WWW ページからの手順に関する箇条書きの抽出

武智峰樹[†,††] 徳永健伸[†††]
松本裕治[††] 田中穂積[†††]

要素技術としての文書分類は，質問応答や Web ナビゲーションにおける主要な構成要素である．特に表層的なテキストの特徴を主に利用する質問応答では，与えられた質問のタイプに応じて適切な回答候補を抽出できる分類エンジンが重要である．また Web ナビゲーションにおいては，従来の質問応答が扱ってこなかった質問も扱う必要があり，そのような質問に対しても適切な回答候補を選び出すための分類技術が求められる．本研究は，Web ナビゲーションが扱う質問のうち，特に手順に関する質問を取り上げ，その回答候補の分類に有効な特徴量を明らかにすることを目的とする．その試みとして Web ページにおいて HTML のリストタグが付与されたテキストを記事集合として，それを手順について書かれたテキストとそれ以外のテキストに分類するタスクを考える．検索エンジンを用いて箇条書きを収集し，機械学習の一手法である Support Vector Machine を用いた文書分類を行い，その結果の観察に基づいて手順について書かれた箇条書きの抽出に有効な特徴量を考察した．N-gram や語の頻度情報をベースにした手法により，コンピュータ分野に関しては 90%以上の精度で分類可能な特徴量の組合せを得た．

Extracting Lists of Procedural Expressions from Web Pages

Mineki Takechi,[†,††] Takenobu Tokunaga,[†††] Yuji Matsumoto[††]
and Hozumi Tanaka[†††]

Text categorization is an essential component to allow for efficient navigation techniques and to get query-relevant information on the Web. Especially in the context of Question-Answering, it requires the right features to categorize the documents and to allow for efficient knowledge acquisition according to the types of queries. In the queries addressed in such navigation, we focus on those asking for procedural knowledge and aim at clarifying the specification of the answers. To solve this problem we exploit procedural descriptions in the form of itemized expressions tagged with the HTML list tags. Applying Support Vector Machines to the set of list expressions gathered from WWW by a search engine, we examine the obtained model in order to to find the relevant features for the extraction of an answer that explains relevant procedures. By exploiting the features based on word frequencies, such as N-gram and the sequences of words, we obtained a feature set for a computer domain that can categorize more than 90% in recall and precision.

1. はじめに

要素技術としての文書分類は，近年質問応答や Web ナビゲーションという応用分野を得て，質問のタイプに応じた分類など，索引語の分野依存性に基づいた従来の方法論だけでは解決できない問題を扱っている．最近の質問応答の研究では，表層的なテキストの特徴を主に利用するアプローチが成功しており，その方法論において，ユーザから与えられる質問のタイプ判定の精度と，それに応じて回答候補を含むパッセージを適切に分類する精度が重要な課題となっている[12]．深い意味理解が困難な現段階においては，このような方向性は当分続くものと考えられ，質問のタイプに応じた文書分類の研究が現時点では重要である．

一方，Web ナビゲーションの研究に目を移すと，Google[4] などの検索エンジンにおけるページランキングや情報フィルタリングの研究，また検索結果をユーザに分かりやすく整理して提示するための自動分類や自動要約の研究[3] などがある．近年の質問応答の研究においても，新聞や論文を対象にした研究に加え，WWW を知識源として積極的に利用する試みが

† 富士通株式会社
FUJITSU LIMITED
†† 奈良先端科学技術大学院大学情報科学研究科
Graduate School of Information Science, Nara Institute of Science and Technology
††† 東京工業大学情報理工学研究科
Graduate School of Information Science and Engineering, Tokyo Institute of Technology

始まっている[14),15)]．本論文は，Web ナビゲーションのための質問応答において必要とされる文書分類を適用分野としている．

従来の質問応答の研究では，What，When など事実を問うタイプの質問が主に扱われてきた．しかし，Web 上の情報を用いて回答する質問応答では，ユーザに与えられる回答は不正確で不確かな情報を多数含む可能性がある．これによってユーザ自身が適切な回答を選択する必要が生じるため，そうしたユーザ行動を想定した質問のタイプについても考える必要がある．また，WWW の利用者の拡大にともなって多様なニーズが生まれており，事実を問う質問以外の質問タイプについても扱う必要がある．こうした研究は現在その端緒にあり，研究のためのコーパスをどのように集めるかという点も重要な課題である．本論文は上記のような質問タイプのうち，**手順を尋ねる質問**を取り上げ，回答の抽出・分類のために有効な特徴量を明らかにすることによって，このような問題の解決に寄与することを研究の動機としている．

手順に関する質問応答において，WWW を情報源として用いた先行研究が少ないために，(1) Web ページにおいて手順がどのような形式によって示されているか，(2) それらの表現をどのように抽出するか，(3) 抽出した表現をどのように回答として出力すべきか，についてほとんど分かっておらず，現時点では一般的な手法を考えることが難しい．

一方，WWW 上には多数の QA 記事や QA 集が存在し，そこには多くの**箇条書き**が含まれていることが知られている[10)]．箇条書きは，要点を端的に示すために人によって行われた一種の要約と見なせるため，重要な情報が読みやすいかたちで得られると期待される．我々はまず Web ページにおいて箇条書きされたテキストに着目し，箇条書きされた記事を回答候補として抽出することを考えた．箇条書きされたテキストにフォーカスすることによって他の形式で書かれた回答候補を抽出することが不可能になるが，一方で得られた箇条書きの集合が，回答候補を比較的高い比率で含んでいることが期待できる．また本論文の動機は，適切かつ信頼性の高い情報へのユーザナビゲーションであって，条件を満たす文書を漏れなく検索することではない．よって，回答するテキストを箇条書きに限定することは本論文の目的に反しない．本論文では，上記のような質問応答を行うための最初のステップとして，与えられた箇条書きを**手順タイプと非手順タイプに分類するタスク**について議論する．ここで手順タイプの箇条書きとは，何らかの具体的な手順について書かれた箇条書きのことを指す．

我々は日本語を対象に，Web 検索エンジンを用いてキーワードとして「手順」および「方法」を含む Web ページを検索し，得られた Web ページから HTML（Hyper Text Markup Language）のリストタグが付与された箇条書きを収集した．次に収集した箇条書きを人手によって分類し，その一部を用いて手順タイプの箇条書きに特徴的な表現を調べた．その結果に基づき手順タイプの箇条書きの記述スタイルを特徴づけると考えられる語の出現頻度を中心とした特徴量と，テキスト中で離れた位置にある語の出現順序を考慮して，機械学習の 1 つである SVM（Support Vector Machine）に用いて学習を行い，得られたモデルを検討することによって分類に有効な特徴量を考察した．この結果コンピュータ分野に関しては 90%以上の精度で分類可能な特徴量の組合せを得た．また，手順に関する箇条書きの分類における特徴量としての機能語の有効性について，新たな知見を得た．

次章では本論文に関連した先行研究をいくつか紹介する．3 章では，Web ページにおける手順の説明を含むテキストの特徴について詳しく解説する．4 章では実際に実験に用いた特徴量と機械学習手法について簡単に説明する．5 章では箇条書きを手順タイプと非手順タイプに分類する実験の結果とその考察を示し，6 章において本論文のまとめを行う．

2. 関 連 研 究

手順に関する質問応答は，エキスパートシステムの研究において早い時期から扱われてきた[2)]．しかしながらエキスパートシステムにおいては，限られた分野の専門的知識を，人手によって形式的な知識表現に変換して知識ベース化する必要がある．これに対し大規模で変化しやすい性質を持つ WWW からの情報を用いる質問応答では，人手によって知識ベースを構築・保守することが難しく，自然言語とマークアップ言語で記述されたより低い構造化レベルの知識ベースを前提とした技術が要求される．

