文と原文を別々に読んで重要かそうでないかをそれぞれ判定し，評価者ごとに本当に重要だと判断される原文に対する再現率と適合率を求めた．ただし，実際のところは送信者の情報などは重要であるが，重要文自身を適切に評価するために，重要文抽出に用いていないヘッダおよび引用部と署名は原文には含まれていない．重要かどうかの判断は，着信通知タスクを考慮しすぐに電子メールを読んでおく必要があるかどうかを基準とした．原文を読んで重要と判断した電子メールを正解データとし，重要文を読んで重要と判断した電子メールを抽出データとした．抽出データと正解データの重なりを正解抽出数，正解データ数に対する正解抽出数の割合を再現率，抽出データ数に対する正解抽出数の割合を適合率と定義した．評価者ごとの再現率と適合率を表 3.1 に示す．

原文を読んで重要と判断した正解データが 3 名の合計で 235 通なので，全体（評価者 3 名，117 通）に占める割合は，67%である．3 名の合計で，正解データ 235 通のうち重要文を読んでそれが重要だと判断できた正解抽出数は 217 通だったので，92%の再現率で各ユーザにとって重要な電子メールを網羅できる．一方，3 名の合計で，重要文を読んで重要と判断した抽出データ 239 通のうち，原文を読ん

で重要だと判断できた正解抽出数は 217 通だったので，91%の適合率で重要文からその電子メールが各ユーザにとって重要だと推定できる．いずれの評価者についても再現率適合率ともに高い値を示しているので，重要文だけを見て電子メールの重要性が判断でき，要約による着信通知タスクにおいては電子メールの取捨選択に十分使えるということがいえる．これは，SummaryBIFF システムを利用すれば，重要な電子メールにだけ原文にアクセスすればよいので，効率的な電子メールアクセスが実現可能となることを意味する．

3.4　考察

3.4.1　ユーザへのアンケートによる評価

ここまでに示した評価は，小規模な主観評価によるものであるため，提案手法の有効性や，実装した重要文抽出ルールの実フィールドでの網羅性を直接実証するものとは言いがたい．しかしながら，メールという個人性の高い情報を対象とすることから，大規模な実験を行うことも困難である．そこで，SummaryBIFF システム[5]を 19 名のユーザに 3ヶ月利用してもらい，その使用感・有効性をアンケートにより分析した．本システムのユーザは研究開発業務に携わるビジネスマンであり，要約の対象とする電子メールはユーザが業務において日々受信している電子メールである．電子メールの要約の満足度を図 3.8 に，重要文の抽出についての評価を図 3.9 に，重要でない文の誤抽出についての評価を図 3.10 にそれぞれ示す．

これらの結果より，要約の品質に対して不満を持つユーザは 21%にとどまり，大変満足と満足を合わせ 58%のユーザは要約結果に満足し，特に不満を持たないユーザを含めると 79%に上ることが分かった．また，重要文抽出の精度についても「あ

[5]実際に運用したシステムでは，セキュリティ上の制約から，携帯端末に要約を転送するのではなく，携帯端末からシステムへのログイン後にまず要約を提示し，ユーザは必要に応じて全文を参照する方式を取った．

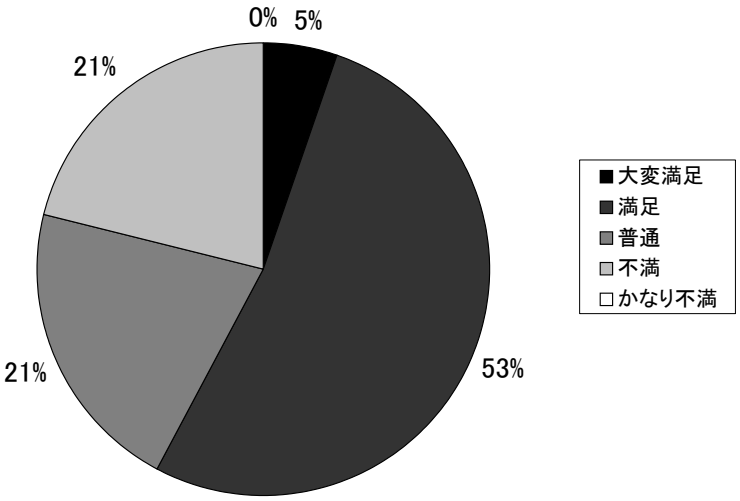

図 3.8: 電子メールに対する要約の満足度についてのアンケート結果

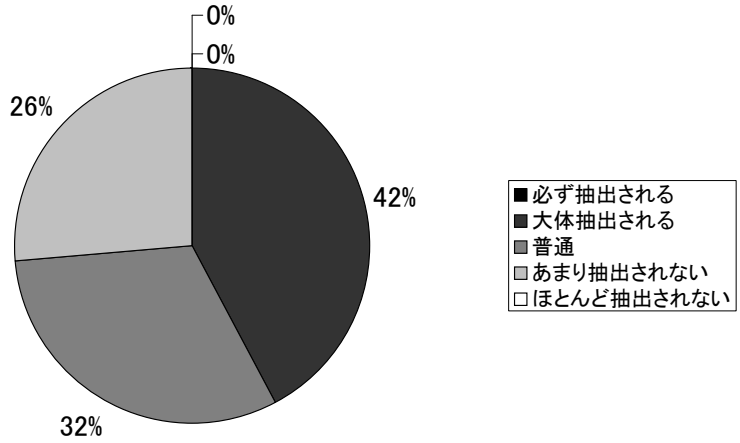

図 3.9: 重要文の抽出精度についてのアンケート結果

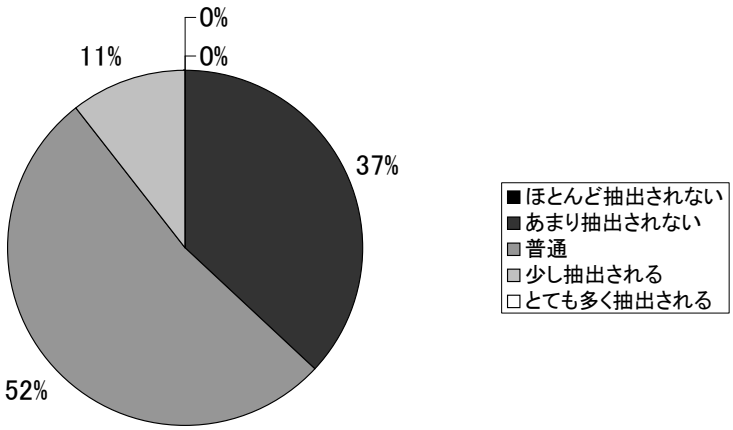

図 3.10: 非重要文の非抽出精度についてのアンケート結果

まり抽出されない」と回答した 26%を除く 74%のユーザが重要文の抽出に対して肯定的な印象を持っていたことがわかる．反対に，重要でない文を誤って抽出するかどうかについては「少し抽出される」と回答した 11%を除く 89%が重要でない文は抽出しないという肯定的な印象を持っていたことがわかる．これらより，電子メールに特有なスタイルと表現の特徴を利用した高々23 個のルールによる重要文抽出であっても，本システムはほぼ実用レベルに達しているものと考えられる．

3.4.2　重要部分の選択の単位

本節では，着信通知というタスクにおける重要文抽出の単位について検討する．本研究では抽出の単位を文としたが，文への分割を行うことなく，パラグラフ (空行で区切られる形式段落) を抽出の単位とすることも可能である．仮に，抽出単位をパラグラフとした場合は，文分割の必要がないため原文のつながりがそのまま維持され可読性に優れる反面，制限文字数が限られている場合には他の重要な部

表 3.2: 重要文を含むパラグラフの長さの平均値と重要文の占める割合

	すべて	重要文を含む	複数文に限定
個数	947	443	258(58%)
バイト数 (byte)	105.5	120.9	159.3
重要文の割合 (%)	–	73.1	64.9

分を落としてしまうという欠点が存在する．一方，抽出単位を文にした場合は，文分割時の誤りや脱落文のために文のつながりが失われ可読性が犠牲になるが，短い制限文字数の中においても重要な箇所は網羅できるという利点が存在する．我々は，以上を検証するために，テストセットについてパラグラフに関する調査を行なった．

バイト数を基準にして，各評価者の選択した重要文を含むパラグラフの長さの平均値とその重要文のパラグラフ当たりに占める割合を調査した結果を表 3.2 に示す．比較のため，重要文を含まないものも含めてすべてのパラグラフの個数と長さの平均値についても示した．重要文を含むパラグラフ当たりに占める重要文のバイト数の割合は 73.1%であった．しかし，この中には文を単位とした場合と共通となる，1 文だけで構成されるパラグラフが含まれている．1 文だけのパラグラフを比較対象から除外すると，重要文を含むパラグラフのうち 58%のパラグラフが複数の文から構成されることがわかり，重要文の占める割合は 64.9% であった．これは，抽出の単位をパラグラフとした場合には，重要でない部分が 35%程度混入することを意味する．さらに，実際のアプリケーションでは，単純に空行で分割可能なパラグラフを単位とするので，本研究では対象としなかった記号のみの行や英数字のみの行のバイト数も加算されるため，パラグラフ内の非重要な部分はさらに多くなるであろう．

また，表 3.2 に示すように，すべてのパラグラフの長さの平均値が 105.5byte であるのに対して，重要文を含むパラグラフの長さの平均値は 120.9byte であった．

一方，すべての文の長さの平均値が 52.8byte であったのに対して，重要文の長さの平均値は 63.9byte であったことから，その分だけ重要文を含むパラグラフは長くなったと考えられる．特に，重要文を含む複数文からなるパラグラフだけに限れば，その長さは 159.3byte となり，重要文の長さである 63.9byte の約 2.5 倍になる．これは，パラグラフを抽出単位とした場合，少ない制限文字数の中では選択できるパラグラフの個数が少なくなることを意味する．以上より，パラグラフ内の非重要な部分の割合とパラグラフ全体の長さという 2 つの面から，携帯電話向け要約システムへの適用時には，抽出単位を文単位として重要箇所の網羅性を上げる方が有利であると考えている．

第4章　Webページの要約

4.1　Webページからの概要文生成

本章では，携帯端末向けに Web ページを要約する方法について述べる．携帯端末向けの Web ページの概要文は，Web ページに対する query-biased な要約であり，重要文抽出だけでなく文短縮までをも必要とする非常に短い要約を Web ページから生成することが課題となる．第 2 章で述べた Web ページの要約に関する従来研究は，この課題を十分に解決しているとは言えない．そこで，非常に限られた要約長において可読性と内容網羅性を両立させた Web ページからの概要文生成の方法を提案する．はじめに要約のアプローチについて全体の流れを述べ，その後に個々の技術について詳細を説明する．

Web ページを要約するときの基本的な考え方は，検索質問を考慮した重要文を抽出するために検索質問に対する各文の適合度を計算しておき，目的の要約長に達するまで最も適合度の高い文から順に抽出していくものとする，ただし，画面の小さな携帯端末での表示に適応するために，目的の要約長が短い，つまり，要約率が低いことを考慮して，次のような改良を加える．(a) 検索結果は Web ページのタイトルとスニペットが同時にユーザに提示されるので，内容の重複を避けるために Web ページのタイトルと類似している文は抽出しないようにする．(b) 文を単位とする抽出では単位あたりの長さが長くなるために目的の要約長を有効に使えない場合もあるので，目的の要約長を最大限利用するために必要に応じて文を短縮する．具体的には，選択しようとしている文を新規にあるいはすでに選択済みの文に追加することによって目的の要約長を超える場合には，その文を追

加しても目的の要約長の範囲内に収まるように短縮する．つまり，目的の要約長からすでに選択している文の長さを引いた長さが，選択しようとしている文に許される最長の長さになり，この長さの範囲でその文を短縮する．例えば，1 番目に重要な文の長さが目的の要約長を越えるならば，その文は目的の要約長の範囲で短縮する．1 番目に重要な文の長さが目的の要約長よりも短い場合は，1 番目に重要な文をそのまま追加した後に 2 番目に重要な文を追加しようとするが，2 番目の文を追加することで目的の要約長を越える場合は，目的の要約長から 1 番目の文の長さを引いた長さの範囲において 2 番目の文を短縮する．抽出した文および短縮した文は出現順序に従って並び替え，これらの間で原文から落ちた部分を記号「...」で置換する．抽出した文および短縮した文の長さの合計が目的の要約長のある一定の割合を満たせば，重要文抽出と文短縮を終了する．このように重要文抽出と文短縮を動的に組み合わせるアプローチを取る．

以下に要約の流れを示す．

1. Web ページからの HTML タグの除去・本文抽出
2. 本文に対する文分割
3. 検索質問に対する本文中の単語の重要度の計算
4. Web ページのタイトルとの冗長性排除による重要文抽出
5. 検索質問を考慮した文短縮

Web ページの HTML タグの除去は，単純に HTML タグを削除するのみである．Web ページから HTML タグを除去したテキストを本文とする[1](1)．本文に対する文分割は句点と飾り記号および改行で区切られている箇所を特定し文を切り出す (2)．検索質問に対する単語重要度の計算 (3) と冗長性排除による重要文抽出 (4)，検索質問を考慮した文短縮 (5) については以下でそれぞれ詳述する．

