自然言語処理 152－2
（2002．11．12）

手順の説明を含む箇条書きを抽出するための手がかり分析

武智峰樹[*1*2]　徳永健伸[*3]　松本裕治[*2]　田中穂積[*3]

[*1]富士通（株）　[*2]奈良先端科学技術大学院大学　[*3]東京工業大学
{mineki-t, matsu}@is.aist-nara.ac.jp　{take, tanaka}@cl.cs.titech.ac.jp

質問応答システムの研究は,主に回答のタイプとして固有名詞や数値表現を扱う段階から,パッセージによる回答が必要な問題を扱う段階へ入ってきた。この新たな研究の端緒として,Web 上の HTML(Hyper Text Markup Language)のリストタグが付与されたテキスト（箇条書き）を,手順タイプと非手順タイプに機械的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１手法である SVM(Support Vector Machine)を用いて学習を行い,得られたモデルを検討することにより分類に有効な特徴量を考察した。筆者推定のタスクに用いられる手法をベースに手法の改良を試み,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み合わせを得た。

Keywords :質問応答, Web 文書, パッセージ, 手順, サポート・ベクター・マシン, 著者推定

Extraction of Procedural Expressions in a List Using Surface Linguistic Cues

Mineki Takechi[*1*3]　Takenobu Tokunaga[*3]　Yuji Matsumoto[*2]　Hozumi Tanaka[*3]

[*1]Fujitsu Ltd.　[*2]Nara Advanced Institute and Science and Technology
[*3]Tokyo Institute of Technology
{mineki-t, matsu}@is.aist-nara.ac.jp　{take, tanaka}@cl.cs.titech.ac.jp

In these days, the research of Question-Answering(QA) has advanced from the generation that focuses on the answer types of named entities and numerical representation to the new one that focuses on the answer types of passages. For the initial step of our work to such type of QA, we set the first task on procedural expressions in a list form tagged with list-tags of HTML found in various Web pages. Applying SVMs(Support Vector Machines) on documents gathered by a search engine, we investigated the obtained model to exploit useful features for the extraction task. Improving on the features for authorship identification as the baseline, we figure out the feature set that achieves more than 90% of recall and precision of extraction in computer domain.

Keywords :QA task, Web documents, Passage, Procedure, Support Vector Machine, Authorship Identification

1. まえがき

近年の電子化文書の爆発的な増加をうけて,ユーザの情報探索を支援する研究はより知的な情報アクセスを指向して進んでいる。このような研究動向として,Semantic Web などインターネット上で複雑な情報探索を可能にするインフラや,質問応答システムなど従来の情報検索に比べてより高精度な情報アクセスを目指すシステムの研究がある。

質問応答システムについては,TREC,NTCIR において質問応答をタスクとしたコンペがここ数年毎年開催され,盛んに研究発表が行われている。このような研究の場においては主に事実を尋ねる質問応答が扱われてきた。こうした研究成果を受けて,質問応答システムにおける幾つかの主要な課題が回答を含むパッセージを決定する処理の部分にあることが明らかになってきている。また TREC における list-question など,回答として複数の句や文などからなるパッセージを返す質問応答の研究も始まっている。

しかし従来の研究においては,パッセージを回答として返すべき質問について十分に議論されてきたとは言えない。 我々はパッセージによって回答すべき質問を手順,比較,例示など５つのタイプに分類し,それぞれのタイプにおいてパッセージをどのように処理すべきかを研究することにした。

研究の端緒として,我々は手順を尋ねるタイプの質問を選んだ。このような質問に対して Web を情報源として回答を返す質問応答を想定し,Web ページに含まれる手順に関する箇条書きを候補として,その中から回答を選び出すタスクを考えた。この研究の準備段階として,Web ページに含まれる箇条書きを集めて,手順に関するものとそうでないものとに自動分類する研究を行っている。本稿ではその進捗状況を報告する。現状のデータセットは小規模であるが,従来から著者推定に用いられてきた特徴量を利用することにより,同一ジャンルの箇条書きについては９０％以上の精度で自動分類が可能であることが分かった。