近年，TREC や NTCIR などの評価型ワークショップを通じて，Web 検索および質問応答の研究発表がさかんに行われている[5)~7),18)]．これらの研究において成功を収めているアプローチは，自然言語の深い理解を求めず，主にテキストの表層的な手がかりを多数組み合わせて利用することによって，検索対象文書から質問に対する回答を抽出するものである．このような質問応答においては，質問のタイプをできるだけ細かく判定し，それに応じて異なる処理を行うことによっ

て精度を高める方策がとられる[12]．また，検索対象文書から回答候補を含むパッセージを検索する段階では，ユーザの質問が定義を求めるものか事実を求めるものかなど質問のタイプによって利用できる言語表現が異なるため，質問のタイプの判定とそれを考慮したパッセージの検索精度の向上が求められている[12]．これは，ドメインやジャンルに基づくカテゴリに対する従来の文書分類とは異なる観点からの研究が必要であることを意味する．

従来質問応答の研究において扱われてきた質問のタイプは，主に事実や定義を尋ねるものである．Web ページを対象とした質問応答に本格的に取り組んだ事例としては，定義タイプの質問を扱った藤井ら[23),24)]の研究がある．また，WWW 上にすでに存在する QA 集などに検索対象を限定し，そのなかから回答を探すというアプローチも存在する[10),29)]．浜田ら[29)]は，Web ページにおける料理レシピ集から調理方法の説明に用いられる用語辞書を手作業で作成し，辞書と機能語を中心としたルールを用いてレシピにおける料理手順の構造解析を行っている．他にユーザクエリに適合した複数の Web 文書の要約を自動的に作成する研究[3),8)]，公的機関のヘルプデスクに寄せられる質問に対する自動応答の研究などがある[9)]．

上記のような先行研究にもかかわらず，手順を尋ねる質問を扱う質問応答において WWW を用いて回答する研究はほとんど知られていない．浜田らの研究は我々の知る限り唯一の類似研究であるが，料理レシピだけを対象にしている点，Web ページからの料理レシピの抽出を人手によって行っている点など，様々なドメインの手順に適応可能で，頻繁に更新される Web ページから効率良く手順についての説明を収集することを目的とする本論文とは異なる．

また文書分類の研究においては，従来から相互情報量[19)]，ルールベース[20)]など様々な特徴量選択が行われてきた．決定木[11)]をはじめ多くの学習アルゴリズムが試され，それらの比較検討も進んでいる[19)]．日本語を対象とする文書分類の研究においても，ブースティングや SVM（Support Vector Machine）を用いた研究[25),26)]，文書より小さなパッセージを分類の対象とする研究[21)]など多くの報告がある．

従来の文書分類の問題では，主にドメインやジャンルに基づいたカテゴリへ文書を分類する課題が扱われる．分類に利用する文書の属性は，ジャンルやドメインの特徴を反映した名詞や動詞などの内容語であり，機能語や頻出する形式的表現は不要語として排除されることが多い．一方で，文書の著者推定など，ドメインとは異なる観点に基づいた文書分類では，助詞や助動詞，あるいはその組合せなど内容語以外の文書の属性が分類に寄与する可能性が示唆されている[16),22),27)]．

しかし，手順に関する箇条書きを抽出するタスクは，一般的な文書分類や，著者推定など他の観点からの文書分類とは問題が異なっており，同様の手法が有効であるかは自明ではない．また，著者推定における先行研究では機能語の役割が指摘されているものの，それらを積極的に文書分類に利用する試みは少ない．本論文は，ドメインに依存しない質問応答のための要素技術として文書分類を研究する立場から，手順に関する箇条書きの抽出に必要な特徴量を明らかにするとともに，特徴量としての機能語の利用可能性に着目するものである．従来の文書分類の研究とはタスクの性質および特徴量の選択基準において異なっている．

3. 箇条書きを用いた手順に関する質問応答

3.1 WWW を用いた手順に関する質問応答

1 つの例として，「Linux のインストール方法を知りたい」というユーザを考える．このようなユーザが Web 検索エンジンを用いて具体的な手順が書かれた文書を探すことはインターネットが普及した現在では自然な探索行動であると考えられる．

現在の Web 検索エンジンは手順だけを優先的に集める機能を持たないため，ユーザは思いつく限りの単語を与えて試行錯誤するほかない．上記の質問の場合には，分野に関する単語（「Linux」「インストール」など），手順に関する単語（「手順」「方法」「やり方」など）を与えるとの予測が立つ．しかし，従来の検索エンジンの返す検索結果は具体的な手順を含んでおらず，十分な回答になっていないことも多い．たとえば，単なるリンク集や，手順について書かれてはいるが選択肢を列挙するにとどまる場合などである．本論文では，このような場合に具体的な手順だけを回答として返すことを目指す．

3.2 箇条書きによる回答

手順を説明するテキストは検索対象のテキストのなかで，連続した一部分を占めているとは限らない．また，Web ページにおいて手順がどのような形式によって示されているのかも十分には分かっていない．

一方，Web ページにおける QA 記事や WWW 上に数多く存在する QA 集には多くの箇条書きが含まれていることが知られている．箇条書きには重要な情報が読みやすく提示されていると考えられるため，まず箇条書きされた記事を回答候補として抽出することを考えた．箇条書きされたテキストにフォーカスするこ

表 1 収集した Web ページ（URL）数
Table 1 Result of collection of web pages.

キーワード	検索結果	収集したページ	有効ページ
手順	748	3,713	629
方法	916	5,998	929

とにより他の形式で書かれた回答候補を抽出することが不可能となるが，一方で得られた箇条書きの集合に多くの回答候補が含まれていることが期待できる．

これに加えて Web ページの箇条書きには機械的な処理を行ううえで以下のような利点がある．

- 箇条書きの前後にタイトルなどの手がかりがある．
- <OL>，<UL>など，HTML のリストタグを利用して比較的容易に抽出が可能．

本論文では，記事の収集において上記のような利点を持つ Web ページ上の箇条書きの集合を，手順について書かれた箇条書（**手順タイプ**）とそれ以外の箇条書き（**非手順タイプ**）に分類するタスクを考える．手順タイプの箇条書きと非手順タイプの箇条書きを自動分類することによって，手順タイプの箇条書きの抽出に役立つ特徴量を明らかにすることを目的とする．

3.3 WWW からの箇条書きの収集

Web ページにおける箇条書きの特徴を調べるため，Web 検索エンジンを用いて以下の Step 1〜4 のようにして箇条書きを収集した．収集の結果を**表 1** に示す．

Step 1 Google に対してキーワードとして「手順」および「方法」をそれぞれ別々に与え，収集の起点となる URL を得た（表 1：検索結果）．この際，1 回の検索で 1 度に取得できた URL のみを，手作業で保存して検索結果とした．

Step 2 得られた URL からリンクされているページを 2 回まで再帰的に探索し，HTML ファイルを収集した（表 1：収集したページ）．

Step 3 Step 2 で得た HTML ファイルから<OL>または<UL>タグで囲まれたパッセージを取り出し，リストタグが階層を持つ場合には各階層を 1 つの箇条書きとして分割した．

Step 4 2 つ以上の項目を持つ箇条書きについて，1 つの箇条書きを 1 記事とする記事セットを作成した（表 1：有効ページ）．

このようにして得られた記事セットを人手によって箇条書き単位に手順タイプと非手順タイプに分類した．分類にあたり，手順タイプの箇条書きの定義を，便宜的に,「何らかの目的を達成するまでの複数の行為または動作を，それぞれ項目に記述して実行すべき順序に並べたもの」として与えた．

ここで項目とは，箇条書きにおける 1 つの条項のことで，記号から始まり 1 つまたは複数の文からなる．また，分類にあたっては以下の制約を与えた．

- 項目の半分以上において，何らかの行為・動作が述べられているものを手順タイプとする．
- 分類は箇条書き部分だけを用いて行い，前後に現れるテキストは判断に用いない．

表 2 各キーワードにより抽出した箇条書きのドメインとタイプ
Table 2 Domains and types of extracted lists.