[1]Web ページから HTML タグを除去するだけなので，広告等を含む Web ページでは広告等がそのまま残るが，広告等を含めて本文として扱う．

4.1.1　検索質問との相関に基づく単語重要度の計算

検索質問を考慮した重要文を抽出するために，検索質問を考慮した単語重要度を計算する．検索質問を構成する単語はもちろん重要であるが，検索質問はたかだか数単語なのでそれ以外にも検索質問と関連性の高い単語も重要であると考えている．我々は検索質問と関連性の高い単語とは検索質問と共起しやすい単語であると考え，検索質問と Web ページに含まれる任意の単語との共起のしやすさを単語重要度に取り入れることを提案する．本研究では，[13] で提案された単語のバイグラムについての仮説検定の手法を流用し，検索質問 q と文書内の単語 t の共起のしやすさ（相関の高さ）を計算する．まず，何らかのコーパスを用いて，コーパスの全文書集合において単語 t を含む文書の出現確率 p と，検索質問 q を含む文書集合における単語 t を含む文書の出現確率 $p1$，および，検索質問 q を含まない文書集合における単語 t を含む文書の出現確率 $p2$ を計算する．もし検索質問 q と単語 t が独立（無相関）であれば，出現確率 $p1$ と $p2$ は等しくなり，p とも等しくなるはずである．逆に検索質問 q と単語 t が独立でない場合は，出現確率 p と $p1$，$p2$ は等しくならないはずである．この仮定を利用し，検索質問 q と単語 t の相関を調べるために，検索質問 q と単語 t が独立に出現するとした仮説と，検索質問 q と単語 t が独立でなく出現するとした仮説との尤度比を計算する．それぞれの仮説の尤度は 2 項分布を用いて計算する．計算式を式 4.1 に示す．式 4.1 の分子は検索質問 q と単語 t が独立とする仮説の尤度，分母は検索質問 q と単語 t が独立でないとする仮説の尤度である．検索質問 q と単語 t の共起のしやすさ（相関の高さ）として，これらの仮説の尤度比 (λ) による対数尤度比 $(-2log\lambda)$ を採用する．検索質問 q と単語 t が独立でない，つまり相関が高い場合に，尤度比 (λ) は小さくなり，対数尤度比 $(-2log\lambda)$ が大きくなる．

$$\lambda = \frac{L(c_{12}, c_1, p)L(c_2 - c_{12}, N - c_1, p)}{L(c_{12}, c_1, p_1)L(c_2 - c_{12}, N - c_1, p_2)} \tag{4.1}$$

ただし，

$$p = \frac{c_2}{N} \tag{4.2}$$

$$p_1 = \frac{c_{12}}{c_1} \tag{4.3}$$

$$p_2 = \frac{c_2 - c_{12}}{N - c_1} \tag{4.4}$$

$$L(k, n, x) = x^k (1 - x)^{n-k} \tag{4.5}$$

とする．ここで N は全文書数である．c_1 は検索質問の文書頻度，c_2 は単語の文書頻度，c_{12} は検索質問と単語の共起頻度を表す．検索質問と同じまたは検索質問に含まれる単語についても，他の単語と同様に検索質問と共起するとみなして対数尤度比を計算する．ただし，計算の都合上共起頻度から 1 を減じておく．

単語重要度は，対数尤度比 (LLR) と文書頻度の逆数 (idf) の積とする．検索質問 q における単語 t の単語重要度を式 4.6 に定義する．

$$imp(t) = -2 \log \lambda(q, t) * \log \frac{N}{df(t)} \tag{4.6}$$

ここで，$-2log\lambda(q,t)$ は検索質問 q と単語 t の対数尤度比であり，$df(t)$ は単語 t を含む文書頻度である．

4.1.2　タイトルとの重複を排除した重要文抽出

検索質問を考慮した単語重要度に基づいて文を抽出するが，Web ページのタイトルやすでに選択した文との内容の重複を避けるために，Maximal Marginal Relevance (MMR)[8] を用いる．MMR の式 4.7 に従って重要文を抽出する．

$$\begin{aligned} MMR(Q, R, S) \;=\; & \underset{P_i \in R \backslash S}{\operatorname{argmax}} \{ \alpha Sim_1(P_i, Q) \\ & -(1 - \alpha) \max_{p_j \in S} Sim_2(P_i, P_j) \} \end{aligned} \tag{4.7}$$

ただし，Q は検索質問，R は本文のすべての文，S はすでに重要文として選択されている文，$R \backslash S$ はまだ重要文として選択されていない文とする．$Sim_1(P_i, Q)$

は検索質問 Q と文 P_i との適合度を算出する関数で，Web ページ内のメニューなどの短い文を避けて内容を多く含む長い文を抽出するために，文を構成する単語の単語重要度の総和を用い，総和の最大値で正規化した値を採用する．$Sim_2(P_i, P_j)$ は S に含まれる文 P_j と抽出候補の文 P_i の間の内容の重複を避けるために計算する類似度で，コサイン類似度を用いる．

検索エンジンは Web ページのタイトルの情報も利用して Web ページのランキングを行っていることが想定され，Web ページのタイトルに検索質問を含む Web ページが上位にランキングされることが予想される．このため，Web ページのタイトルを 0 番目の文として考え，はじめから S に入っている，つまりはじめからすでに選択された文として扱うことで，Web ページのタイトルとの冗長性を排除しながら，検索質問に適合する文を抽出する．

4.1.3　検索質問を考慮した文短縮

文を抽出していく際に目的の要約長の範囲に収まらない場合に文を短縮する．抽出対象の文を係り受け解析して得られる木構造に基づき，目的の要約長の範囲で述部をルートとする最適な部分木を選択する．最適な部分木を選択するために，以下の 2 つの数値を組み合わせたスコアを提案する．

- 検索質問を考慮した単語重要度に基づく文節網羅率
- 文節同士のつながりやすさである文節連接確率

一方で，文節網羅率を考慮することにより，文の内容を網羅するよう重要な単語を含む文節を多く含む部分木のスコアが高くなり，他方で，文節連接確率を考慮することにより，文としての読みやすさを向上するよう各文節同士のつながりやすさが大きい部分木のスコアが高くなる．計算されたスコアが最も大きい部分木を短縮文とする．なお，文節連接確率では文節が文頭や文末になる確率も計算する．

文節網羅率は式 4.8 に定義する．

$$P_{imp}(w_i) = \frac{\sum_{t_k \in w_i} imp(t_k)}{\sum_{w_j \in X} \sum_{t_l \in w_j} imp(t_l)} \tag{4.8}$$

ただし，X は文全体，w_i と w_j は文節を表し，t_k は文節 w_i に含まれる単語を，t_l は文節 w_j に含まれる単語を表す．

文節連接確率は式 4.9 に定義する．

$$P_{adj}(w_i|w_{i-1}) = P(f_i|f_{i-1}) \tag{4.9}$$

ただし，f_i は文節の内容語の主辞を表し，f_{i-1} は文節の機能語の主辞を表す．文節の内容語の主辞とは，文節に存在する内容語列のうち最後尾のものを指す．これは，日本語では文節の代表単語はより後方に存在するためである．文節の機能語の主辞とは，文節に存在する機能語列のうち句読点等の記号類を除く最後尾のものを指す．これは，日本語では文節がどの文節と係り受け関係になるかを表す重要な指標となるためである．ただし，副詞文節のように機能語がないものは内容語の主辞と同じ単語になる．式 4.9 は文短縮した結果の短縮文において，各文節同士が隣り合う尤もらしさを表し，$P(f_i|f_{i-1})$ はコーパスから最尤推定により求める．ある文節の機能語の主辞はその文節の最後尾となるので，次に並ぶ文節の内容語の主辞との連接確率を計算する．これは，係り受け解析で得られる部分木のうち，どの深さまでのノード（文節）を選択するかを決定するのに寄与する．また，係り受け解析に誤りがある場合でも，文節の連接を考慮することで，その誤りを回避することも期待される[2]．

[2]例えば,「いただいた/お薬は、/弱い/種類であるものの/ステロイド系なので、/生まれて/間もない/赤ちゃんに/使う/ことを/迷っています。」という文（文節の境を”/”とする）を例に取ると，仮に「生まれて」が「使う」に係ると誤って解析された場合には,「使う」には他に「間もない/赤ちゃんに」が係るとすると,「間もない」だけが落ちて,「生まれて/赤ちゃんに/使う」という望ましくない部分木を含む非文が生成される可能性がある．係り受け解析に部分的に誤りがあったとしても，文節連接確率を用いることにより,「生まれて」と「赤ちゃん」の連接確率が小さいのであれば，このような非文が生成される確率も小さくなる．

文節網羅率と文節連接確率に基づいて各部分木のスコアを式 4.10 に定義し，このスコアが最大となる述部をルートとした部分木を要約対象である文の短縮文とする．

$$W^* = \underset{W \in G(X)}{\mathrm{argmax}} \{ \log \sum_{i=0}^{node(W)-1} P_{imp}(w_i) + \frac{\sum_{i=0}^{node(W)} \log P_{adj}(w_i|w_{i-1})}{node(W)} \} \qquad (4.10)$$

ここで，X は要約対象の文を，$G(X)$ は文 X における部分木の集合を，W は部分木を，$node(W)$ は部分木 W を構成する文節数を表す．第 2 項の文節連接確率は文節の少ない部分木のスコアが大きくなるため，部分木の文節数で幾何平均を取る．文節網羅率と文節連接確率の重みは，両者が同等であることに価値があると考え 1:1 とした．

4.2　実験

4.2.1　実験の設定

提案した Web ページの概要文を生成する手法について評価を行うために，実験を行った．検索質問は NTCIR-4 Web Info1 情報指向検索タスク[3]のサーベイ型クエリ 35 個のうち 20 個を用いた．Web ページは検索エンジン，モバイル goo[4]を用いて各検索質問で検索され重複を除いた上位 10 件（検索質問のひとつのみ 9 件）の携帯端末向けページを収集した．各 Web ページから HTML タグを除去した後の全 199 ページのテキストの平均の長さは 2736 バイト，標本分散は 5823899 であった．概要文の生成の条件として目的の要約長を 70 バイトとした．要約率に換算すると平均 2.56%となり極めて低い要約率となる．

比較対象として KWIC (keyword-in-context) に基づくベースラインを採用した[5]．KWIC は，基本的には検索質問で与えられるキーワードとその前後の文字列を抜

[3]http://research.nii.ac.jp/ntcir/permission/ntcir-4/perm-ja-WEB.html

[4]http://mobile.goo.ne.jp

[5]商用の検索エンジンの概要文生成は基本的には KWIC による方法と考えられるが，その具体的な方式は公開されていない．このため，商用の検索エンジンの概要文との比較により提案手法の

粋するが，複数のキーワードがマッチする場合についてはキーワードの選び方が多様になるため，次のような手順とした．まず，本文中に出現するすべてのキーワードの位置を取得し，それに基づき各キーワード，および，あるキーワードから別のキーワードまでの文字列の範囲をそれぞれパートとする．パートは，目的の要約長を超えない範囲で，3 つ以上のキーワードを含んでもよい．次に，目的の要約長を超えない 1 つ以上のパートの組み合わせを候補とし，候補が 1 つになるまで以下の優先規則を順次適用する．1) キーワードの異なり数が最も多い．2) パート数が最も少ない．ただし，パート数は 2 までとし，パート同士をつなぐ箇所には記号「...」を挿入する．3) キーワードの総数が最も多い．最後に，目的の要約長に達するように，決定した候補の各パートの前後から同じ長さの文字列を付加する．

提案手法では，単語重要度の計算に対数尤度比を用いるものと用いないもの，MMR を用いるものと用いないもの，文短縮を用いるものと用いないもので組み合わせを作り，表 4.1 に示す 4 つのバリエーションを評価した．単語重要度が LLR*idf 値の場合は，式 4.6 を用いる．対数尤度比の計算は形態素解析による品詞が名詞または未知語である単語に限定し，それ以外の単語の単語重要度は idf 値とした．idf 値の計算には，品詞が名詞または未知語である単語の文書頻度を用い，それ以外の単語には一定の小さな値を与えた．単語重要度が tf*idf 値の場合は，対数尤度比の変わりに文書内単語頻度 $tf(t)$ を用い，式 4.6 において $-2log\lambda(q,t)$ の変わりに $\log tf(t)$ で計算した．ただし，検索質問 q に含まれる単語については計算結果を二乗することで検索質問にバイアスをかけた．MMR を行う場合は式 4.7 の α の値を 0.5 に設定し，MMR を行わない場合は α の値を 1 に設定して第 2 項を無視した．文短縮を行う場合は，目的の要約長を超えて文を選択する際に式 4.10 に従って目的の要約長を超えないように文を短縮し，目的の要約長の 8 割を満たせば重

効果を測定することは難しい．そこで，KWIC に基づくベースラインを実装することにより提案手法との比較を行った．

表 4.1: 提案手法のバリエーション

	単語重要度	MMR	文短縮
提案手法 (1)	tf*idf 値	行う	行わない
提案手法 (2)	tf*idf 値	行わない	行う
提案手法 (3)	tf*idf 値	行う	行う
提案手法 (4)	LLR*idf 値	行う	行う

要文抽出と文短縮を終了させた．文短縮を行わない場合は，要約長を超えて文を選択する際に文短縮の代わりに，抽出する文の文頭から目的の要約長までの残りのバイト数分だけカットした．提案手法で利用したデータを表 4.2 に示す．単語重要度を計算するための各単語についての文書頻度および検索質問との共起頻度は表 4.2 のすべてを用いた．また，文短縮に用いる文節連接確率を求めるための文節連接モデルを学習する際には，毎日新聞と Yahoo!知恵袋データに対して係り受け解析器 [21] を実行し，得られた文節の内容語と機能語のそれぞれの主辞の品詞の連鎖からモデルを学習した．ただし，機能語については表記も利用した．

評価方法は，被験者 3 名による主観評価とした．評価基準については，Web ページの概要文として有効かという検索タスクの評価と，テキスト要約としての観点から要約の評価を行った．検索タスクの評価では，NTCIR-4 Web Info1 情報指向検索タスクの <NARR> タグに記載されている判定基準に基づき，部分適合を含む適合と不適合の 2 値による適合性判定を行った．要約の評価では，文法性の評価，タイトルを含めた非冗長性の評価，内容網羅性の評価を各 5 段階で行った（5 が最も良く 1 が最も悪い）．文法性の評価では，単語の途中で文が途切れていないか，日付や記号を含まないかを評価した．非冗長性の評価では，Web ページのタイトルを含めて不必要な情報の繰り返しがないかを評価した．内容網羅性の評価では，概要文として不必要な情報を含んでいないか，検索質問に対して無関係の

表 4.2: 実験に用いたデータ

データ	文書数
毎日新聞 2003 年-2007 年	496,606
Wikipedia	511,120
Yahoo!知恵袋	16,593,794
ブログ	13,556,798

情報を含んでいないかを評価した[6].