2. 質問応答システムの問題点

質問応答システムの先行研究において,その主要な課題として次のような整理が行われている[1]。

・回答タイプの判定： 入力された質問に対してどのような回答タイプ（人名,国名など）が求められているかを判定する際に利用できる頑健なモデルが存在しないため,従来の情報抽出の手法に頼らざるを得ない。このため,抽出用のテンプレートが適用できない質問の場合には回答タイプの判定が難しい。

・パッセージ検索の精度[1]： 正解を含む記事セットから回答を探し出す処理フローにおいて,それぞれの記事から質問文中のクエリの内容と強いつながりをもつ連続した一部分（**パッセージ**[2]）を収集し,その後の処理をパッセージに対してのみ行うシステムでは,パッセー

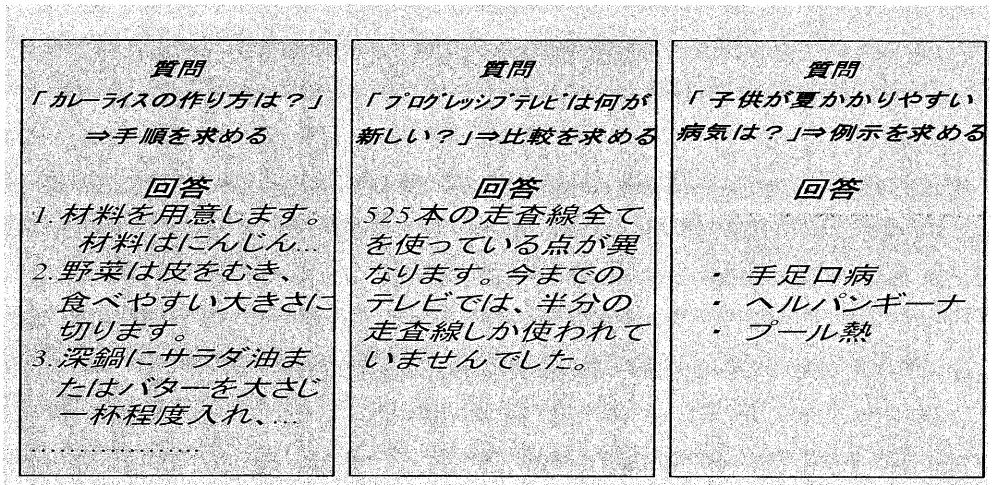

図１　パッセージによる回答が必要な質問応答の例

ジ検索における精度がシステム全体の性能に大きく影響する。

現在までの質問応答は事実に関する質問を主に扱ってきたが,実際にはこれ以外にも様々なタイプの質問が存在する。黒橋ら[3]は, ソフトウェア製品のヘルプデスクに寄せられる質問を6つのタイプに分類し,このうち事実を尋ねる質問(What-type), 方法を尋ねる質問(How-type), 問題の解決策を尋ねる質問(Symptom-type)に対して対話を用いた質問応答を行っている。How タイプや Symptom タイプの質問は文によって回答できるものもあるが,手順,比較,例示などが要求されている場合には,パッセージによって回答する方が自然である場合も少なくない（図1）。

論理的には1つの文を用いて幾つものステップや選択肢を記述したり,1文ずつ部分的な回答を対話によって与えることは可能だが,文が長くなることで人間による可読性が悪くなることや,完全な回答を得るまでのインタラクションが長くなるなどの問題がある。How タイプや Symptom タイプの質問に対しては,複数の文から構成されるパッセージによる回答を含めて考える方が自然である。

2.1 パッセージを単位とするテキスト処理

複数の文からなるパッセージを主な処理単位とするテキスト処理として,次のような先行研究がある。

・パッセージ分割,分類,検索： サブトピック毎にテキストを分割するテキストタイリング[4], 長い文書を対象とした文書分類においてパッセージを分類対象とすることで分類精度を向上するパッセージ分類[2], 特定のトピックに関連したパッセージだけを検索結果として返すパッセージ検索[5]などがある。これらの研究は,パッセージによって回答する必要がある質問応答において, どのようなパッセージを選ぶべきかについては考慮していない。