	手順 「手順」/「方法」	非手順 「手順」/「方法」	総数
コンピュータ	295/263	724/942	2,224
その他	64/99	476/1,257	1,896
総数	721	3,399	4,120

分類作業は箇条書き部分だけを抜き出した記事セットを 2 つ用意して，それぞれについて分類を行った．記事セットの 1 つは Web 検索エンジンに与えるキーワードとして「手順」を用いて収集したものであり，もう一方は「方法」を用いて収集したものである．分類作業は，簡単なタグ付けエディタを作成して，手順タイプの箇条書き全体を単純にタグで囲んだ．また，上記の 2 つの記事セットそれぞれについて，箇条書きが含まれていた Web ページのドメインに基づいた分類も行った．コンピュータ分野の記事が多数を占めたため，コンピュータ分野以外の記事はその他分野として 1 つにまとめた．その他分野の記事は，教育，医療，冠婚葬祭など複数のドメインの記事を含む．その他分野の Web ページに含まれる箇条書きでも，ホームページにおけるサービス画面の操作方法などを説明したものはコンピュータ分野に分類した．以上の箇条書きについての人手による分類結果を，検索エンジンに与えたキーワード別に**表 2** に示す．表中で '/' の左右の数値は，それぞれキーワードに「手順」を用いた場合と「方法」を用いた場合に対応している．

検索エンジンを用いた記事の収集においては，特にドメインの指定をしていないにもかかわらず，実際に収集できた箇条書きの半数以上がコンピュータ分野のものであったことが分かる．

キーワードに「手順」を用いた場合の記事セットについては，1 つの記事につき 2 人の作業者が分類を行い，Kappa 統計値を指標として記事単位で分類の揺れを調べた．Kappa 統計値は，談話構造解析や要約評価の先行研究[30] において用いられてきたもので，観測一致率 P(A) から偶然一致率 P(E) を除いた式 (1) によって定義される．ここで観測一致率とは，2 人の作業者の分類結果が一致した記事数が，分類対象の記事の総数に対して占める割合である．本研究において

表 3 人手による分類の一致率
Table 3 Result of categorization by human judgment.

	全分野	コンピュータ分野	その他分野
観測一致率	0.94	0.95	0.93
偶然一致率	0.65	0.59	0.80
Kappa 値	0.83	0.87	0.66

は，ある記事が手順タイプであるか否かについて，2 人の作業者の判断が一致した場合に分類結果が一致したと見なす．一方偶然一致率は，各作業者がそれぞれの分類カテゴリを選んだ回数を分類対象の記事の総数で割り，カテゴリごとに 2 人の作業者の値を掛け合わせてその和をとったものである．これは，各作業者がそれぞれの分類カテゴリを選択した割合を，作業者が任意にカテゴリを選んだ場合の経験的確率と見なし，2 人の作業者の判断が偶然に一致すると期待される割合を求めたものである．

$$Kappa = \frac{P(A) - P(E)}{1 - P(E)} \quad (1)$$

人手による分類結果について Kappa 値を計算したところ**表 3** に示す結果が得られた．この指標においては，Kappa 値が，1）0.41〜0.60 の場合に適度な一致（moderate），2）0.61〜0.80 の場合にかなりの一致（substantial），3）0.81〜1.00 場合には完全に近い一致（near perfect）という基準が用いられている．この基準を用いれば，コンピュータ分野の分類に関してはその結果について完全に近い一致が得られている．一方その他の分野については，コンピュータ分野には及ばないものの Kappa 値はかなりの一致の範囲に入っており，様々なドメインの箇条書きについて，手順の説明であるか否かの判断が比較的高い精度で行える可能性を示している．

3.4 箇条書きにおける手順の表現

人手により分類した記事セットを調べた結果，箇条書きが含まれる Web ページのドメインにかかわらず，箇条書きにおける手順の表現として次の特徴が見られた．

- 項目の 1 文目に行為や動作が現れることが多い．
- 1 文目においては，(1) 文末が主に行為や動作を表す名詞句で終わるタイプと，(2) 文が使われるタイプがある．
  名詞句タイプの例）
  1. ダウンロード
  2. インストール
  3. 設定
- 名詞句タイプの場合，サ変名詞で終わることが多い．
- ガ格，提題助詞，否定辞が少ない，ヲ格が多い，文末や読点前に繰り返し同じ表現が多用される．

ドメインにかかわらず前記のような機能語を中心とした特徴があるとすれば，特定のドメインから得られた分類に有効な特徴量を用いて他のドメインについても手順・非手順の分類が可能となる．文末表現や，読点前に繰り返し使用される表現[27] はテキストの記述スタイルを特徴付けるとされており，手順タイプの箇条書きには特定のスタイルがあるのではないかと考えた．

テキストの記述スタイルを特徴付けるために，従来文学作品を対象とした研究では単語 N-gram や品詞の頻度分布，頻繁に使用される文字の種別などの情報が用いられてきた[22]．またこれに加えて，Web 文書を対象にした最近の研究では，頻出する語の系列を特徴量として用いるものも見られる[16]．我々は，手順タイプと非手順タイプの分類に有効な特徴量を検討するにあたり，手順タイプの箇条書きの観察に基づく特徴量に加えて，テキストの記述スタイルの特徴を獲得するために先行研究において用いられてきた特徴量についても利用した．

4. 分類精度向上のための特徴量

4.1 ベースライン

記述スタイルを定量的に特徴付けるために，(1) 文字・形態素 N-gram による特徴量，および，(2) 形態素 N-gram に加えてシーケンシャルパターンマイニングの一手法である PrefixSpan [13] を用いて取り出した文単位での語の頻出パターンを組合せた特徴量セット，の 2 系統の特徴量を採用した．シーケンシャルパターンマイニングを文に対して用いることにより，文内で離れた位置関係にある語の間の関係を，その出現順序も含めて分類に利用することができる．これらを機械学習の一手法である SVM（Support Vector Machine）[17] に与えることにより学習および文書分類実験を行った．SVM は，高次元の特徴空間に対しても高い汎化能力を持つ二値線形分類器である．SVM の理論的背景については文献 17) を参照いただきたい．手順タイプの判定においては，どのような特徴量が有効であるか明らかになっていないため，多くの特徴量を用いて有効な特徴量の組合せを探っていく必要がある．SVM が示す，事例数に比べて高い次元の特徴空間でも過学習が起こりにくい性質は，他の学習器との比較において，高次元の特徴空間から有効な特徴量を絞り込んでいく本研究に適していると考えた．

4.2 シーケンシャルパターンマイニング

シーケンシャルパターンマイニングは次のように

定義される処理である．今，**アイテム**（リテラル）の集合 I= $i_1, i_2, \ldots, i_n$ を考える．集合 I の空でない部分集合を**エレメント**という．また，ある閾値 $\xi > 0$ が与えられたとき，集合 I において ξ 回以上現れるアイテムを**頻出アイテム**という．エレメントの順序列を**シーケンス**という．さらに，シーケンス $\alpha =< a_1, a_2, \ldots, a_n >$ とシーケンス $\beta =< b_1, b_2, \ldots, b_m >$ に対して $a_1 \subseteq b_{j1}, a_2 \subseteq b_{j2}, \ldots, a_n \subseteq b_{jn}$ なる整数 $1 < j_1 < j_2 < \ldots < j_n < m$ があるとき，α を β の**サブシーケンス**といい $\alpha \subseteq \beta$ と書く．シーケンス ID sid とシーケンス s の組<sid,s>の集合を**シーケンスデータベース**と呼ぶ．閾値 ξ（最小サポート値と呼ぶ）以上の個数の<sid,s>に含まれているシーケンスをシーケンシャルデータベースにおける**シーケンシャルパターン**という．シーケンシャルパターンマイニングとは，シーケンスデータベース S と最小サポート値 ξ が与えられたときに，S におけるシーケンシャルパターンをすべて見つけることである．代表的なシーケンシャルパターンマイニングの手法として，Apriori アルゴリズム[1]，PrefixSpan[13] などがある．