4.2.2　検索のための評価

適合性判定の評価結果を表 4.3 に示す．適合性判定の評価は概要文（要約），Web ページ（原文）の順序で行った．原文の適合性判定が適合となる文書を正解として用い，再現率，適合率，F 値で評価を行った．再現率と適合率は以下のように定義する．

$$\text{再現率} = \frac{\text{概要文が適合かつ原文が適合の文書数}}{\text{原文が適合の文書数}}$$
$$\text{適合率} = \frac{\text{概要文が適合かつ原文が適合の文書数}}{\text{概要文が適合の文書数}}$$

評価対象の 199 文書を評価者 3 名により評価したところ，延べ文書数 597（$199*3 = 597$）のうち 146 文書が適合と判定された．2 名 1 組とした 3 組の kappa 値の平均値は 0.58(moderate) であった．表 4.3 から，Web ページの本文を文として扱った

[6]内容網羅性の評価については,「概要文として必要な情報とは何か」，また,「検索質問に対して関係のある情報とは何か」を規定することが難しかったため,「概要文として不必要な情報を含んでいないか」と「検索質問に対して無関係の情報を含んでいないか」という 2 つの基準を考慮して 5 段階の主観評価で評価した．これら 2 つの基準を両方とも満たしていれば良い評価が，どちらも満たさなければ悪い評価が与えられる．

表 4.3: 適合性判定の評価

	再現率	適合率	F 値
KWIC (ベースライン)	0.760	0.457	0.571
提案手法 (1)	0.582	0.531	0.556
提案手法 (2)	0.623	0.532	0.574
提案手法 (3)	0.644	0.537	0.586
提案手法 (4)	0.719	0.522	0.605

ときに，(1) を除く提案手法はベースラインの F 値を上回ることが確認できた．最も F 値が高かったのは提案手法の (4) であった．tf*idf 値と MMR による重要文抽出と文頭からのカット (1) では，再現率が低いためにベースラインの F 値を上回ることはできず，MMR を実行しないが文短縮を実行すること (2) ではじめてベースラインの F 値を上回ることがわかった．提案手法の (1) と (3) は文短縮を実行するか否かの違いがあり，これらを比べると再現率，適合率ともに向上したので，文短縮の有効性が明らかになった．また，提案手法の (2) と (3) は MMR を実行するか否かの違いがあり，これらを比べると文短縮ほどではないものの再現率，適合率ともに向上し，MMR の有効性が確認できた．提案手法の (3) と (4) では対数尤度比を用いたか否かの違いがあり，適合率は下がるもののそれ以上に再現率が上がり F 値が向上したことから，対数尤度比の効果を確認できる結果となった．

4.2.3　要約としての評価

要約としての評価を表 4.4 に示す．ベースラインの KWIC と提案手法の各バリエーションに対して，文法性，非冗長性，内容網羅性を 5 段階で評価した．文法性については，KWIC が最も評価が低く，次に文短縮を行わない提案手法の (1) の評価が低かった．文法性の評価では評価基準に単語の途中で切れているかどうか

を採用した．このため，KWIC では文境界を考慮せずに検索質問を含む周辺の文字列を抜き出すので，評価を下げたと考えられる．文短縮を行わない提案手法の (1) も文頭から指定の文字数を抜き出すために，単語の途中で切れてしまう場合が発生するので，評価を下げたと考えられる．実験では文抽出のみの文法性評価を行わなかったが，これは KWIC でも文を単位とする抜粋ではないことと，文抽出のみに限定すれば文法性は向上するかもしれないが目的の要約長を大きく割り込むことで内容網羅性が著しく低下する可能性を避けることによる．表 4.4 における文法性の括弧内の数値は文を短縮あるいは先頭から抜粋したもののみを対象とした評価結果であるが，それでも提案手法はいずれもベースラインを上回っている．非冗長性では，提案手法の (2) と (3) を比べると，(2) よりも (3) の方が非冗長性のスコアが大きいことから，MMR は非冗長性に有効であるという結果が得られた．ただし，非冗長性については文法性ほど差が出なかった．これは，実験に用いたすべての検索質問が複数のキーワードから構成されていたので，Web ページのタイトルに複数のキーワードを含む場合自体が少なく，概要文との冗長性が大きくはなかったためと考えられる．内容網羅性については，不適合文書は検索質問に適合する情報を含むことは少ないと考え，適合と判定された 146 文書のみで評価した．70 バイトという非常に限られた長さのために全体的に内容網羅性は低いが，提案手法のいずれもベースラインの KWIC を上回った．提案手法のバリエーションの比較により，文短縮や MMR，対数尤度比の利用がそれぞれ効果的であり，それらを統合した提案手法の (4) が最も高い評価であることがわかった．

4.3　考察

4.3.1　Web ページの検索結果に対する考察

実験結果について考察する．適合性判定の F 値が最も高かった提案手法の (4) とベースラインの KWIC とを比較する．まず，提案手法とベースラインにおける適

表 4.4: 要約としての評価

	文法性	非冗長性	内容網羅性
ベースライン	2.85	4.27	1.75
提案手法 (1)	3.73(3.62)	4.68	2.16
提案手法 (2)	4.09(4.02)	4.61	2.14
提案手法 (3)	4.00(3.96)	4.82	2.23
提案手法 (4)	4.11(4.08)	4.81	2.32

合性判定の差分を表 4.5 に挙げる．ベースラインでは正しく適合性判定が出来なかったページが提案手法では正しく適合性判定ができるようになった場合と，反対にベースラインでは正しく判定できていたが提案手法では誤って判定した場合について，それぞれの判定の変化の件数を示した．特に顕著なのは表 4.5 の分類 B で，原文が不適合な文書に対して適合から不適合に正しく判定できた事例が他よりも多かった．このことは提案手法の適合率が向上したことを裏付けているといえる．評価者 3 名のうち 2 名の判定結果が一致する例を分類ごとに表 4.6 に示す．ただし，分類 B と D は一部のみを示している．分類 A の例は，KWIC では 2 つのキーワードが離れていたために文脈が不明であるため誤って不適合と判定された例である．提案手法では 2 つのキーワードを含まないが 1 つのキーワードを含む文における文短縮が効果的に機能し，正しく適合すると判定された．分類 B の例では，KWIC は 3 つのキーワードをすべて含むことから誤って適合と判定された例である．提案手法では 1 つのキーワードしか含まないが述部をルートとする文短縮により文脈が明らかになり，正しく不適合と判定された．実験結果では分類 B が最も多かったが，不適合文書に対して不適合であることを示す概要文を生成するには，たとえキーワードを落としても述部をルートとする文短縮が有効であることを確認した．一方，分類 C では，検索質問が「料理，切り方，名称」の例では,「料理」の近傍に「講習会」があるため正しく不適合と判定された例であ

る．提案手法ではこのような不適合となりえる情報が落ちたために誤って適合と判定された．分類 D のひとつめの例では，KWIC では 3 つのキーワードが十分な文脈を伴って含まれているために正しく適合と判定された例である．提案手法では，KWIC の抽出対象と同じ文について文短縮を行ったが，目的の要約長の範囲で述部をルートとする部分木を採用するという制約により，キーワードのひとつである「訴訟」が落ちてしまい，Web ページのタイトルにもこのキーワードがなかったため誤って不適合と判定された．このことから，キーワードによっては落としてよいものと落としてはいけないものがあることがわかり，Web ページのタイトルの情報を含めてこれらをどのように判別するかという課題が明らかになった．

さらに，3 名の評価者全員の適合性判定が一致する文書のみを対象とした場合についての評価結果を示し，提案手法の有効性を考察する．評価尺度は，延べ文書数で行ったときと同様に，再現率と適合率とした．再現率は，原文の評価が 3 名とも適合で一致する文書数に対する，原文と概要文の評価が 3 名とも適合で一致する文書数の割合と定義した．適合率は，概要文の評価が 3 名とも適合で一致する文書数に対する，原文と概要文の評価が 3 名とも適合で一致する文書の割合と定義した． 3 名の適合性判定が一致する場合では，ベースラインが再現率 0.46・適合率 0.34 で，提案手法は再現率 0.65・適合率 0.52 であった．つまり，再現率は 1.4 倍，適合率は 1.5 倍に向上した．一方，3 名全員の原文の評価が適合で一致する文書数が 26 であったのに対し，3 名の評価者それぞれが適合と判定した原文の文書数は 44, 53, 49 であった．これは，ある評価者の原文を適合と判定した 44 文書から原文の情報とまったく過不足のない情報を含む概要文が作られたと仮定しても，別の評価者によってこのうちの 26 の概要文は原文が 3 名とも適合と判定されているので概要文も適合と判定されるが，残りの 18 の概要文は原文が 3 名とも適合と判定されていないので概要文は不適合と判定されることを意味する．これを 3 名の評価で平均すると，3 名それぞれの評価者が適合と判定した文書数に対する，3 名全員が適合の判定で一致した 26 文書の割合の平均は，$(26/44 + 26/53 + 26/49)/3 = 0.53$

表 4.5: ベースラインと提案手法における適合性判定の差分

差分	原文の判定	要約の判定の変化	件数	分類
改	適合	不適合 → 適合	13	A
善	不適合	適合 → 不適合	52	B
改	不適合	不適合 → 適合	16	C
悪	適合	適合 → 不適合	19	D

であり，この値が適合率の上限値に相当するといえる．以上により，提案手法は適合性判定において有効である．

次に，適合と判定された原文において KWIC に比べて提案手法の内容網羅性が 2 以上向上した件数と 2 以下低下した件数を表 4.7 に示す．2 以上向上するとは，例えば 2 から 4 に上がった場合や 1 から 5 に上がった場合を指す．表 4.7 より提案手法では内容網羅性を 10 件低下させはしたが 37 件も向上させていることから，適合性判定の結果が同じ適合であっても提案手法はより内容のある概要文を多く生成できたことがわかる．また，すべての文書と評価者の組み合わせにおいて，内容網羅性と適合性判定の正解の相関を求めた．ただし，適合性判定の正解として，原文と要約のいずれも適合であるとした．計算の結果，0.504 の相関係数が得られた．このことから，内容網羅性が高い概要文は，検索質問に適合するとともに，Web ページ自身も検索質問に適合する傾向にあると言える．評価者 3 名のうち 2 名の評価が一致した例の一部を表 4.8 に示す．これらの例からわかるように，KWIC では Web ページのタイトルと内容が重複していることがわかる．内容が重複する分だけ概要文に盛り込める情報が少なくなり，KWIC の内容網羅性は低かったと判断できる．これに対して，提案手法では重要文抽出の過程において MMR により Web ページのタイトルとの重複を避けていることで，内容網羅性を高めることができたといえる．