・テキスト自動要約： テキスト自動要約は,原文の大意を保持したまま, テキストの長さ, 複雑さを減らす処理[6]とされる。なかでも動的要約と呼ばれるテキスト自動要約の一分野は質問応答と密接な関連にある。要約の際に用いられるテキストは,要約対象となる文書の連続した一部分とは限らないが,要約済みのテキストはクエリと関連した文の集合となっており,自動要約はパッセージを生成していると言える。しかしこれらの先行研究も,パッセージによって回答すべきタイプの質問応答を想定した研究ではない。

2.2 パッセージによる質問応答

パッセージによる応答が必要な回答のタイプを網羅的かつ厳密に定義・分類することは困難である。多様な内容や表現を含む現実の質問に対しては,当面実用的な観点から役に立つ分類を与えることの方が有益であると考える[3]。我々は先行研究に照らし,パッセージで回答すべき質問応答のうち部分的には既に実現可能性であると考えられる課題を中心に,次のような分類を試みた。

・手順・方法：目的達成のための手順・方法を尋ねる
・時間的順序：時間的な順序関係を尋ねる
・例示　　：実例のリストを要求する
・比較　　：類似性や新規性を尋ねる
・その他　　：上記以外の下記のような回答タイプ
　-目次など単なるラベルの集合
　-複雑な因果,時間,空間的関係
　-意味的・概念的な包摂など

ここに示した回答タイプの質問応答は,どれも複数の正解が許されるという点で一致している。したがって,ここに挙げたパッセージによる回答が必要な質問応答は,唯一の正解を返す問題ではなく,正解の集合を返す問題としてとらえるべきである。

一方,上記の分類においては同じ条件を共有するカテゴリがあるが,それらは同時に異なる制約条件を有している。例えば,手順と時間的順序は共に動作や事象の順序が正しく回答できなければならないが,時間的順序においては順序関係が正しく回答できれば正解であるのに対し,手順は目的達成に必要なステップがユーザの要求に応じた粒度で含まれていなければ正解とは言えない。したがって,ここに挙げたパッセージによる質問応答は,質問のタイプ毎に個別に検討されるべきで課題であると考える。我々は,この研究の端緒として,まず手順を回答するタスクを扱うことにした。

2.3 手順の回答

手順について,日本語を対象に複数の文からなるパッセージを回答するタスクを直接扱う研究は我々の知る限りない。

例えば,「RedHat のインストール方法を知りたい。」というユーザが,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（「RedHat」「インストール」），手順に関する単語（「手順」「方法」「やり方」）を与えると考えられる。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まず十分な回答になっていないことも多い。例えば,単なるリンク集となっていたり,手順について書かれはいるものの選択肢を列挙するに留まる場合もある。

ex) …インストールのやり方は3つあります
1）パッケージを買ってくる
2）ホームページからダウンロードする
3）雑誌の CD-ROM を使う

本研究では,このような場合でも具体的な手順を回答として返すことを目指す。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,あらかじめ連続したテキストの一部分となっているとは限らないため,手順を回答する質問応答は,一般的には要約が必要となる。しかし,この方面での研究は十分に進んでおらず,どのような要約を行うべきであるか,またテキスト中でどのような表現が用いられるのかが明確ではない。そこで今回は,回答が検索対象のテキスト中において連続した一部分として存在する場合を考えた。さらに,回答候補として扱うパッセージのスタイルを箇条書きに制限することによって,問題を扱いやすくした。

箇条書きは人手によってあらかじめ要約されたテキストであると考えられるため,重要な情報が含まれていることが期待できる。加えて機械的な処理を行う上で以下のような利点がある。

・ＱＡ集やWebページにおいて数多く使われている。
・箇条書きの前後にタイトルなどの手がかりがある。
・HTMLタグを利用して比較的容易に抽出が可能。
ex）<OL>, <UL>などのリストタグ