Apriori アルゴリズムは，シーケンシャルパターンマイニングの手法のなかでも最も広く使われ，今までに多くの拡張手法が提案されている．しかしながら Apriori アルゴリズムは，最小サポート値の条件を満たすアイテム集合を計算するたびに，シーケンスデータベース全体を走査するため，計算コストの面で改善の余地がある．PrefixSpan アルゴリズムは，射影と呼ばれる操作によって走査の繰返しを頻出アイテムだけで構成されるサブシーケンスの集合に限定することよって，Apriori アルゴリズムに比べてマイニングに必要な計算量を減らすことに成功している．PrefixSpan アルゴリズムの詳細については文献 13) を参照いただきたい．

4.3 項目を超えて現れるパターン

我々は，記述スタイルに基づく文書分類において従来から用いられてきた特徴量に加えて，箇条書きの項目を超えて現れるパターンについても利用した．すでに指摘したように，箇条書きには特定の助詞の省略や多用，文末表現の連続が見られるため，複数の項目にわたって繰り返し使用されるパターンは，分類精度の向上に役立つことが期待される．箇条書き全体を 1 つの単位として以下のような手順で文字列を作成して PrefixSpan に与え，項目をまたがって繰り返し現れる語の出現パターンを獲得して特徴量とした．

表 4 パターンマイニングの前処理で使用したタグセット
Table 4 Tag set for document annotation.

	タグ名	付与する単位
文書タグ	dv	箇条書き
	p	項目
	su	文
品詞タグ	np	名詞☆ 接頭詞
	snp	名詞-サ変接続
	vp	動詞
	adp	助詞☆☆ 副詞 連体詞 接続詞
	ajp	形容詞
	aup	助詞-終助詞 助動詞 接尾語
	ij	感動詞
	seg	その他

Step 1 茶筌[28] を用いて形態素解析を行い，その結果に基づき**表 4** に示す品詞タグを箇条書きに付与する．文書タグは，記事セットを作成する段階で付与した．表 4 における品詞名は茶筌の品詞体系に基づく．タグの形式は，XML（eXtensible Markup Language）の仕様に従う．

Step 2 各項目からそれぞれ n 文を取り出し項目タグを含めて 1 記事を 1 つの文字列とする．

形態素および文書タグを 1 アイテムとする．**図 1** に箇条書きに対するタグ付け例を示す．<p>タグが付与された項目の先頭における番号（<seg>1</seg>など）は，Web ページから箇条書きを抽出する段階で，<LI>タグを適当に置き換えた結果である．<UL>タグによって囲まれた場合は・（中黒）を，<OL>タグによって囲まれている場合は「1.」,「2.」のように昇順に番号を付与した．この処理はデータの視認性の向上と処理の簡略化のために行ったものであり，本質的な処理ではない．**図 2** は，図 1 の箇条書きからシーケンシャルパターンを取り出すための文字列の作成例である．図 2 では各項目の第 1 文目だけを使用する例を示した（Step 2 において n=1 の場合）．<p>タグは 1 つのアイテムとして扱い，それ以外は開始タグと終了タグで囲まれた文字列をタグを含めて 1 つのアイテムとした．なお，文字 N-gram および形態素 N-gram を作成する際にはこれらのタグは使用しない．ただし，形態素解析の結果異なる品詞タグが割り当てられた場合に

☆ サ変接続以外
☆☆ 終助詞以外

```
<p><su><seg>1</seg><seg>.</seg><vp>必要</vp><aup>な</aup><np>内容</np>
        <adp>を</adp><snp>入力</snp><vp>し</vp><aup>ます</aup>
        <seg>。</seg></su>
    <su><seg>「</seg><np>お</np><vp>読み</vp><vp>ください</vp>
        <seg>」</seg><adp>を</adp><vp>見る</vp><adp>と</adp>
        <ajp>よい</ajp><aup>でしょ</aup><aup>う</aup><seg>。</seg></su>
</p>
<p><su><seg>2</seg><seg>.</seg><seg>「</seg><np>アプリケーション</np>
        <seg>」</seg><adp>の</adp><np>ラジオ</np><np>ボタン</np>
        <adp>を</adp><vp>押し</vp><aup>ます</aup><seg>。</seg></su>
</p>
```

図 1 品詞タグを付与した後の箇条書きの例
Fig. 1 An example of tagged list by part-of-speech tagger.

```
<p> <vp>必要</vp> <aup>な</aup> <np>内容</np> <adp>を</adp> <snp>入
力</snp> <vp>し</vp> <aup>ます</aup> <p> <np>アプリケーション</np>
<adp>の</adp> <np>ラジオ</np> <np>ボタン</np> <adp>を</adp> <vp>押し
</vp> <aup>ます</aup>
```

図 2 PrefixSpan に渡す文字列の例（各項目の 1 文目を用いる場合）
Fig. 2 Transaction to PrefixSpan corresponding to Fig. 1.

は，同じ表記の形態素であっても異なる特徴量として区別する．

5. 分類実験

5.1 実験設定

まずコンピュータ分野の箇条書きだけを用いて，手順タイプと非手順タイプの文書分類を行った．評価には 5 分割交差検定を用いた．また，コンピュータ分野のデータセットを学習データとし，その他の分野のデータセットを評価データとして，オープンドメインでの文書分類を行った．

実験に用いた箇条書きは，リストタグに囲まれた部分のテキストのみから構成した．Web ページにおいてリストタグの前後にはそれが置かれた文脈を示す手がかりが含まれている場合があるが，本研究では箇条書き部分の特徴をより詳しく知るためにこれらの手がかりを意図的に排除した．記事セットにおける箇条書きの概要を**表 5** に示す．手順/非手順の別およびドメイン別に 4 グループに分け，それぞれ記事数（箇条書きの数）とそれらが含む 1 記事あたりの項目数，1 項目あたりの文数，1 文あたりの文字数を示した．各欄において '/' の左側が平均，右側が偏差である．項目数と 1 文あたりの文字数において，若干の異なりが見られる．SVM および PrefixSpan の実装は工藤拓氏による Tiny-SVM☆および PrefixSpan☆☆を使用した．SVM のカーネル関数には 2 次の多項式関数を使用し，PrefixSpan の最小サポート値には，実験の条件を満

☆ http://cl.aist-nara.ac.jp/~taku-ku/software/TinySVM/
☆☆ http://cl.aist-nara.ac.jp/~taku-ku/software/prefixspan/

表 5 記事セットの概要
Table 5 Statistics of dataset.