表 4.6: ベースラインと提案手法における適合性判定の変化の事例

分類 検索質問	タイトル KWIC	URL 提案手法
A 作家の値打ち，福田和也	スーパーダイアローグ は行 - ふ - 福田和也 □本 - ジャ... の回。特に「作家の値打ち」で最低点	http://imc.mechamilk.com/item/4898151361/ ... 福田和也の... 知識量と、頭の回転の速さに感心させられる。
A 花粉症，予防法	消痛術 粉 バスターズ 予防法から治療法まで... きましょう。 1 花粉症の薬 2 治療方法	http://www.bamcreation.com/i/kafun/ スギ花粉... 予防法から治療法まで完全網羅。... しておきましょう。
B 料理，切り方, 名称	料理の専門学校以外で 料理の専門学校以外で 料理板 Ads ... 菜の切り方（むき方までいろいろ名称が	http://www.casphy.com/bbs/test/r.cgi/cook/1208700437/-10 ... 包丁の使い方やだしの取り方、... 本などでも勉強できると思います。
B タバコ，害，訴訟	column をつけ、タバコメーカーを相手に訴訟を起こす...『吸いすぎれば害である』と分	http://www.a-park.com/column/column.php?mode=disp&yyyy=2005&mm=05&dd=13 ... 大量喫煙をした結果の責任が、... 困難さを引き起こしたのでした。
C 料理，切り方, 名称	岡崎市/市民協働推進課/平成 18 年度市民協働事業公募事業選定事業一覧 もを対象にした料理講習会の開催 第...：魚屋さん－魚の名称・種類、魚をさ	http://keitai.city.okazaki.aichi.jp/auto/www.city.okazaki.aichi.jp/yakusho/ka2605/ka067.htm ... 魚屋さん－魚の名称・種類、魚をさわる... さしみの切り方、試食）
C 世界遺産，日本	世界遺産ブームの日本 世界遺産ブームの日本 世界遺産ブームの日本 日本の世界遺産はいくつあるの	http://lipta.net/sekaiisanboom/ 日本にある世界遺産とこれから登録されそうな世界遺産を紹介します。
D タバコ，害，訴訟	ヒロコラム こ会社を訴える」訴訟が...。タバコを吸っている人の多くは、タバコの害に	http://www.fmyokohama.co.jp/i/column/154.html ... 人の多くは、タバコの害について知っているだろうし... 同じでしょ？
D タバコ，害，訴訟	未成年喫煙問題・藤沢市回答 くでタバコを... 害金額を算出して訴訟に持ちこんでいるのですが、藤沢市の損害	http://www.cityfujisawa.ne.jp/~559-mori/simin/Fmondai/tabako/kituenM.html ... 害から守ることを... するもので、... 注意することがうたわれています。

表 4.7: 内容網羅性に大きな変化があった件数

差分	件数	分類
内容網羅性が 2 以上向上	37	向上
内容網羅性が 2 以下低下	10	低下

表 4.8: 内容網羅性に大きな変化があった事例

分類 検索質問	タイトル KWIC	URL 提案手法
向上 低カロリー, お菓子, 食品	低カロリーで超満腹！無敵のダイエット食品 !! ガマンのダイエットはもう終わり !! 低カロリーで超満腹！無敵のダイエット食品 !! ガマ... めんやらお菓子やらを	http://shopping46.6japan.net/ ... お菓子やらを食べていたのをやめて... 豆乳クッキーにした結果。
向上 タバコ, 害, 訴訟	Anti-Smoke Site BLOG ＜東京地裁＞事業者に煙害防止配慮義務　タクシー禁煙訴訟で 裁＞事業者に煙害防止配慮義務 タクシー禁煙訴訟で [タバコ問題関連] Comme	http://www.anti-smoke-jp.com/blog/?ID=123 ... からだとして、... 損害賠償を求めた訴訟の判決で、... 請求を棄却した。
向上 点 数 制 度, 運転免許	運転免許の点数制度について 運転免許の点数制度について 運転免許の点数制度について。 主な点数表は	http://www.eonet.ne.jp/ da910/unten2.html ... 免許停止の前科がある場合は、... 免許停止や取り消しの... ことも
向上 花粉症, 予防法	花粉症の対策と予防 花粉症の対策と予防 花粉症の対策と予防法 適切な花粉症対策や予防法でつら	http://kafun.moba-info.net/ 適切な花粉症対策や予防法でつらい花粉症の季節を乗り切りましょう!
向上 花粉症, 予防法	花粉症をやっつけろ [無料 HP MINX] 花粉症をやっつけろ [無料 HP MINX] 花粉症をやっつけろ イヤ〜な花粉が飛ん	http://my.minx.jp/kahun ... 鼻炎に関して... ひどくなっちゃった時は…... 花粉症予防茶の最終兵器
向上 世 界 遺 産, 日本	旅行情報.mobi｜日本の世界遺産 ｜日本の世界遺産 世界遺産を巡る 過去から受け継がれる地球の宝物 日本国	http://www.ryokou-j.mobi/world_heritage/ 日本国内にも 14 の世界遺産があります。... 検索・宿泊予約も可能です。

4.3.2　FAQ 要約への適用に対する考察

提案手法はキーワードを与えれば関連する文書から非常に短い要約を作成する技術であり，その応用先として QA サイトを携帯端末に表示するための質問と回答の簡潔な要約の作成が考えられる．質問の要約は最初の一文を表示すれば十分と考え，回答の要約に提案手法を適用したときの有効性を調査するために，FAQ データを用いた追加実験を行った．質問とそれに対する回答からなる FAQ データに対して，質問からキーワードを抽出し，そのキーワードを検索質問とみなして回答の要約に適用した．FAQ データは，公開されている Yahoo!知恵袋データを用い，2004 年 7 月の質問でカテゴリが「レシピ・調理法」,「料理・食材」,「家事」のいずれかで，対応する回答（ベストアンサー）の長さが 500 バイト以上のうち，方法や手段を問う How 型質問と思われる質問を人手で抽出した 274 文書を対象とした．how 型を対象としたのは，理由を問う why 型や定義を問う definition 型に比べて要約の手がかりとなる明示的な情報が少なく，特に手がかり語などを用いない提案手法の評価に適していると考えたからである．

文書頻度と共起頻度は知恵袋データの質問と回答のすべてから獲得し，文節連接確率のモデルは毎日新聞 2003 年から 2007 年までを対象に学習した．各質問に存在する単語について tf*idf 値を計算し，上位 3 つをその質問のキーワードとした．提案手法との比較として，先頭から指定したバイト数だけを抽出するリード手法と，質問に依存せず回答に含まれる単語の tf*idf 値による重要文抽出と文短縮による方法を用いた．重要文抽出と文短縮の適用に対する考え方は提案手法と同様に，目的の要約長を越えるまで文を抽出し，選択済みの文が目的の要約長の 8 割に満たず次に選択する文を含めると目的の要約長を越える場合に，次に選択する文を残りの要約長の範囲で短縮する．提案手法との違いは，提案手法は質問に含まれるキーワードと単語の相関として対数尤度比を用いるのに対して，tf*idf 値による重要文抽出と文短縮の方法は質問の情報を用いない点である．目的の要約長は携帯端末への表示を前提とし，質問の要約と同時に提示することを想定して，さら

表 4.9: FAQ の回答の要約に対する可読性と応答性の評価

	可読性	応答性
リード手法（ベースライン）	2.15	2.69
tf*idf 値による文抽出・文短縮	3.29	2.65
提案手法	3.51	2.70

に短い 50 バイトに設定した．評価基準は回答の要約の読みやすさである可読性と答えとしての良さである応答性を各 5 段階とし，4 名の評価者により評価した．

表 4.9 に可読性と応答性に対する 4 名のスコアの平均値を示す．提案手法は，ベースラインであるリード手法と同等の応答性を保ちながら可読性を向上させたことがわかる．また，tf*idf 値による重要文抽出・文短縮ではベースラインよりも可読性が上がるものの応答性が下がることがわかった．これは質問に対する回答は比較的文章の先頭に現れるために応答性は下がらないが，途中で文が切れてしまうことによる可読性の問題が大きいと考えられる．一方，tf*idf 値による重要文抽出・文短縮では，述部をルートとする文抽出により文の途中で切れることはなくなるので可読性は上がるが，質問から抽出したキーワードとの対数尤度比を用いる提案手法ほどには応答性が上がらず，tf*idf 値では不十分であることがわかった．提案手法は，質問から抽出したキーワードが回答に存在しない場合もある中で，キーワードと相関の高い文を抽出することでベースラインと同等の応答性を維持した．また，述部をルートとする文短縮により文の途中で切れる問題を解消し可読性を向上させた．これらより，携帯端末への表示を想定した 50 バイトという非常に短い要約長の制限のもとで，提案手法は可読性と応答性を両立させたといえる．

第5章　テキストからの情報抽出

5.1　教師なし学習による関係抽出

本章では，情報抽出のためのテンプレートを獲得するコストが少ない情報抽出の手法について述べる．第 2 章では，アブストラクトによる要約が携帯端末への情報提示を指向したテキスト要約に有効であることを延べ，そのために必要な情報抽出の関連技術について述べた．情報抽出の課題は，情報抽出に必要なテンプレートを低コストで獲得することである．第 2 章で述べた情報抽出の先行研究は，この課題を十分に解決しているとは言えない．そこで，情報抽出のためのテンプレートを構成する情報のひとつとして，テキスト中の人名や地名等の固有表現の間に存在する関係に焦点を当て，大規模なタグなしコーパスから教師なし学習によって低コストで関係を抽出する方法を提案する．

本研究は固有表現の対が出現する文脈のクラスタリングに基づいている．これは，似たような文脈で出現する固有表現の対は，同じクラスタのインスタンスとしてまとめられ，それらのインスタンスは同じ関係にあると考えることができるからである．つまり，クラスタリングのプロセスを経て固有表現の対の間に存在する関係を発見する．固有表現の対の間にある関係が出現する文脈が複数の関係を表現している場合には，どのクラスタにも属さないか，最も頻度の高い関係を表すクラスタに属することを想定している．なぜならば，このような文脈は低頻度の関係を表す文脈には類似することはないと考えているからである．有益な関係は大規模コーパスにおいては頻繁に言及されているであろうし，逆に 1 度や 2 度しか言及されない関係は重要でないと考えている．

基本的なアイデアは次の通りである．

1. テキストコーパスから固有表現を抽出する
2. 共起する固有表現の対ととそれらが出現した文脈を獲得する
3. 固有表現の対同士で文脈の類似性を測定する
4. 固有表現の対をクラスタリングする
5. 固有表現の対が属する各クラスタにラベルを付与する

図 5.1 に提案手法の全体像を示す．初めに，新聞記事コーパスに対して固有表現抽出器を実行し，ORGANIZATION (ORG) A と B の対と ORGANIZATION (ORG) C と D の対を見つける．ある規定の距離の間に出現する A と B の対のすべてのインスタンスを収集する．次に，文脈として A と B の間に挟まれている単語列，例えば”be offer to buy”や”be negotiate to accuire”[1]を蓄積する．同様に C と D の間に挟まれている単語列も蓄積する．もし，A と B の文脈と C と D の文脈が類似していれば，これらの 2 つの対は同じクラスタに属し，A-B と C-D は同一の関係，この場合では M&A の関係にある．つまり，これらの ORGANIZATION の間の関係を発見できることを意味する．

5.1.1　固有表現抽出

本研究は完全に教師なし学習であり，関係に関する情報がタグ付けされたコーパスや，機械学習に用いるための人手で選択したシードを必要としない．その代わりに，大規模なコーパスから低コストで固有表現を発見するために固有表現抽出器を用いる．昨今の固有表現抽出器は実用レベルに達していることに加え，識別できるタイプも細分化されている．例えば，関根らは ORGANIZATION を COMPANY,

[1]POS タガーによってステミングされた単語の基本形を収集する．ただし，能動態と受動態を区別するために，動詞の過去分詞形は他と区別する．

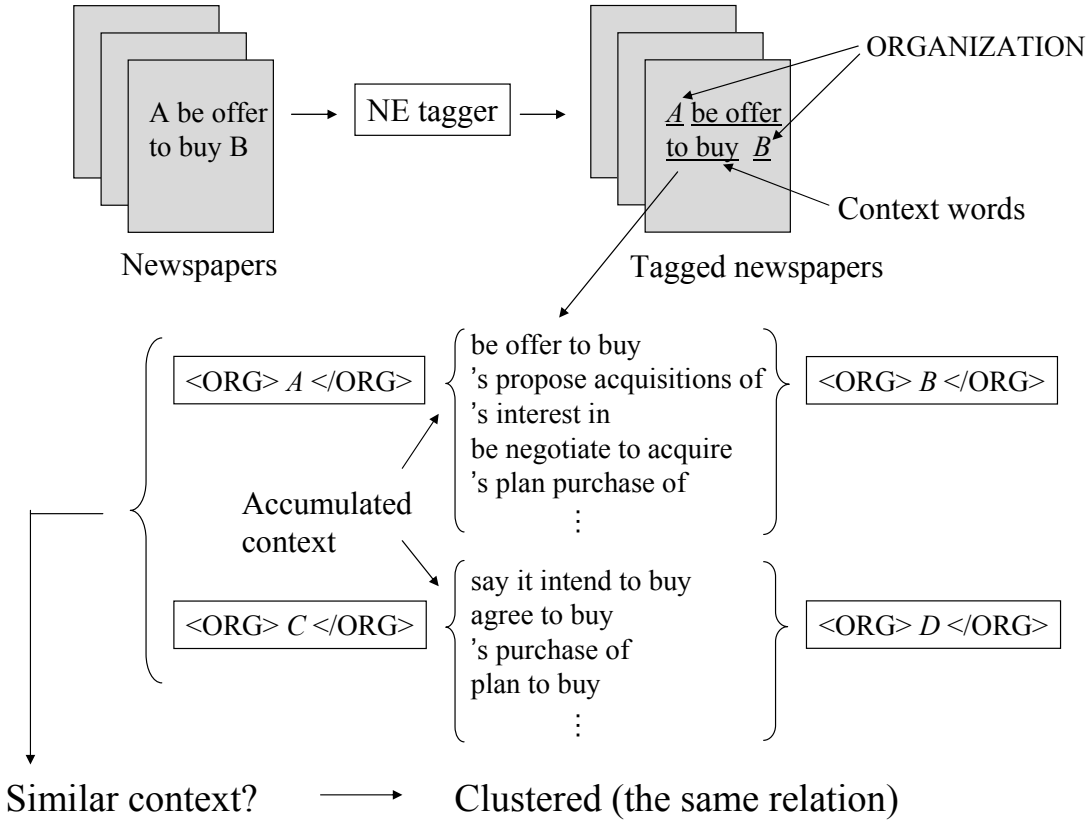
A be offer
to buy B
NE tagger
A be offer
to buy B
ORGANIZATION
Context words
Newspapers
Tagged newspapers
<ORG> A </ORG>
be offer to buy
's propose acquisitions of
's interest in
be negotiate to acquire
's plan purchase of
<ORG> B </ORG>
Accumulated
context
<ORG> C </ORG>
say it intend to buy
agree to buy
's purchase of
plan to buy
<ORG> D </ORG>
Similar context?
Clustered (the same relation)