	手順	非手順	コンピュータ	その他
記事数	721	3,399	2,224	1,896
項目数	4.6/2.8	4.9/5.7	4.8/6.1	4.9/4.4
文数	1.8/1.7	1.3/0.9	1.5/1.1	1.3/1.1
文字数	40.3/48.6	32.6/42.4	35.6/40.1	32.6/48.2

たすパターンの件数が得られるように経験的な値を設定した．シーケンシャルパターンについては，文単位のマイニングと箇条書き全体を 1 つの文字列としたマイニングの両方を行った．それぞれで得られたパターンについて出現頻度順に 1,000〜5,000 個までを使用し，頻度において 1 位から 5000 位までのパターンを 1,000 個ずつ使って異なる頻度レンジのパターンを使った評価を行った．シーケンシャルパターンマイニングによって得られたパターンの利用に際しては，頻度とは異なる基準を用いて学習器への入力をあらかじめ絞り込む立場と，そうした事前の絞り込みを行わない立場がある．本論文では，本実験固有の事情を考慮して，(1) 手順タイプの箇条書きの分類においては有効な特徴量が明らかになっていないため，文書分類の先行研究などで利用されている特徴量選択を行ってしまうと，未知の性質に気付きにくい可能性がある，(2) SVM を用いた文書分類においては，相互情報量などのフィルタリングがあまり効果がなかったとの報告や，(3) 品詞のみの選択を行って該当する単語をすべて入力として与えた場合に，15,000 属性までの異なり語を用いた実験で良い結果が得られた[25] との報告があったことなどから，後者の立場を選択した．N-gram は頻度を，パターンはその有無により 1 または 0 を値とし，ベクトルの長さが 1 になるように正規化を行って箇条書きごとに特徴ベクトルを作成した．用いた特徴量は次の 4 つのグループからなる．

F1-3 文字 N-gram (1, 1+2, 1+2+3)
F4-6 形態素 N-gram (1, 1+2, 1+2+3)
F7 形態素 N-gram(1+2)+文単位のパターン
F8 形態素 N-gram(1+2)+箇条書きごとのパターン

ここで，'1+2' は uni-gram と bi-gram の 2 つを，'1+2+3' は uni-gram，bi-gram，tri-gram の 3 つを特徴量として同時に用いることを示す．

実験に先立ち，タスクの難しさを見積もるために，実験に用いる記事セットを単純な基準で分類した場合の分類性能を調べた．1 つの可能性として箇条書きが手順について書かれたものであれば，ドメインにかかわらず項目の先頭においてナンバリングが行われていることが考えられる．また逆に，「・」（中黒）が項目

表 6 SVM による箇条書きの自動分類結果（多項式関数の次元数 d=2）
Table 6 Result of categorization of lists on SVM: polynomial kernel d=2.

	コンピュータ分野			その他分野		
	1 文	2 文	3 文	1 文	2 文	3 文
F0	0.25/1.00	−	−	0.09/1.00	−	−
F0-1	0.65/0.71	−	−	0.17/0.62	−	−
F0-2	0.65/0.76	−	−	0.17/0.60	−	−
F1	0.90/0.86	0.90/0.86	0.89/0.85	0.65/0.50	0.66/0.45	0.64/0.44
F2	0.91/0.87	0.90/0.85	0.89/0.85	0.66/0.50	0.65/0.45	0.68/0.45
F3	0.91/0.87	0.91/0.86	0.90/0.85	0.66/0.48	0.66/0.44	**0.70**/0.42
F4	0.92/0.91	0.92/**0.90**	0.92/0.88	**0.70**/0.52	0.67/0.47	0.68/0.42
F5	**0.93**/0.91	**0.93**/0.89	**0.93/0.89**	0.69/0.52	**0.74**/0.50	**0.70**/0.47
F6	**0.93**/0.90	**0.93**/0.89	0.92/0.88	0.67/0.55	0.71/**0.53**	0.68/**0.52**
F7	**0.93/0.92**	−	−	0.65/0.56	−	−
F8	**0.93/0.92**	−	−	0.65/**0.58**	−	−

の先頭において使われる場合には，非手順タイプが多いのではないかとも考えられる．<UL>タグに囲まれた箇条書きがブラウザ上に表示された場合は,「・」（中黒）として現れる．よって，次の単純な 2 つの基準によって記事セットを分類した．

F0-1 2 つの連続した項目の先頭において「1.」および「2.」がそれぞれ現れれば手順タイプとする．

F0-2 項目の先頭において中黒が使われた場合は非手順タイプとする．

評価は再現率，適合率によって行い，F 値についても検討時に補助的に考慮した．また F 値の計算は式 (2) によった．

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

5.2 実験結果と評価

はじめに，前節で示した特徴量を用いた実験の結果をまとめて**表 6** に示す．表 6 において F で始まる番号は，前節で述べた特徴量のグループを示している．また，条件 F0 はすべての箇条書きを手順タイプに分類した場合の参考値である．箇条書きの各項目において先頭の文から 3 番目の文まで使用する文の数を増やした場合のそれぞれ分類結果を列記した．各欄において，適合率，再現率を示し，同じ文の数を利用した場合に F1〜F8 で最も良かった値を太字で示した．

ナンバリングおよび・(中黒) の有無を用いた単純な分類でも，コンピュータ分野の場合はすべての評価値で 0.6〜0.7 程度である．一方，その他の分野については 0.1〜0.6 程度にとどまっており，項目の先頭において現れる記号の種別が，手順タイプであることを導く条件として必ずしも有効とはいえないことが分かる．

また，それぞれの分野の箇条書きに含まれる各項目において，分類に使用する文を項目の 1 文目から 3 文目まで増加させた場合，その他分野の条件 F5 および F6 を除いて，1 文目だけを用いた場合に最も良い結果を示した．コンピュータ分野についてはいずれの特徴量を用いた場合でも比較的高い精度で分類できた．今回使用したデータセットは，箇条書きが階層をなしている場合に 1 つの記事として切り出しているため，手順タイプか非手順タイプかにかかわらず，異なる記事の間で共通した内容語が多いと考えられる．このような場合にも，N-gram と SVM の組合せは分類に有効な特徴の抽出に成功している．

F7 については，記事セットの分野を問わず頻度において上位 1,000 パターンを使用した場合に最も良い結果となった．さらに,（条件 1）1 つの項目内においてパターンを適用する文を 1〜3 文までの 3 通り,（条件 2）特徴量として使用する形態素の品詞を，すべて，名詞以外，名詞＋動詞の 3 通り，に変えて同じ実験を行ったが各性能評価値の変化は 0.01 以内であった．

F8 については，コンピュータ分野の場合，頻度において上位 3,000 パターン以上を使用した場合に，またその他分野の場合には上位 1,000 パターンを使用した場合に最も良い結果となった．F8 についても上記条件 1 および条件 2 で実験を行ったが，各性能評価値の変化は 0.01 以内であった．

使用するパターンの頻度レンジを変えた場合の最低評価値との差は，適合率で 0.01，再現率で 0.06，F 値で 0.03 であった．シーケンシャルパターンを用いた特徴量はわずかながら N-gram のみを用いた場合を上回り，文字 N-gram の結果よりも形態素 N-gram の結果の方がおおむね上回った．bi-gram と tri-gram ではほとんど結果に差が見られなかった．一方，その他分野では最良の場合でも F 値で 0.60 程度にとどまった．すでに表 3 において示したように，その他分野の場合には 2 人の作業者による分類でも Kappa 値で 0.7 割程度の一致率であることから，この結果はタグ

（1）多用される言い回しを反映したと考えられるもの

```
<p> * <vp>の</vp> * <seg>を</seg> * <seg>。</seg>
<vp> し</vp> * <seg>、</seg> *
                    <adp>を</adp> * <seg>。</seg>
```

（2）スタイルと動詞のクラスを反映したと考えられるもの

```
<p> * <seg>1</seg> * <adp>を<adp>
<seg>2</seg><seg>. </seg> * <adp>を<adp>
<seg>3</seg> * <adp>を<adp>
```

図 3 使用されたパターンの例
Fig. 3 An example of exploited sequential patterns.

付けの揺れを反映していることが考えられる．

5.3 パターンの効果

実際にパターンを用いたことによって正しく分類できるようになった記事を見ると，コンピュータ分野においては，(1) 操作手順などの慣用的な言い回しを反映したと思われるパターンや，(2) ヲ格をとる動詞が連続的に使われることを反映したと考えられるパターンが多く見られた（**図 3**）．特に（2）のパターンは単純ではあるが，箇条書きの半分近くの部分で手順以外のことを述べているような場合に有効に働いた．

パターンを加えることによって正しく分類できるようになった記事がある一方で，正しく分類できなくなった事例も見られた．1 つの箇条書きの文字数が極端に少ない事例や，形態素解析の誤りがあった場合に単純なパターンに適合してしまう事例であった．後者の例としては，「がしがし整理」などの名詞句で，「が/し/が/し」と単語境界が判定されて，格助詞「ガ」と動詞「する」の連用形である「し」による品詞タグが付与された場合に，コンピュータ分野の手順の説明に比較的多く見られる「画面がポップアップし，」などの表現に現れる「〜が〜し」というパターンに一致する条件で誤りが発生する事例などがあった．

5.4 箇条書き単位のパターンマイニング

一方，F8 では文単位のパターンに加え箇条書き全体にわたるパターンを利用したにもかかわらず，大きな性能の変化は見られなかった．これは，PrefixSpan によるマイニングの過程において，与えるシーケンスの単位を文単位から箇条書き単位に変更する以外に特別な処理を施していないため，高頻度パターンに関しての重なりが多く，現実的な時間内でマイニングが終了するようなサポート値の範囲では，箇条書き全体にまたがる有効なパターンを獲得するに至らなかったものと考えられる．また箇条書き単位の場合には，離れた距離にある語の関係がとれる一方で，マイニングされるパターンが急激に増加するため，実用的な時間内でマイニングを完了するためには，あらかじめ与えるアイテムの種別に制限を加えるなどの対策も必要となる．

最終的に 2 つの分野ともパターンを用いた特徴量が最も良い結果となったものの，その差はわずかであり，いまだシーケンシャルパターンの有効性については明らかでない．そこで本タスクにおける機能語の利用可能性を調べる実験において，シーケンシャルパターンについて引き続きその有効性を調査することにした．

5.5 品詞のグルーピング

手順タイプの箇条書きを分類するための特徴量として，各特徴量の直接的な分類への寄与を見るために，SVM の多項式カーネル関数の次数を 1 次に設定し，N-gram についてもすべて 1 または 0 の 2 値でその箇条書きにおける存在を表して，SVM への入力ベクトルを 2 値ベクトルとして構成し直した．表 6 の実験では文字 N-gram よりも良い結果を残した形態素 N-gram の特徴量を，コンピュータ分野とその他分野のそれぞれにおいて，内容語と機能語の違いで大きく分けたうえで，3.4 節において記載した手順に関する箇条書きの特徴を考慮して以下の 6 つのグループに分けた．

M1 すべての品詞
M2 snp+np+vp+ajp
M3 snp+np+vp+ajp+unknown
M4 aup+adp+seg
M5 aup+adp+seg+unknown
M6 snp+aup+adp+seg

上記において snp などの記号は表 4 の分類に基づく．unknown は分析のために新たに追加した品詞タグであり，茶筌が未知語と判定した場合に付与する．M2 および M3 は内容語を中心とするグループ，M4 および M5 は機能語を中心とするグループである．M6 では 3.4 節において記載した箇条書きの観察結果に基づいて，機能語に加えサ変名詞を追加した．M1〜M6 それぞれを特徴量として用いて同じ条件で箇条書きの分類性能を比較した．まずコンピュータ分野の記事だけを用いたクローズド メインでの五分割交差検定の実験結果を示す（**表 7**）．表 7 に記載した異なり語数はコンピュータ分野での集計である．使用する特徴量として形態素 N-gram の N=1〜3 まで追加していった場合（表中 1，1+2，1+2+3 の列）と bi-gram に加えて文単位にマイニングしたシーケンシャルパターンを頻度の高いものから 1,000 パターン使用して特徴量とした場合（表中 pattern の列）の結果を記した．箇条書きの各項目において利用する文は，先頭の 1 文目のみとした．

表 7 異なる品詞を用いた場合の分類結果（クローズド メイン/次元数 d=1）
Table 7 Result of close domain categorization with POS groupings: d=1.

	コンピュータ分野				
	1	1+2	1+2+3	pattern	異なり語数
M1	0.88/0.88	0.92/0.90	0.93/0.90	0.93/0.92	9,885
M2	0.85/**0.86**	0.90/**0.87**	0.91/0.85	0.89/**0.88**	4,570
M3	**0.87/0.86**	**0.93/0.87**	**0.93/0.86**	**0.91/0.88**	9,277
M4	0.81/0.81	0.85/0.85	0.86/**0.86**	0.86/0.86	608
M5	0.81/0.84	0.86/0.85	0.90/**0.86**	0.89/**0.88**	5,315
M6	0.85/0.87	0.90/0.89	0.91/0.89	0.89/0.89	1,493

表 8 異なる品詞を用いた場合の分類結果（オープンド メイン/次元数 d=1）
Table 8 Result of open domain categorization with POS groupings: d=1.

	コンピュータ分野 → その他分野				その他分野 → コンピュータ分野				
	1	1+2	1+2+3	pattern	1	1+2	1+2+3	pattern	異なり語数
M1	0.60/0.46	0.69/0.45	0.72/0.45	0.66/0.48	0.90/0.52	0.95/0.60	0.97/0.56	0.95/0.64	13,031
M2	0.52/0.42	**0.69**/0.39	**0.72**/0.37	**0.64**/0.41	0.88/**0.51**	0.92/0.44	0.94/0.37	0.94/0.47	7,818
M3	**0.56**/0.46	0.68/0.44	0.70/0.42	0.63/0.45	**0.90**/0.46	**0.95**/0.48	**0.97**/0.41	**0.96**/0.49	12,169
M4	0.46/**0.51**	0.59/**0.58**	0.58/**0.52**	0.53/**0.60**	0.80/0.33	0.79/**0.58**	0.79/**0.55**	0.79/**0.59**	862
M5	0.43/0.50	0.52/0.48	0.61/0.48	0.53/0.53	0.83/**0.51**	0.85/0.54	0.88/0.51	0.87/0.53	5,213
M6	0.53/0.49	0.67/0.53	0.71/0.50	0.61/0.55	0.81/0.51	0.90/0.56	0.94/0.51	0.89/0.56	2,360

5.6 品詞による分類結果の相違

表 7 においては，1+2+3 の場合と pattern における再現率を除いてすべて内容語を中心とするグループが機能語のグループの評価値を上回っている．しかし，内容語を中心としたグループと機能語を中心としたグループとの性能の差はわずかであり，M2 または M3 に対して M4 を比較すると，分類に使用した異なり語の数の違いに比べ，分類性能の差は小さいものになっている．特徴量として語の頻度を用いていないため，この結果は純粋に語とその組合せによるものである．また機能語のグループにサ変名詞を加えることによって若干精度が改善される様子が分かる．パターンによる性能改善はほとんど見られなかった．

次にコンピュータ分野の箇条書きおよびその他分野の箇条書きを，学習と評価の役割を入れ替えて実験した結果を**表 8** に示す．表 8 に示した集計値はその他分野の異なり語数である．内容語を中心とするグループでは，N-gram において利用する単語の組合せを増やすほど，再現率が悪化している．