図 5.1: 提案手法の概要

MILITARY, GOVERNMENT などのタイプに細分化し，150 種類の固有表現のタイプを提案している [40]．固有表現の対に存在する関係の抽出の精度を高めるためには，細分化された固有表現のタイプは有利に働くと考えられる．そこで，この 150 種類の固有表現のタイプを識別できる固有表現抽出器 [39] を用いる．

5.1.2　固有表現対とその文脈の抽出

固有表現の対の共起とは，同一文内の N 単語以内で共起して現れる 2 つの固有表現と定義する．コーパスからそれぞれの共起について，2 つの固有表現に挟まれた単語列を収集する．ここで，単語は基本形に変換して収集する．収集される単語列は，固有表現の対が出現する文脈とみなすことができる．同じ固有表現の対でも出現順序を考慮し，出現する順序が異なる場合は別の文脈として扱う．例えば，e1 と e2 を固有表現とすると，e1…e2 と e2…e1 は別の文脈と考える．共起頻度が小さい固有表現の対は，関係の抽出においては信頼性が低くノイズとなる可能性があるので，あらかじめこれを除外するための閾値を設ける．

5.1.3　固有表現対における文脈の類似性

収集された固有表現の対の文脈の集合の間の類似度を計算するために，ベクトル空間モデルとコサイン類似度を採用する．固有表現の対同士の類似度の計算は，固有表現の対が同じタイプのもの（これをドメインと呼ぶことにする）に限定する．例えば，PERSON と GPE[2]の対は別の PERSON と GPE の対との類似度だけが計算される．つまり，PERSON-GPE ドメインにおいては PERSON と GPE の間の関係を抽出することとする．

[2]情報抽出に関する米国の評価型プロジェクト ACE (Automatic Content Extraction) で導入された固有表現のタイプであり，MUC (Message Understanding Conference) で導入された Location から派生した，統治機能が存在する場所を指す．

文脈ベクトルを作成する前に，ストップワードや，並列の表現およびコーパスに特有な表現は，固有表現の対を表現する際にノイズとなる可能性があるので，これらはあらかじめ除いておく．文脈ベクトルは収集されたフレーズにおける単語の集合 (bag of words) から構成され，ベクトルの各要素の値は単語頻度と文書頻度の逆数の積である tf*idf が用いられる．単語頻度は，収集された文脈における単語の頻度とする．このとき，関係の方向性を議論するために固有表現の出現順序を考慮し，例えば，e1 と e2 を固有表現とすると，単語 w_i が e1…e2 の文脈で L 回出現し，e2…e1 の文脈で R 回出現する場合には，単語 w_i の単語頻度 tf_i を $L - R$ と定義する．なぜならば，関係の 2 つの引数である固有表現のタイプが同じ場合に，固有表現の出現順序の異なりは関係の方向を獲得するのに有効と考えたためである．文書頻度は，コーパスのすべての文書におけるある単語を含む文書数である．文脈ベクトル α の長さ $|\alpha|$ が内容語の欠如により極端に短い場合は，文脈ベクトルに対する信頼性が低下するので，あらかじめ長さの短い文脈ベクトルは除外するための閾値を設ける．

文脈ベクトル α と β のコサイン類似度 $cosine(\theta)$ は，次式により計算される．

$$cosine(\theta) = \frac{\alpha \cdot \beta}{|\alpha||\beta|} \tag{5.1}$$

コサイン類似度は 1 から -1 までの値を取る．コサイン類似度が 1 とは固有表現の対が正確に同じ文脈を共有し出現する順序が同じであることを意味し，コサイン類似度が -1 とは固有表現の対が正確に同じ文脈を共有するが出現する順序が逆であることを意味する．コサイン類似度の値の正負により，2 つの固有表現の対に存在する関係の方向が同じか否かを判別できる．

5.1.4　固有表現対のクラスタリング

固有表現の対の文脈ベクトルの類似度を計算し，文脈ベクトル間の類似度に基づくクラスタリングを行い，固有表現の対のクラスタを生成する．最適なクラス

タ数は事前に予測できないので，階層的クラスタリングの手法を用いる．階層的クラスタリングの手法はいくつかあるが，本研究ではクラスタを保守的に作成していく complete linkage を用いる．complete linkage ではクラスタ間の距離はクラスタ内の最も遠いノードの距離とする．

5.1.5　クラスタへの関係ラベルの付与

同じクラスタに属する多くの固有表現の対は文脈に共通する単語が存在する．このような単語はクラスタの特徴を表しており，言い換えれば，このような単語は特有の関係を表現しているとみなすことができる．

本研究では，同一のクラスタに属する固有表現の対のすべての組み合わせの文脈において共通して出現する単語の頻度を単純にカウントする．頻度は組み合わせ数により正規化される．クラスタにおいて共通して出現する頻度の高い単語は，そのクラスタのラベルとみなすことができる．つまり，もしそのクラスタが同一の関係を持つ固有表現の対から構成されるのであれば，そのような単語はその関係のラベルとみなすことができる．

5.2　実験

5.2.1　実験の設定

コーパスから教師なし学習で関係を抽出する方法を評価するために実験を行った．実験では，対象言語を英語とし，コーパスとして New York Times 1995 年版を用いた．まず，関係抽出のための 3 つのパラメータを定義し，ストップワードと並列の表現および New York Times に特有な表現を規定した．パラメータは経験的に，文脈の長さは 5 単語以内に，固有表現の対の共起頻度の閾値を 30 に，また文脈ベクトルの長さの閾値を 10 に設定した．ストップワードには，3 回未満の

表 5.1: 新聞記事コーパスから抽出された固有表現の対に対する人手による関係の分類

P.-G.	President	Senator	Governor	Prime Minister	Player	Living	Coach
NE 対の個数	28	21	17	16	12	9	8
	Republican	Secretary	Mayor	Enemy	Working	その他 (2, 3)	その他 (1)
	8	7	5	5	4	20	17
C.-C.	M&A	Rival	Parent	Alliance	Joint Venture	Trading	その他 (1)
NE 対の個数	35	8	8	6	2	2	4

低頻度語と 100,000 回を超える高頻度語とした．並列の表現は“ ,.*, ”,“ and ”, “ or ”とし，New York Times に特有な表現には記事の先頭にある日付と発行所を表す“ ） -- ”と規定した．

New York Times 1995 年版から，規定したパラメータや条件を満たすように抽出した固有表現の対と同時に収集された文脈を，評価実験のデータセットとした．提案手法の妥当性を実証するには，多くの固有表現の対が高い頻度で出現することが期待されるドメインでなければならない．実験では，PERSON と GPE の対である PERSON-GPE (以下， P.-G. と呼ぶ) と COMPANY と COMPANY の対である COMPANY-COMPANY (以下，C.-C. と呼ぶ) の 2 つのドメインを対象とした． 関係抽出の評価を行うために，どのような関係の事例がどれくらい存在するのかを調べるため人手でデータセットを分析した．P.-G. ドメインでは，177 個の固有表現の対が得られ，人手により 38 個のクラス（関係）に分類された．C.-C. ドメインでは，65 個の固有表現の対が得られ，人手により 10 個のクラス（関係）に分類された．表 5.1 に 2 つのドメインにおけるクラスと固有表現の対の個数を示す．なお，提案手法を正確に評価するために，固有表現抽出の誤りは除外した．

5.2.2　固有表現対のクラスタリングの評価

固有表現の対のクラスタリングとクラスタへのラベル付けを分けて評価する．最初のステップにおいて，2 つ以上の固有表現の対からなるクラスタを対象とした．評価対象となる各々のクラスタにおいて，クラスタ内の固有表現の対に人手により割り当てられた関係のうち，最も多い関係をそのクラスタの関係 (R) とする．関係 (R) のクラスタの中で関係 (R) が割り当てられている固有表現の対は正解としてカウントし，評価対象の全クラスタにおける正解の固有表現の対の総数を $N_{correct}$ と定義する．正解とカウントされない固有表現の対は誤りとしてカウントし，評価対象の全クラスタにおける誤りの固有表現の対の総数を $N_{incorrect}$ と定義する．クラスタリングの評価は，以下に定義する適合率と再現率および F 値を用いた．

適合率 (P) クラスタリングされた固有表現の対のうち，正解の固有表現の対の割合を求める．

$$P = \frac{N_{correct}}{N_{correct} + N_{incorrect}} \tag{5.2}$$

再現率 (R) データセットの中で人手により付与された関係の事例が 2 つ以上存在する関係における固有表現の対の総数 N_{key} のうち，正解の固有表現の対の割合を求める．

$$R = \frac{N_{correct}}{N_{key}} \tag{5.3}$$

F 値 (F) F 値は適合率と再現率の組み合わせによって定義される．

$$F = \frac{2PR}{P + R} \tag{5.4}$$

コサイン類似度の閾値を下げていくと，クラスタは徐々に結合して大きくなり，ついにはひとつの大きなクラスタになるため，これらの 3 つの値は可変となる．

P.-G. ドメインのクラスタリングの結果を図 5.2 に， C.-C. ドメインのクラスタリングの結果を図 5.3 に示す．グラフを右から左へ見れば，コサイン類似度の閾値が下がるにつれて，適合率は下がり再現率は閾値がほぼ 0 になるまで上がる．閾

表 5.2: コサイン類似度の閾値を 0 の直前としたときの適合率，再現率，F 値

	適合率	再現率	F 値
P.-G.	79	83	80
C.-C.	76	74	75

値が 0 の時に再現率は大きく落ち込むが，これはクラスタ間の距離が 0 となり多くのクラスタが結合されるため，正解の固有表現の対の数が減少するためである．P.-G. ドメインで最も高い F 値は 82 であり，C.-C. ドメインで最も高い F 値は 77 であった．いずれのドメインにおいても，コサイン類似度の閾値が 0 の付近でほぼ一致して最も高い F 値が得られることがわかった．一般的には，最も良いコサイン類似度の閾値を事前に決定することは難しい．しかしながら，2 つのドメインにおいていずれも 0 の直前で高い F 値が得られたことから，ドメイン間で統一したコサイン類似度の閾値を 0 の直前に設定した．閾値が 0 の直前ということは，クラスタ内のすべて固有表現の対には互いに他のいずれとも文脈に共通する単語が少なくともひとつ以上存在することを意味する．このため，他のドメインに対しても高い F 値が得られることが期待できる．コサイン類似度の閾値を 0 の直前に設定したとき，P.- G. ドメインでは 34 個，C.-C. ドメインでは 15 個のクラスタが得られた．この閾値での精度は表 5.2 に示すように，P.-G. ドメインで F 値 80，C.-C. ドメインで F 値 75 を達成し，どちらも最も高い F 値に近い値が得られた．

5.2.3　クラスタに付与される関係ラベルの評価

次に，固有表現の対により構成されたクラスタから関係が特定できるかどうかを評価する．それぞれのドメインにおいて，各クラスタにおいて人手で付与された最多の関係（クラスタにおいて最も多く表現された関係）と，クラスタ内のすべての固有表現の対に対して最多の関係を持つ固有表現の対の割合を表 5.3 の左側