一方機能語に関しては組合せを考慮することによって性能が向上しており，適合率と再現率ともに，内容語に比べると全般に良い結果となっている．式 (2) を用いて F 値を計算して比較すると，その他分野で学習した場合の uni-gram の場合を除いて，機能語のグループが内容語のグループを上回った．異なる分野から得られた機能語を中心とする特徴量を用いて，正例を 10%程度しか含まない記事セットに対し，高精度とはいえないものの，F 値で 0.6 以上の精度が得られており，手順に関する記述を含む箇条書きの候補抽出を目的とした緩やかな絞り込みなどのタスクで，一定の効果が期待できると考えられる．本研究のアプローチが幅広い分野に対する質問応答の前処理として使用できる可能性を示すものである．

また，異なる分野間での比較では，コンピュータ分野の箇条書きで学習した場合に比べて，その他分野の箇条書きで学習した場合の方が良い結果が得られた．機能語は文章の記述スタイルに関する特徴を含むとされており，様々なジャンルを含むその他分野の箇条書きで学習することによって多様な記述のスタイルの特徴を取り込むことに成功しているのではないかと考えられる．

また M2，M3，M4 などにおいて，bi-gram を加えた場合に対して tri-gram を加えた場合では性能が劣化しているにもかかわらず，pattern を加えた場合には改善されていることから，隣接 3 単語よりも遠い距離にある語の間に，手順に関する箇条書きの分類に寄与するパターンが存在する可能性がある．また，オープンド メインの場合にもサ変名詞を特徴量に加えることによって性能が改善されることが表 8 からうかがえる．しかし，この実験においてもパターンによる顕著な性能の改善は見られなかった．

5.7 手がかりとなる表現

箇条書きの観察に基づいて得られた知見についてさらに調査した．形態素 N-gram を用いた分類において，使用する語の品詞の組合せを変えて F 値の変化を調べた（**図 4**）．横軸には 1〜8 まで使用する品詞を減らす

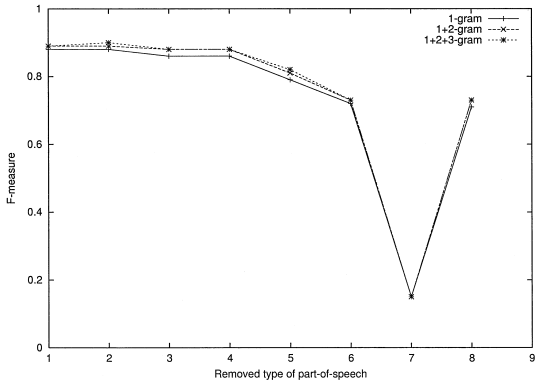

図 4 使用する形態素を変えたときの F 値の変化
Fig. 4 F-measure results, when types of POS are selected.

段階を記した．使用する品詞を 1〜6 まで，snp，np，vp，unknown，adp，seg の順に減らし 7，8 ではそれぞれ aup を除いた場合と ajp を除いた場合を記した．

aup タグに対応する助動詞，終助詞，接尾語を分類に使用するか否かによって大きく分類性能が変わることが分かった．SVM によって得られたサポートベクタに現れる aup に関連した特徴量を調べ，各尤度を計算してランキングを行うと，正例においては「〜たら,」「〜ましたら,」などの読点前に現れる動詞のアスペクト，および「れたら,」「れた後」など受動表現，負例においては「時に」「たとき」などの時点を表す表現や「ん.」「ません.」など否定辞を含む文末表現が上位になった．これは，箇条書きの観察によって得られた知見を支持するものである．

次に，文頭および文末表現に現れる語について分類への影響を調べた．形態素 N-gram を用いた分類において，特徴量として使用する形態素を各文に対して文頭および文末それぞれ 3 つの形態素に制限した．次に，文頭において文末へ向かって使用する形態素を 1 つずつ増やした場合と，文末において文頭へ向かって使用する形態素を 1 つずつ増やした場合の F 値の変化を調べたところ，文末から 2 番目の形態素を使用するか否かによって，F 値が大きく変わることが分かった．この位置に使われていた形態素について頻度の高いものを調べると，正例では格助詞「を」，負例では「いる」「ある」などの状態動詞，助詞「の」，否定辞「ん」「ない」などであった．

シーケンシャルパターンマイニングによって獲得されたパターンを加えることによって，正しく分類できるようになった箇条書きを調べると，固有名詞など特定のクラスの表現を囲んでいる括弧などの記号と機能語との組合せによるパターンが多く見られた（**図 5**）．こうした特定の記号とともに使用される語とそれと共

箇条書きの一部 ：
「メニュー」 を 選択し、「確認」 ボタン を押します。

上記の箇条書きの分類に利用されたパターン：
パターン１： <seg>「</seg> * <seg>」</seg> * <adp>を</adp>
パターン２： <adp>を</adp> * <aup>ます</aup>

図 5 手順タイプの箇条書きに多く見られるパターンの例
Fig. 5 An example of frequent patterns in procedural expressions.

起するパターンの種別に基づいて，さらに詳細に箇条書きが分類できる可能性も考えられ，興味深い．

6. あとがき

WWW 上に多数存在する箇条書きを手順・非手順に分類するために有効な特徴量について検討した．手順に関する箇条書きの分類タスクが，従来の分野に基づいた文書分類とは，それに有効な特徴量や，語の共起に関する性質において異なることを示した．また手順に関する箇条書きから得られた機能語を中心とした特徴量を用いることによって，異なる分野の手順タイプの箇条書き抽出において，一定の絞り込みがが期待できる可能性を示した．

キーワードとして「手順」および「方法」を含む Web ページを，検索エンジンを用いて収集し，N-gram を用いた特徴量と SVM を組み合わせることによって，コンピュータ分野に関しては高い精度で手順タイプの箇条書きが抽出可能であることを示した．しかし今回の実験では，bi-gram に対して tri-gram を加えた場合とパターンを加えた場合の性能変化の異なりや，サポートベクタやマイニングされたパターンの分析から発見できた事例などにおいて，分類に寄与するパターンの存在を期待させる現象が散見されたものの，その有効性を立証するに十分な事実を提示することができなかった．パターンの有効性の検証は今後の課題である．

一方，その他分野の箇条書きについては，今回の実験に用いた記事のドメインごとの事例数では十分とはいい難い．専門性の高い分野においては，正解データのタグ付け作業者が内容を理解することが困難である場合も考えられ，箇条書きのドメインや内容の違いをより詳細に考慮した検討が必要である．

また現在考慮していない問題点の 1 つに，箇条書きの階層構造の扱いがある．現在は箇条書きの各階層を別の箇条書きとして扱っているために，いくつかの階層をあわせることによって手順タイプと判定できるものを非手順タイプとして判定してしまう場合がある．これを防止するため，階層構造を持つ箇条書きについ

て検討を進め，どのような場合に 1 つの箇条書きとして扱うべきかを明らかにする必要がある．

2 つ目には，名詞句タイプの箇条書きの扱いがある．その他分野において 2 人の作業者の判断が分かれた記事について調べたところ，そのいくつかは名詞句タイプの箇条書きであった．このような場合には箇条書き部分だけでは判断が難しく，箇条書き前後のテキストなどを利用して箇条書きの内容を示す手がかりを増やす必要がある．

今後は上記の課題解決に加え，手順について書かれた複数の箇条書きを対応づけることにより，手順の特定の部分だけを詳細化できるようなナビゲーションや，パッセージによる回答が必要な手順以外の質問応答のための箇条書きの分類などを考えている．

謝辞 Tiny-SVM，PrefixSpan をご提供いただいた奈良先端大の工藤拓氏に深く感謝いたします．東京工業大学大学院田中・徳永研究室および奈良先端大松本研究室の皆様のご協力に深く感謝いたします．

参 考 文 献

1) Agrawal, R. and Srikant, R.: Fast Algorithms for Mining Association Rulesr, *Proc. 20th Intl. Conf. Very Large Data Bases* (*VLDB*), pp.487–499 (1994).
2) Barr, A., Cohen, P.R. and Feigenbaum, E.A.（著），田中幸吉，淵　一博（監訳）：人工知能ハンドブック，共立出版，Tokyo (1989).
3) Berger, A.L. and Mittal, V.O.: (OCELOT): A System for Summarizing Web Pages, *Research and Development in Information Retrieval*, pp.144–151 (2000).
4) Brin, S. and Page, L.: The Anatomy of a Large-Scale Hypertextual Web Search Engine, *Proc. 7th Intl. World Wide Web Conf.* (1998).
5) Eguchi, K., Oyama, K., Ishida, E., Kando, N. and Kuriyama, K.: Overview of the Web Retrieval Task at the Third NTCIR Workshop, Technical Report NII-2003-002E, National Institute of Informatics (2003).
6) Fukumoto, J. and Kato, T.: An Overview of Question and Answering Challenge (QAC) of the Next NTCIR Workshop, *Proc. 2nd NTCIR Workshop*, pp.144–151, National Institute of Informatics (2001).
7) Hawking, D.: Overview of the TREC-9 Web Track, *Proc. 9th Text REtrieval Conf.* (*TREC 9*) *NIST Special Publication 500-249*, pp.97–112 (2000).
8) Kan, M.Y., McKeown, K.R. and Klavans, J.L.: Applying Natural Language Generation to Indicative Summarization, *Proc. 8th European Workshop on Natural Language Generation*, pp.92–100 (2001).
9) Kurohashi, S. and Higasa, W.: Dialogue Helpsystem based on Flexible Matching of User Query with Natural Language Knowledge Base, *Proc. 1st ACL SIGdial Workshop on Discourse and Dialogue*, pp.141–149 (2000).
10) Lai, Y., Fung, K. and Wu, C.: FAQ Mining via List Detection, *Proc. Workshop on Multilingual Summarization and Question Answering* (*COLING*) (2002).
11) Lewis, D.D. and Ringuette, M.: A Comparison of Two Learning Algorithms for Text Categorization, *Proc. SDAIR-94 3rd Annual Symposium on Document Analysis and Information Retrieval*, pp.81–93 (1994).
12) Moldovan, D., Pasca, M., Harabagiu, S. and Surdeanu, M.: Performance Issues and Error Analysis in an Open-Domain Question Answering System, *Proc. 40th Annual Meeting of the Association for Computational Linguistics* (*ACL*), pp.33–40 (2002).
13) Pei, J., Han, J., et al.: Prefixspan: Mining Sequential Patterns by Prefix-Projected Growth, *Proc. Intl. Conf. of Data Engineering*, pp.215–224 (2001).
14) Ravichandran, D. and Hovy, E.: Learning Surface Text Patterns for a Question Answering System, *Proc.40th Annual Meeting of the Association for Computational Linguistics* (*ACL*), pp.41–47 (2002).
15) Soubbotin, M.M. and Soubbotin, S.M.: Patterns of Potential Answer Expressions as Clues to the Right Answers, *Proc. 10th Text Retrieval Conf.* (*TREC 2001*), pp.293–302 (2002).
16) Tsuboi, Y. and Matsumoto, Y.: Authorship Identification for Heterogeneous Documents, 情報処理学会研究会報告，NL-148-3, pp.17–24 (2002).
17) Vapnik, V.N.: *The Nature of Statistical Learning Theory*, Springer, New York (1995).
18) Voorhees, E.M.: Overview of the TREC 2001 Question Answering Track, *Proc. 2001 Text Retrieval Conf.* (*TREC 2001*) (2001).
19) Yang, Y. and Pedersen, J.O.: A Comparative Study on Feature Selection in Text Categorization, *Proc.ICML-97 14th Intl.Conf.on Machine Learning*, pp.412–420 (1997).
20) 内野寛治，宗意幸子，橋本三奈子，武智峰樹，松井くにお，菊田泰代：ルールベースを用いたテキスト分類サービス―自動分類技術のビジネスへの応用，情報の科学と技術，Vol.50, No.10, p.502 (2000).

21) 岩山 真，徳永健伸：確率モデルに基づくパッセージ分類とその応用，自然言語処理，Vol.6, No.3, pp.181–198 (1999).
22) 吉田篤弘，延澤志保，平石智宣，斉藤博昭：著者判別に有効な特徴量の推定，情報処理学会研究会報告，NL-145-13, pp.83–90 (2001).
23) 藤井 敦，石川徹也：IT 技術者試験を対象とした質問応答システム―事典情報に基づく用語問題の解法，言語処理学会第 7 回年次大会発表論文集，pp.514–517 (2001).
24) 藤井 敦，石川徹也：World Wide Web を用いた事典知識情報の抽出と組織化，電子情報通信学会論文誌（D-II），Vol.J85-D-II, No.2, pp.300–307 (2002).
25) 平 博順，春野雅彦：Support Vector Machine によるテキスト分類における属性選択，情報処理学会論文誌，Vol.41, No.4, pp.1113–1123 (2000).
26) 平 博順，春野雅彦：トランスダクティブ・ブースティング法によるテキスト分類，情報報処理学会研究報告，NL-139-10, pp.69–76 (2000).
27) 金 明哲：助詞の n-gram モデルに基づいた書き手の識別，計量国語学，Vol.23, No.5, pp.225–240 (2002).
28) 松本裕治，北内 啓，山下達雄，平野善隆，松田 寛，浅原正幸：形態素解析システム『茶筌』version2.0 使用説明書，Naist Technical Report NAIST-IS-TR99009, Nara Institute of Science and Technology (1999).
29) 浜田玲子，井手一郎，坂井修一，田中英彦：料理テキスト教材における調理手順の構造化，電子情報通信学会論文誌（D-II），Vol.J85-D-II, No.1, pp.79–89 (2002).
30) 竹内和広，松本裕治：自動要約を視野にいれたテキスト構造解析実験，情報処理学会研究報告，NL-133-9, pp.61–68 (1999).

(平成 14 年 12 月 27 日受付)
(平成 15 年 6 月 27 日採録)

（担当編集委員 安達 淳）

武智 峰樹（学生会員）

1991 年埼玉大学理学部数学科卒業．同年富士通（株）入社，現在に至る．産業向け情報サービスの技術開発に従事．2000 年より奈良先端科学技術大学院大学情報科学研究科博士後期課程在学．人工知能学会，言語処理学会，ACM 各会員．

徳永 健伸（正会員）

1983 年東京工業大学工学部情報工学科卒業．1985 年同大学院理工学研究科修士課程修了．同年（株）三菱総合研究所入社．1986 年東京工業大学大学院博士課程入学．現在，同大学院情報理工学研究科助教授．自然言語処理，計算言語学の研究に従事．工学博士．人工知能学会，言語処理学会，計量国語学会，ACL，ACM SIGIR 各会員．

松本 裕治（正会員）

1977 年京都大学工学部情報工学科卒業．1979 年同大学院工学研究科修士課程情報工学専攻修了．同年電子技術総合研究所入所．1985 年～1987 年（財）新世代コンピュータ技術開発機構に出向．京都大学助教授を経て，1993 年より奈良先端科学技術大学院大学教授，現在に至る．工学博士．専門は自然言語処理．人工知能学会，日本ソフトウェア科学会，言語処理学会，認知科学会，AAAI，ACL，ACM 各会員．

田中 穂積（正会員）

1964 年東京工業大学理工学部制御工学科卒業．1966 年同大学院修士課程修了．同年電気試験所（現，産業技術総合研究所）入所．1983 年より東京工業大学工学部助教授．現在，同大学院情報理工学研究科教授．自然言語処理，人工知能に関する研究に従事．工学博士．電子情報通信学会，認知科学会，人工知能学会，計量国語学会，言語処理学会各会員．