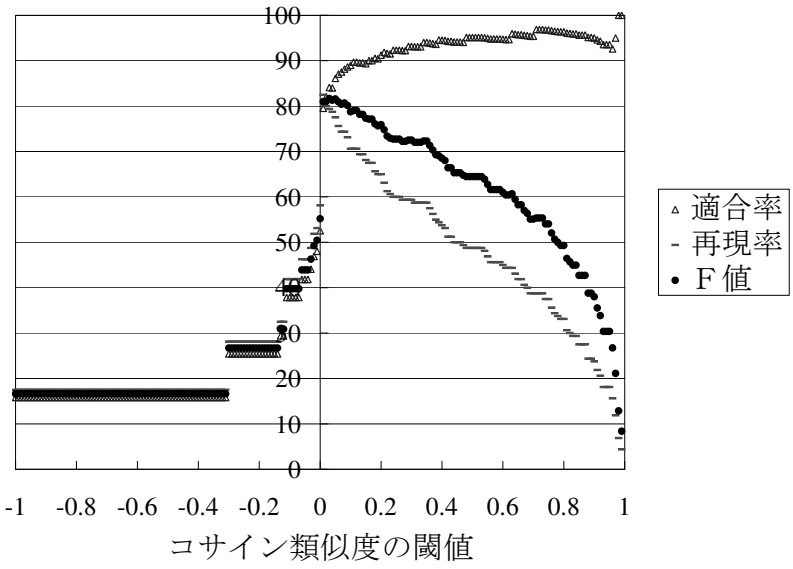

図 5.2: P.-G. ドメインでのクラスタリングにおけるコサイン類似度の閾値と適合率，再現率，Ｆ値の関係

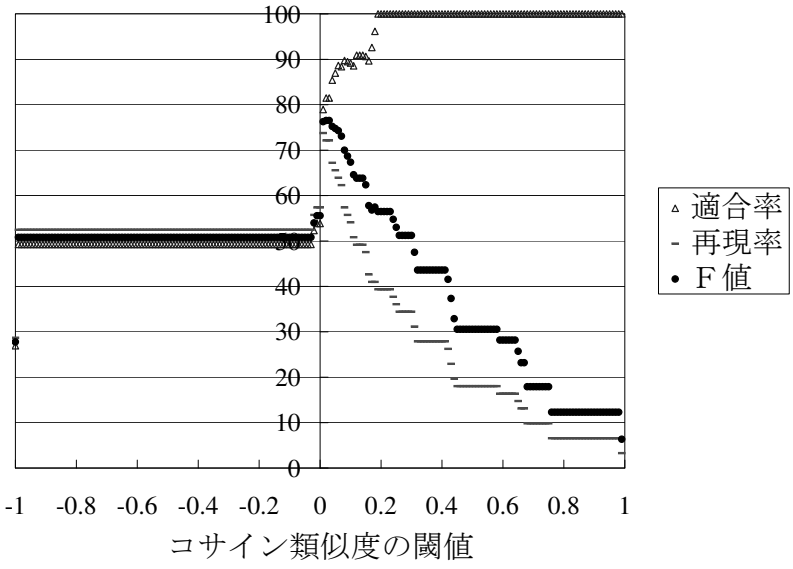

図 5.3: C.-C. ドメインでのクラスタリングにおけるコサイン類似度の閾値と適合率，再現率，Ｆ値の関係

表 5.3: 各クラスタにおける最多の関係と頻度の高い文脈に共通な単語

最多の関係	割合	文脈に共通な単語 (相対頻度)
President	17 / 23	President (1.0), president (0.415), ...
Senator	19 / 21	Sen. (1.0), Republican (0.214), Democrat (0.133), republican (0.133), ...
Prime Minister	15 / 16	Minister (1.0), minister (0.875), Prime (0.875), prime (0.758), ...
Governor	15 / 16	Gov. (1.0), governor (0.458), Governor (0.3), ...
Secretary	6 / 7	Secretary (1.0), secretary (0.143), ...
Republican	5 / 6	Rep. (1.0), Republican (0.667), ...
Coach	5 / 5	coach (1.0), ...
M&A	10 / 11	buy (1.0), bid (0.382), offer (0.273), purchase (0.273), ...
M&A	9 / 9	acquire (1.0), acquisition (0.583), buy (0.583), agree (0.417), ...
Parent	7 / 7	parent (1.0), unit (0.476), own (0.143), ...
Alliance	3 / 4	join (1.0)

に示す．また，各クラスタにおいて文脈に共通する単語とその相対頻度を表 5.3 の右側に示す．もし，2 つの固有表現の対が文脈に特定の単語を共有するならば，これらはその単語を介してリンクされると考える．1 つのクラスタにおいて可能な限りのすべてのリンクを考えると，N 対のクラスタには $N(N-1)/2$ 個のリンクが存在する．単語の相対頻度は，考えられるすべてのリンク数に対する，特定の単語により張られたリンクの数の割合である．相対頻度が 1.0 であれば，特定の単語はすべての固有表現の対の文脈に共通して出現する．実験では要素数の小さなクラスタにおいても意味のある関係を得ることができたが，要素数の小さなクラスタにおいて文脈に共通する単語は信頼性が低いために要素数の小さなクラスタは対象外とした．実験結果から，要素数が大きいクラスタについては，固有表現の対のクラスタリングにより，各クラスタの大部分を占める固有表現の対に存在する関係を表す適切なラベルが得られ，関係が特定できたことを意味する．言い換えれば，コーパスから同一の関係にある固有表現の対だけでなく，その固有表現の対の間にある関係を表すラベルも高い精度で発見できたと言える．

5.3　考察

実験結果は高い精度で関係を抽出することができることを示した．PER-GPE ドメインの方が COM-COM ドメインよりも少し精度が高かった．おそらくこれは COM-COM ドメインよりも PER-GPE ドメインの方がコサイン類似度の高い固有表現の対が多かったからである．しかしながら，いずれのドメインにおいても特にコサイン類似度が 0.2 以下ではグラフは類似していた．2 つのドメインにおける相違と教師なし学習による関係抽出について以下の観点について議論する．

- 関係の特性
- 適切な文脈の長さ
- 最適なクラスタリング手法の選択
- 低頻度の固有表現対の網羅

以下では順にこれらの観点について述べる．

5.3.1　関係の特性

実験の結果，C.-C. ドメインは P.-G. ドメインよりも F 値が低かったが，これには 2 つの理由があると考えている．ひとつは，関係の類似性の問題で比較的近い関係を判別することの難しさがあげられる．例えば，C.-C. ドメインでは，“ M&A ”の関係にある会社の対はその後“ Parent ”の関係として現れることもある．もうひとつは，関係の方向の問題で正しい方向まで含めて抽出することの難しさがあげられる．多くの関係には方向があり，関係の引数となる固有表現の順序は入れ替え不可能な非対称性を持つ．P.-G. ドメインでは， 2 つの固有表現のタイプが異なるために関係の方向の問題は顕在化しなかったが，C.-C. ドメインでは 2 つの固有表現のタイプが同じなので関係の方向を考慮しなければならない．このため我々

は，関係の方向まで抽出することを試みた．例えば，会社 *A,B,C,D* がある場合に，$A \rightarrow B$ か $B \rightarrow A$ かまた $C \rightarrow D$ か $D \rightarrow C$ かという関係の方向を識別するために，$A \rightarrow B$ と $C \rightarrow D$ および $A \rightarrow B$ と $D \rightarrow C$ のそれぞれで文脈ベクトルの類似度を比較することにより，方向が同じになるようにクラスタリングを行った．しかしながら，正しい方向よりも誤った方向でより多く文脈の単語が共有される現象も見られ，15 個のうち 2 個のクラスタで方向の誤りがあった．

5.3.2　文脈の長さ

いずれのドメインでも関係抽出における誤りには，クラスタが生成されないために固有表現の対に存在する関係が得られない抽出不能なケース (Miss) と，クラスタにおける多数の関係とは異なる関係にある固有表現の対が混じっている抽出誤りのケース (False Alarm) がある．これらの原因はいずれも文脈に共通な単語の有無に起因する．Miss では，固有表現の対が持つ関係を明確に表すような単語が文脈に共通して現れないことが原因である． False Alarm では，固有表現の対が持つ関係とは異なった単語が偶然に文脈に共通して出現することが原因である．実験では 2 つの固有表現の間の 5 単語以内を文脈としたが，文脈の長さを延長したりあるいは 2 つの固有表現の外側まで拡大すれば文脈に共通な単語が得られ，Miss を減らせる可能性が高くなる．しかしながら，偶然に文脈に共通する単語も増えて，逆に False Alarm が増える可能性もあるため，最適な文脈の長さは慎重に決めなければならない．

5.3.3　クラスタリングの方法

クラスタリングの手法として，complete linkage の他にも single linkage と average linkage を試した．complete linkage が最も高い F 値を出し，最適なクラスタリング手法であった．さらに，他の 2 つのクラスタリング手法では最もよい F 値となる

コサイン類似度の閾値が 2 つのドメインで異なったが，complete linkage では最適な閾値が 2 つのドメインでほぼ一致した．complete linkage の最適な閾値は 0 の直前に決定された．閾値がちょうど 0 に達すると固有表現の対はどの単語も共有する必要がなくなるので F 値は急激に落ち込む．閾値を 0 の直前に設定するということは，同一のクラスタ内の固有表現の対の組み合わせは少なくとも 1 つの単語を共有し，これらの多くは関係を表すのに適切な単語であることを意味する．これは文脈の長さに関連があると考えている．本研究では，異なる関係に横断的に共通する単語がノイズとなることを防ぐために，5 単語という比較的短い文脈の長さを用いた．complete linkage と短い文脈の長さの組み合わせは関係抽出において有効であることを証明した．

5.3.4　低頻度の固有表現対

実験では固有表現が共起する頻度の閾値を 30 に設定したので，低頻度の固有表現の対は抽出していない可能性があり，その中のいくつは価値のある関係にあるかもしれない．低頻度の固有表現の対では，文脈のバリエーションが少なく文脈ベクトルの長さが短くなるので，これらの関係を精度良く抽出することは難しい．この問題を解決する 1 つの方法はブートストラッピングであると考えている．ブートストッピングの問題ははじめにシードをどのように選択するかという問題があるが，提案手法によってこの問題を解決できると考えている．クラスタ内の固有表現の対における文脈に共通して出現する単語を多く持つ固有表現の対は，有望なシードになるはずである．一度このようにシードを獲得できれば，低頻度の固有表現の対もこれらのシードが持つ文脈の単語との重なりに基づいてシードと同じクラスタに属するようになるだろう．

第6章　結論

本論文では，携帯端末への情報提示を指向したテキスト要約の手法を提案した．実社会で有用なアプリケーションに適用するために，要約の対象とするテキスト，要約の利用目的，要求されている要約長という観点に基づいて，それぞれに適切なアプローチを選択した．実社会に導入するためには，低コストでありながらも高い精度を有することが求められる．しかしながら，これらは精度を上げるにはコストがかかるというトレードオフの関係にある．提案した手法は，いずれも低コストでありながら，実社会における有用性としては一定以上のレベルに達していることを確認した．

電子メールに対しては，電子メールに特有な引用や署名，箇条書きなどのスタイルの特徴を考慮して比較的短い文の単位に分割し，電子メールに特有な表現の特徴を利用したルールに基づいて重要な箇所を網羅的に抽出する重要文抽出の手法を提案した．評価実験では，携帯電話の画面の大きさを考慮して要約長を 500byte に設定することを想定し，6 文以上の文書から 5 文を選択するタスクを用意した．データの多くの割合を占める 1.2KB 以内の比較的短い文書において，文書の先頭から 5 文を選択するリード手法に比べて，提案手法により選択した 5 文の方が人手により選択された 5 文との文の重なりが大きいことを示した．また，着信通知のタスクに適用することを想定し再現率と適合率で評価したところ，再現率，適合率ともに 90%以上の高い精度が得られることを確認した．さらに，要約による着信通知システムを 3ヶ月利用した 19 名のユーザにアンケートを取った結果，79%のユーザが電子メールの要約に対して満足あるいは特に不満を持っていないことがわかった．

Web ページに対しては，タイトルとの重複を排除しながら検索質問との相関に基づき検索質問に適合する文を選択する重要文抽出と，選択した文の述部をルートとした部分木の中から文節網羅率と文節連接確率に基づくスコアが最も大きい部分木を出力する文短縮との組み合わせにより概要文を生成する手法を提案した．Web ページの検索結果を携帯端末へ提示するため，要約長を 70byte に設定して概要文を生成し評価を行った．評価実験の結果，検索の適合性判定においては KWIC による方法の F 値が 0.571 に対して提案手法の F 値が 0.605 であり，要約の文法性，非冗長性，内容網羅性のすべてにおいて 5 段階の主観評価で KWIC による方法を上回ることを確認した．適合性判定では，特に不適合な Web ページから正しく不適合と判定できる概要文を多く生成したことにより適合率が上昇した．また，内容網羅性が高い概要文は検索質問に適合し Web ページもまた検索質問に適合する傾向があり，適合という同じ判定結果においても内容網羅性が大きく向上する概要文が増加した．以上から，提案手法は，電子メールの着信通知のための要約，ならびに，検索のための Web ページの概要文に対して有効であることを確認した．

また，携帯端末向けの短い要約を生成するために有利となる情報抽出を用いたアブストラクトの生成によるテキスト要約のアプローチでは，情報抽出に必要なテンプレートを低コストで獲得することが必要である．本論文では，テンプレートを構成する情報のひとつとして 2 つの固有表現の間に存在する関係を取り上げ，テキストから低コストで関係を抽出するために，大規模なテキストコーパスを用いた教師なし学習の手法を提案した．固有表現の対が出現する文脈の類似度に基づいて固有表現の対をクラスタリングすることにより，固有表現の対に存在する関係を発見する．新聞記事 1 年分のテキストコーパスを用いて，人物 (Person) と統治機能が存在する場所 (GPE) の間の関係，および，会社 (Company) と会社 (Company) の間の関係を対象に評価を行った．評価実験の結果，同一の関係を有する固有表現の対に対して，人物と統治機能が存在する場所の関係では F 値 80 が，会社と会社の関係では F 値 75 が得られ，高い精度でクラスタリングできることを確認した．

さらに，クラスタに付与すべき関係のラベルについては，クラスタ内の文脈に最も多く共通する単語は平均して 89%の固有表現の対において正しい関係ラベルであることを確認し，クラスタリングされた固有表現の対に対して高い精度で関係のラベルを自動的に与えることができることを示した．

最後に今後の課題について述べる．すでに述べたように，本論文では電子メールの着信通知としての要約と Web ページの検索結果に用いる概要文を目的とした手法を提案したが，これら以外の他の目的についても幅広く本手法が適用できるかどうかを検証する必要がある．電子メールについては，受信する内容があらかじめ想定できない場合が多いため generic な要約を議論してきたが，もし特定の情報に関する内容だけ知ることができればよいというユーザに対しては，query-biased な要約が適している可能性もある．この場合は事前にユーザの関心の高いキーワードを取得することにより，Web ページの要約で提案した検索質問との相関を計算する方法が適用できるかもしれない．Web ページについては，検索結果を提示するための概要文を生成する手法を議論したが，概要文から Web ページの原文にアクセスする場合にも Web ページの全文ではなく，例えば 500byte 程度の概要文より長い要約を提示することが考えられる．この目的には検索質問を考慮した query-biased な要約が適しているのであれば，単に要約長を長くしただけの提案手法が適用できる可能性がある．しかしながら，そもそも検索質問を考慮した query-biased な要約を提示すべきなのか，あるいは，検索質問とは無関係に generic な要約を提示すべきなのかという方針の選択は検討に値するだろう．

要約の品質についてはおおむね満足できるものであると考えているが，さらに精度を上げるためには以下の課題がある．まず，本論文で対象としたテキストは，一般のユーザにより記述される電子メールや Web ページが含まれるが，これらのテキストの質は概して高くはない．本論文で論じた提案手法では，テキストからの重要文抽出の単位として文を用いるため，後の処理に影響を及ぼさないためには高い精度でテキストを文の単位に分割することが求められる．しかしながら，質

の高くないテキストから正確に文の単位で分割することはそう容易ではない．正確に文分割ができるようになれば，さらに要約の品質も上がるであろう．

次に，電子メールの要約におけるルールに付与するスコアや，Web ページからの概要文の生成におけるパラメータ，および，教師なし学習による関係抽出におけるパラメータの設定に関する問題が残る．スコアやパラメータの値を変化させると結果が変わる可能性があるため，さらに精度を上げるためには実際に対象とするテキストを用いながらこれらの値をチューニングすることが必要となるであろう．

また，本論文で論じたテキスト要約の手法は，テキストの先頭にある文を抽出するリード手法や検索質問の周辺を抜粋する KWIC による方法などの既存の一般的な方法に比べると計算量が多い．このため，本手法を実用化するには，実時間を意識した実装と運用が必要になる．特に Web ページの検索では，大規模な文書集合に対して現実的な時間で処理することが求められる．しかしながら，検索時におけるオンラインでの処理ではなく，オフラインでの処理で事前に計算できる部分も多く存在する．例えば，各単語について検索質問との相関を計算するためには，文書集合全体から検索質問の頻度，検索される Web ページ内の単語の文書頻度および検索質問と各単語との共起頻度を取得しなければならないが，Web ページのテキストが長くなり単語の異なり数が増えるにしたがって，これらの処理の計算コストは大きくなる．しかしながら，検索質問と検索結果として返される Web ページが事前に与えられるならば，検索質問の文書頻度，検索結果として返される Web ページ内の各単語の文書頻度および検索質問と各単語との共起頻度はオフラインの処理で取得することができ，検索質問と各単語の相関もオフラインで計算することが可能である．本手法を実用化するためには，オフラインでの処理を含めた運用を考慮した実装を検討する必要がある．

一方，本論文で論じた情報抽出の手法は，大規模なテキストデータを用いることで比較的出現頻度の高い固有表現の対を対象として，それらに存在する関係を

抽出した．しかしながら，出現頻度が小さくても重要な関係を持つ固有表現の対が存在することも考えられる．このため，半教師有り学習の方法を組み合わせ，本手法で得られる固有表現の対をシードとしたブートストラッピングを行うことによって，低頻度の固有表現の対をさらに獲得できる可能性がある．さらに，抽出した情報に基づいてテキストを生成する手法とを組み合わせ，固有表現だけでなく専門用語などにも関係の対象を拡張することにより，同一の事象に言及している複数のテキストからアブストラクトとしての短いテキストを低コストで生成することができるようになるかもしれない．

このように残された課題は多いが，以上に述べた点を考慮しながら，今後もテキスト要約の研究に取り組んでいくことが必要である．本論文で論じたテキスト要約および情報抽出の手法が携帯端末を用いた効果的な情報アクセスの実現の一助となり，実社会の多くのユーザに利便性をもたらすことに貢献できれば幸いである．

謝辞

本論文をまとめるのにあたり，種々のご指導，ご教示を賜りました東京工業大学大学院総合理工学研究科の奥村学教授に深く感謝いたします．また，貴重なご助言を数多く頂きました東京工業大学大学院総合理工学研究科の新田克己教授，小野田崇連携教授，渡邉澄夫教授，長谷川修准教授に心より御礼申し上げます．

本研究はNTTサイバースペース研究所において，多くの方々のご理解とご支援のもとに行われたものです．研究の機会を与えて頂くとともに数々のご援助を頂きました，東田正信氏（現在，NTTソフトウェア株式会社），大山芳史氏（現在，NTTアドバンステクノロジ株式会社），小原永氏（現在，NTTアドバンステクノロジ株式会社），今村明弘氏（現在，NTTアイティ株式会社），児島治彦氏に厚く御礼申し上げます．

本論文の第3章の研究を行うにあたっては，NTTサイバースペース研究所の林良彦氏（現在，大阪大学）と山崎毅文氏（現在，NTTサイバーソリューション研究所）より数多くのご助言やご支援を頂きました．同研究所の廣嶋伸章氏（現在，NTTサイバーソリューション研究所）には評価実験システムの運用やアンケート調査において多大なるご尽力を頂きました．また，同研究所の松岡浩司氏（現在，NTTソフトウェア株式会社），堀井統之氏（現在，NTTサイバーソリューション研究所）には研究の動機を与えて頂きました．同研究所の浅野久子氏にはプロトタイプの実装においてご協力を頂きました．ここに深く御礼申し上げます．

本論文の第4章の研究を行うにあたっては，NTTサイバースペース研究所の菊井玄一郎氏，今村賢治氏，西川仁氏からのご助言やご支援がなくては完遂できるものではありませんでした．ここに深く感謝いたします．また，本研究の実施にあ

たっては，国立情報学研究所が提供している NTCIR-4 Web テストコレクションの Web 検索評価データと，ヤフー株式会社が国立情報学研究所に提供した Yahoo! 知恵袋データを利用させていただきました.

本論文の第 5 章の研究は，米国ニューヨーク大学客員研究員として行われたものです．米国滞在中において熱心にご指導頂き，数多くのご教示，ご支援を賜りましたニューヨーク大学の Ralph Grishman 教授と関根聡准教授に深く感謝いたします.

また，研究者としての基本的な姿勢を身に付けていなければ本論文をまとめることはできませんでした．慶應義塾大学在学中にご指導頂きました，慶應義塾大学の土居範久教授（現在，中央大学），土居研究室の高田眞吾氏（現在，慶應義塾大学）と，NTT 基礎研究所の岡田美智男氏（現在，豊橋技術科学大学）と栗原聡氏（現在，大阪大学）に深く感謝いたします．そして，日本電信電話株式会社への入社以降にご指導頂きました NTT 情報通信研究所の加藤恒昭氏（現在，東京大学），中野有紀子氏（現在，成蹊大学），高木伸一郎氏（現在，NTT ソフトウェア株式会社）と NTT サイバースペース研究所の永田昌明氏（現在，NTT コミュニケーション科学基礎研究所），松尾義博氏に心より感謝いたします.

本論文で述べた研究を進めるにあたり，議論に参加していただいた，NTT サイバースペース研究所音声言語メディア処理プロジェクトの皆様，ニューヨーク大学 Proteus Project の皆様，東京工業大学奥村研究室の皆様に謹んで感謝の意を表します.

最後に，私を温かく見守ってくれた両親と，常に私の心の支えとなってくれた妻・亜紀，娘・瑞穂，息子・慶信に心より感謝の意を捧げます.

業績一覧

学術論文

- 長谷川 隆明, 西川 仁, 今村 賢治, 菊井 玄一郎, 奥村 学. 携帯端末のためのWebページからの概要文生成. 人工知能学会論文誌, Vol.25, No.1, pp.133-143, 2010.1.
- 長谷川 隆明, 林 良彦, 山崎 毅文. 電子メールにおける重要文抽出と携帯電話向け要約システムへの適用. 情報処理学会論文誌, Vol.45, No.7, pp.1745-1754, 2004.7.
- 長谷川 隆明, 高木 伸一郎. 文書構造の認識と言語の特徴の利用に基づく電子メールからのスケジュールと ToDo の抽出. 情報処理学会論文誌, Vol.40, No.10, pp.3694-3705, 1999.10.
- 廣嶋 伸章, 長谷川 隆明, 奥 雅博. Web ページのヘッドライン生成のための統計的要約. 自然言語処理, Vol.12, No.6, pp.113-128, 2005.11.

国際会議（査読付き）

- Takaaki Hasegawa, Satoshi Sekine, Ralph Grishman. Discovering Relations among Named Entities from Large Corpora. Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL 2004), pp.415-422, 2004.7.

- Takaaki Hasegawa, Hisashi Ohara. Automatic Priority Assignment to E-mail Messages Based on Information Extraction and User's Action History. Proceedings of the 13th International Conference on Industrial and Engineering Applications of Artificial Intelligence and Expert Systems (IEA/AIE 2000), pp.573-582, 2000.6.

- Takaaki Hasegawa, Yukiko I. Nakano, Tsuneaki Kato. A Collaborative Dialogue Model Based on Interaction between Reactivity and Deliberation. Proceedings of the 1st International conference on Autonomous Agents (Agents 1997), pp.83-87, 1997.2.

- Tsuneaki Kato, Yukiko I. Nakano, H. Nakajima, Takaaki Hasegawa. Interactive Multi modal Explanations and their Temporal Coordination. Proceedings of the 12th European Conference on Artificial Intelligence (ECAI 1996), pp.261-265, 1996.8.

解説記事

- 長谷川 隆明. 私のブックマーク:テキストマイニング. 人工知能学会誌, Vol.16, No.6, pp.893-896, 2001.11.

- 市村 由美, 長谷川 隆明, 渡部 勇, 佐藤 光弘. テキストマイニング:事例紹介. 人工知能学会誌, Vol.16, No.2, pp.192-200, 2001.3.

- 長谷川 隆明, 中野 有紀子, 永田 昌明, 小原 永. マルチメディア時代を支える言語処理技術. NTT R&D, Vol.49, No.3, pp.172-180, 2000.3.

- 長谷川 隆明, 浅野 久子, 堀井 統之. 電子メールのインテリジェントサービス. 人工知能学会誌, Vol.14, No.6, pp.951-958, 1999.11.

研究報告

- 長谷川 隆明, 今村 賢治, 菊井 玄一郎. 検索キーワードとコンテキストとの相関に基づく検索文書のリランキング. 言語処理学会 第 14 回年次大会, pp.107-110, 2008.3.
- 長谷川 隆明, 関根 聡, Ralph Grishman. 教師なし学習による関係抽出に基づくパラフレーズの獲得. 言語処理学会 第 11 回年次大会，pp.1145-1148, 2005.3.
- 長谷川 隆明，林 良彦. 隠れマルコフモデルに基づく音声認識結果からの固有表現抽出. 言語処理学会 第 9 回年次大会，pp.533-536, 2003.3.
- 廣嶋 伸章, 長谷川 隆明, 山崎 毅文. 統計的手法に基づく Web ページからのヘッドライン生成. 情報処理学会研究報告「自然言語処理（NL)」 Vol.2002, No.44, pp.45-50, 2002.5.
- Takaaki Hasegawa, Takefumi Yamazaki, Yoshihiko Hayashi. SummaryBIFF: An E-mail Summarizer for Mobile Phones. Proceedings of the Sixth Natural Language Processing Pacific Rim Symposium (NLPRS2001), pp.761-762, 2001.11.
- 長谷川 隆明，小原 永. 単語情報を用いたパターンマッチングに基づくテキストからのモダリティ情報の抽出. 第 12 回 AI シンポジウム，pp.19-24, 2000.12.
- 長谷川 隆明. 送受信履歴と情報抽出に基づく電子メールの個人適応型ランキング. 情報処理学会研究報告「自然言語処理（NL)」 Vol.1999, No.62, pp.17-24, 1999.7.
- 長谷川 隆明，高木 伸一郎. 情報抽出とユーザの行動履歴に基づく電子メールのランキング. 情報処理学会第 56 回全国大会前期 Vol.2, pp.259-260, 1998.3.

- 長谷川 隆明, 高木 伸一郎. 電子メールコミュニケーションにおけるスケジュール情報抽出. 情報処理学会研究報告「自然言語処理 (NL)」 Vol.1998, No.123. pp.73-80, 1998.1.

- 長谷川 隆明, 中野 有紀子, 加藤 恒昭. 反射と熟考の相互作用に基づく協調的対話モデル. 情報処理学会研究報告「自然言語処理（NL）」 Vol.1996, No.65, pp.127-134, 1996.7.

- 加藤 恒昭, 中野 有紀子, 中嶋 秀治, 長谷川 隆明. 対話的マルチモーダル説明とその時間的協調. 情報処理学会研究報告「自然言語処理（NL)」 Vol.1996, No.65, pp. 135-142, 1996.7.

- 長谷川 隆明, 中野 有紀子, 加藤 恒昭. 反射と熟考の相互作用に基づく協調的対話モデル. 情報処理学会第 52 回全国大会前期 Vol.2, pp.411-412, 1996.3.

- 長谷川 隆明, 栗原 聡, 岡田 美智男, 土居 範久. 創発的な振舞いに基づく対話過程のモデル化. 情報処理学会第 50 回全国大会前期 Vol.3, pp.87-88, 1995.3.

- 長谷川 隆明, 高田 真吾, 土居 範久. 係り受けによる曖昧性の解消を支援する推敲システム. 情報処理学会第 45 回全国大会後期 Vol.3, pp.153-154, 1992.9.

特許

- 長谷川 隆明. 言い換え表現獲得システム、言い換え表現獲得方法及び言い換え表現獲得プログラム. 特許第 4252038 号.

- 長谷川 隆明, 林 良彦. 認識誤り訂正方法、装置、およびプログラム. 特許第 4171323 号.

- 長谷川 隆明, 林 良彦. クラス同定モデル生成方法、装置、およびプログラム、クラス同定方法、装置、およびプログラム. 特許第 4008344 号.

- 廣嶋 伸章, 長谷川 隆明. テキスト要約方法、装置、およびテキスト要約プログラム. 特許第 3790187 号.
- 廣嶋 伸章, 長谷川 隆明. キーワード決定方法、装置、プログラム、および記録媒体. 特許第 3787310 号.
- 長谷川 隆明, 高木 伸一郎. 文書構造解析方法及び装置及び文書構造解析プログラムを格納した記憶媒体. 特許第 3767180 号.
- 長谷川 隆明, 高木 伸一郎. 文書優先度付与方法及び装置及び文書優先度付与プログラムを格納した記憶媒体. 特許第 3740826 号.
- 長谷川 隆明. 文書表示方法及び文書表示プログラムを格納した記憶媒体. 特許第 3724270 号.
- 長谷川 隆明. 返信送出方法及びシステム及び返信送出プログラムを格納した記憶媒体. 特許第 3716548 号.
- 長谷川 隆明, 小原 永. 電子メール着信通知システム，電子メール着信通知方法，電子メール着信通知プログラムおよびそのプログラム記録媒体. 特許第 3697174 号.
- 長谷川 隆明, 小原 永. 電子化文書転送装置. 特許第 3690959 号.
- 長谷川 隆明, 高木 伸一郎. 日時表現正規化装置及び日時表現正規化プログラムを記録した記録媒体. 特許第 3628160 号.
- 長谷川 隆明. 話題別関心度計算方法及び装置及び話題別関心度計算プログラムを格納した記憶媒体. 特許第 3622602 号.
- 長谷川 隆明, 高木 伸一郎. 情報抽出方法、情報抽出装置及び情報抽出プログラムを記録した記録媒体. 特許第 3574551 号.

- 長谷川 隆明. 文書分類方法、文書分類装置、および文書分類プログラムを記録した記録媒体. 特許第 3471253 号.
- 長谷川 隆明, 松岡 浩司, 高木 伸一郎. 電子メール案内方法及び装置及び電子メール案内プログラムを記録した記録媒体. 特許第 3449893 号.
- 長谷川 隆明, 小原 永. 感情情報抽出方法および感情情報抽出プログラムの計算機読み取り可能な記録媒体. 特許第 3372532 号.

参考文献

[1] Eugene Agichtein and Luis Gravano. Snowball: Extracting relations from large plain-text collections. In *Proc. of the 5th ACM International Conference on Digital Libraries (ACM DL'00)*, pp. 85–94, 2000.

[2] 浅野久子, 加藤恒明, 高木伸一郎. Signature の局所的パターンマッチによる電子メールからの送信元住所録情報抽出とそれを用いた住所録管理システム. 情報処理学会論文誌, Vol. 39, No. 7, pp. 2196–2206, 1998.

[3] M. Banko, V. Mittal, and M. Witbrock. Headline generation based on statistical translation. In *Proc. of the 38th Annual Meeting of the Association of the Computational Linguistics (ACL-2000)*, pp. 318–325, 2000.

[4] M. Banko, V. Mittal, and M. Witbrock. Headline generation based on statistical translation. In *Proc. of the 38th Annual Meeting of the Association for Computational Linguistics (ACL-2000)*, pp. 318–325, 2000.

[5] A. Berger and V. Mittal. Ocelot: A system for summarizing web pages. In *Proc. of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR-2000)*, pp. 144–151, 2000.

[6] A. Berger and V. Mittal. Query-relevant summarization using faqs. In *Proc. of the 38th Annual Meeting of the Association for Computational Linguistics (ACL-2000)*, pp. 294–301, 2000.

[7] Sergey Brin. Extracting patterns and relations from world wide web. In *Proc. of WebDB Workshop at 6th International Conference on Extending Database Technology (WebDB'98)*, pp. 172–183, 1998.

[8] J. Carbonell and J. Goldstein. The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *Proc. of the 21th Annual International ACM/SIGIR Conference on Research and Development in Information Retrieval*, pp. 335–336, 1998.

[9] Giuseppe Carenini and Jackie Chi Kit Cheung. Extractive vs. nlg-based abstractive summarization of evaluative text: The effect of corpus controversiality. In *Proc. of Fifth Internationnal Natural Language Generation Conference (INLG-08)*, pp. 33–41, 2008.

[10] S. Corston-Oliver. Text compaction for display on very small screens. In *Proc. of the Workshop on Automatic Summarization at the 2nd Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL-2001)*, pp. 89–98, 2001.

[11] H. T. Dang. Duc 2005: Evaluation of question-focused summarization systems. In *Proc. of the Workshop on Task-Focused Summarization and Question Answering*, pp. 48–55, 2006.

[12] Defense Advanced Research Projects Agency. *Proceedings of the Sixth Message Understanding Conference (MUC-6)*. Morgan Kaufmann Publishers, Inc., 1995.

[13] T. Dunning. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, Vol. 19, No. 1, pp. 61–74, 1993.

[14] T. Fuchi and S. Takagi. Japanese morphological analyzer using word co-occurrence –jtag–. In *Proc. of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics (COLING-ACL'98)*, pp. 409–413, 1998.

[15] 福本淳一, 桝井文人. メールロボ：インターネットメールからの情報抽出. 沖電気研究開発第 181 号, Vol. 66, No. 2, pp. 55–58, 1999.

[16] 畑山満美子, 松尾義博, 白井諭. 重要語句抽出による新聞記事自動要約. 自然言語処理, Vol. 9, No. 4, pp. 55–70, 2002.

[17] 平尾努, 鈴木潤, 磯崎秀樹. 軽量な文短縮手法. 言語処理学会第 14 回年次大会講演論文集 (NLP2008), pp. 484–487, 2008.

[18] 堀智織, 古井貞熙. 講演音声の自動要約の試み. 話し言葉の科学と工学ワークショップ, pp. 165–171, 2001.

[19] 伊知地宏, 倉部淳. メイルを用いたソフトウェア開発を支援するツール. 情報処理振興事業協会 (ipa) 平成 13 年度成果報告集, 情報処理振興事業協会 (IPA), http://www.ipa.go.jp/NBP/13nendo/reports/explorafft/mailide/mailide.pdf, 2001.

[20] 池原悟, 宮崎正弘, 白井諭, 横尾昭男, 中岩浩巳, 小倉健太郎, 大山芳史, 林良彦（編）. 日本語語彙体系. 岩波書店, 1997. NTT コミュニケーション科学研究所 監修.

[21] 今村賢治. 系列ラベリングによる準話し言葉の日本語係り受け解析. 言語処理学会第 13 回年次大会講演論文集 (NLP2007), pp. 518–521, 2007.

[22] 稲垣博人, 早川和宏, 井上孝史, 田中一男. モバイル端末の表示特性に応じたメッセージ要約方式の提案. 情報処理学会第 56 回全国大会講演論文集 (分冊 2), pp. 255–256, 1998.

[23] D. Lam, S. L. Rohall, C. Schmandt, and M. Stern. Exploiting e-mail structure to improve summarization. IBM Watson Research Center Technical Report 02-02, IBM Watson Research Center, Cambridge, MA, USA, 2002.

[24] Dekang Lin and Patrick Pantel. Dirt - discovery of inference rules from text. In *Proc. of the 7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2001)*, pp. 323–328, 2001.

[25] H. Luhn. The automatic creation of literature abstracts. *IBM Journal of Research and Development*, Vol. 2, No. (2), pp. 159–165, 1958.

[26] 望月源, 奥村学. 読みやすさの向上と冗長性の排除を考慮した重要箇所抽出型要約. 情報処理学会自然言語処理研究会報告 139-3, pp. 17–24, 2000.

[27] 森辰則. 情報検索表示向け文書要約における情報利得比に基づく語の重要度計算. 自然言語処理, Vol. 9, No. 4, pp. 3–32, 2002.

[28] 村越広亨, 山見太郎, 島津明, 落水浩一郎. 電子メールを利用した学習者間のコミュニケーション支援技術の開発. 教育システム情報学会誌, Vol. 18, 3・4, pp. 308–318, 2001.

[29] S. Muresan, E. Tzoukermann, and J. Klavans. Combining linguistics and machine learning techniques for email summarization. In *Proc. of the Fifth Workshop on Computational Language Learning (CoNLL-2001)*, pp. 152–159, 2001.

[30] 中川裕志, 渡部聡彦. 携帯端末向けコンテンツ変換と自然言語処理. 情報処理, Vol. 43, No. 12, pp. 1300–1304, 2002.

[31] 仲尾由雄. 見出しを利用した新聞・レポートからのダイジェスト情報 の抽出. 情報処理学会自然言語処理研究会報告 117-17, pp. 121–128, 1997.

[32] National Institute of Standards and Technology. Automatic Content Extraction. http://www.nist.gov/speech/tests/ace/index.htm, 2000.

[33] T. Nomoto. A generic sentence trimmer with crfs. In *Proc. of the 46th Annual Meeting of the Association for Computational Linguistics (ACL-2008)*, pp. 299–307, 2008.

[34] 奥村学, 難波英嗣. 知の科学 テキスト自動要約. オーム社, 2005. 人工知能学会 編集.

[35] 大森岳史, 増田英孝, 中川裕志. Web 新聞記事の要約とその携帯端末向け記事による評価. 情報処理学会 研究報告 NL153-1, pp. 1–8, 2003.

[36] Dragomir R. Radev and Kathleen R. McKeown. Generating natural language summaries from multiple on-line sources. *Computational Linguistics*, Vol. 24, No. 3, pp. 470–500, 1998.

[37] Deepak Ravichandran and Eduard Hovy. Learning surface text patterns for a question answering system. In *Proc. of the 40th Annual Meeting of the Association for Computational Linguistics (ACL-2002)*, pp. 41–47, 2002.

[38] 酒井浩之, 篠原直嗣, 増山繁, 山本和英. 連用修飾表現の省略可能性に関する知識の獲得. 自然言語処理, Vol. 9, No. 3, pp. 41–62, 2002.

[39] Satoshi Sekine. OAK System (English Sentence Analyzer). http://nlp.cs.nyu.edu/oak/, 2001.

[40] Satoshi Sekine, Kiyoshi Sudo, and Chikashi Nobata. Extended named entity hierarchy. In *Proc. of the Third International Conference on Language Resources and Evaluation (LREC-2002)*, pp. 1818–1824, 2002.

[41] 高見真也, 田中克己. ウェブ検索結果における検索目的に応じたスニペット生成. 情報処理学会論文誌, Vol. 49, No. 4, pp. 1648–1656, 2008.

[42] A. Tombros and M. Sanderson. Advantages of query biased summaries in information retrieval. In *Proc. of the 21st Annual International ACM/SIGIR Conference on Research and Development in Information Retrieval*, pp. 2–10, 1998.

[43] K. Yamagata, S. Fukutomi, K. Takagi, and K. Ozeki. Sentence compression using statistical information about dependency path length. In *Proc. of the 9th International Conference on Text, Speech and Dialogue, TSD 2006 (Lecture Notes in Computer Science / Lecture Notes in Artificial Intelligence, Vol. 4188)*, pp. 127–134, 2006.

[44] 山本和英, 池田諭史, 大橋一輝. 「新幹線要約」のための文末の整形. 自然言語処理, Vol. 12, No. 6, pp. 85–111, 2005.

[45] Dmitry Zelenko, Chinatsu Aone, and Anthony Richardella. Kernel methods for relation extraction. In *Proc. of the Conference on Empirical Methods in Natural Language Processing (EMNLP-2002)*, pp. 71–78, 2002.