本研究では,記事の収集において上記のような利点をもつ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とそれ以外の箇条書きに分類するタスクを考える。手順について書かれた箇条書きとそれ以外の箇条書きを自動分類することで,分類に役立つ特徴量を調べることを目的とする。

	手順	非手順
コンピュータ	295	724
その他	64	476

表1 人手による箇条書きの分類結果

3 Webページ上の箇条書き

3.1 Webページ上の箇条書きの収集

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ようにして箇条書きを集めた。

ステップ1　Google に対してキーワードとして「手順」を与え,収集の起点となるURL 748 件を得た。

ステップ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 3713 件を収集した。

ステップ3　ステップ2で得た HTML ファイルから<OL>又は<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もつ場合には各階層を1つの箇条書きとして分割した。

ステップ4　2 つ以上の項目をも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。最終的に 1559 件の箇条書きを収集した。

このようにして得られた記事セットを人手によって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について便宜的に以下のような定義を与えた。

「ある目的を達成するまでの複数の行為又は動作を,それぞれ項目に記述して実行すべき順序に並べたもの」

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つ又は複数の文からなる。

また,得られた箇条書きが含まれているWebページのジャンルについても人手により分類した。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ジャンルの記事を含む。その他分野のWebページに含まれる箇条書きでも,ソフトウェアの操作方法等を説明したものはコンピュータ分野に分類した。分類結果を表1に示す。

3.2 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ページのジャンルに関わらず,箇条書きにおける手順の表現として次のような特徴が見られた。

・項目の1文目に行為や動作が示されることが多い
・1文目において,文末が主に行為や動作を表す名詞句で終わるタイプと,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。

ex 1)「（スタート）を（押）します。」

ex 2)「～を（閉じ）る。」

ex 3)「～を（押）し,（番号）を（入力）し,（ノブを回）す。」

ジャンルに関わらず前記のような特徴があるとすれば,特定のジャンルから得られた分類に有効な特徴量を用いて他のジャンルについても手順・非手順の分類が可能となる。文末や読点前に現れる機能語や活用語尾には,テキストの記述スタイルが現れるとされており,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。

上記のような特徴量を用いるテキスト処理として,著者推定がある。著者推定では,複数の著者によるテキストを含む記事セットを著者毎に分類する。われわれは,手順タイプと非手順タイプを分類する問題と著者推定の問題を,共にテキストの記述スタイルを推定する問題と見なし,我々のタスクに対して著者推定の手法をベースに手法の改良を試みた。

4．分類精度向上のための特徴量

4.1 ベースライン

我々は,まず日本語を対象とした著者推定の先行研究において取り上げられた特徴量が,手順タイプの推定にどの程度有効であるかを調べた。吉田ら[7]は,従来から文学作品の筆者推定に用いられてきた単語 N-gram や,文頭・文末・読点前などに限ったときの品詞の出現頻度,一文辺りの文字数,漢字数などの特徴量について,その有効性を検討している。

表２　使用したタグセット

	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－＊1 接頭詞
	snp	名詞－サ変接続
	vp	動詞
	adp	助詞－＊2 副詞 連体詞 接続詞
	ajp	形容詞
	aup	助詞－終助詞 助動詞 ＊3－接尾
	ij	感動詞
	seg	その他

＊1:サ変接続以外
＊2:終助詞以外
＊3:任意の品詞

また,坪井[8]らは形態素N-gramに加えて,Sequential pattern mining の１手法である PrefixSpan[9]と機械学習の１手法である SVM(Support Vector Machine)[10]を用いて,同一著者のテキストに頻出する文を単位としたときの語の出現パターンを特徴量として用いる手法を提案している。Sequential pattern mining を用いることにより,隣接しない語の間の関係を積極的に分類に利用できる利点がある。

4.2 項目を超えて現れるパターン

われわれは従来の著者推定の特徴量に加えて,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。既に指摘したように,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,複数の項目に渡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

図２　箇条書きへのタグ付け例

タグ付け後の箇条書き:

<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>

<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>

PrefixSpanに与える文字列（ｎ ＝ １ の場合）:

 <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> <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>

て次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。

ステップ1　茶筌[11]を用いて形態素解析を行い，表2のように文書タグと品詞タグを付与した。表2における品詞名は茶筌の品詞体系に基づく。

ステップ2　その他を除いた後,各項目の1文目からn文を取り出し項目タグを含めて1記事を1つの文字列とし（図2）,形態素及び文書タグを1アイテムとしてSequential pattern miningを行った。

5．Sequential Pattern Mining

Sequential pattern miningでは,あるアイテムの集合I={i1,i2,...in}上の集合si(エレメント)の順列<s1,s2,...si>（シーケンス）を考える。Sequential pattern miningとは,シーケンスを要素とする集合S（シーケンス・データベース）があるとき,任意の正の整数ξ（サポート）に対してSにおいてξ回以上現れるシーケンスを全て数え上げる操作を言う。代表的なSequential pattern miningの手法として,Apriori アルゴリズム[12],PrefixSpanなどがある。PrefixSpanは深さ優先で頻度の高いパターンを探索するため,自然言語処理のように種類が多くかつスパースなアイテムの集合を扱う場合には,幅優先に探索を行うAprioriアルゴリズムよりも効率的なため,今回はPrefixSpanを利用した。PrefixSpanは,与えられたシーケンスデータベースに含まれるシーケンスの先頭から末端へ向かって走査し,サポート値以上の頻度で出現するアイテム（高頻度アイテム）それぞれについて,それに引き続く高頻度アイテムからなるシーケンスだけを残す操作（射影）を繰り返す（図3）。射影によって作成されたシーケンスデータデースに対して高頻度アイテムを再度計

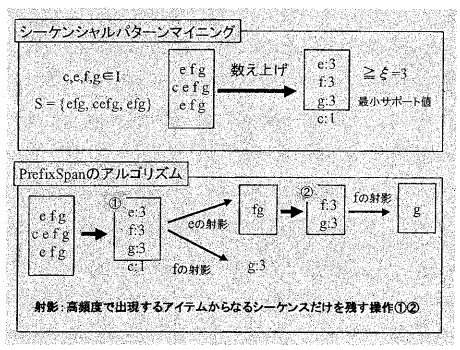

図3　シーケンシャルパターンマイニングとPrefixSpan

算することにより,最初に与えられた大規模なシーケンスデータベースに対する全探索を回避することが可能となる。

6．Support Vector Machine

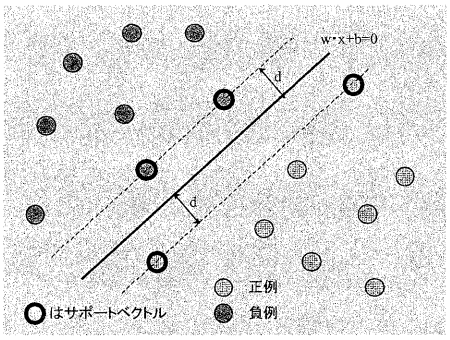

図4　サポートベクトルマシン

SVM(Support Vector Machine)は,高次元の特徴空間に対しても高い汎化能力をもつ2値線形分類器である。少ない事例数による学習でも高い分類精度が得られることが知られている。SVMはマージン最大化と呼ばれる学習方略に基づき,特徴空間における正例と負例の分離平面$\mathbf{w}\cdot\mathbf{x}+b=0(\mathbf{w},\mathbf{x}\in\mathbf{R}^n)$から学習事例までの距離d（マージン：式（1））が最大になるように分離平面を決定する(図4)。

$$d=\min\frac{|\mathbf{w}\cdot\mathbf{x}_i+b|}{\|\mathbf{w}\|}+\min\frac{|\mathbf{w}\cdot\mathbf{x}_i-b|}{\|\mathbf{w}\|}=\frac{2}{\|\mathbf{w}\|} \quad (1)$$

また,線形分離できない学習データについては,高次元空間に写像することによって線形分離を行う。高次元空間における分離平面のパラメータ計算を,カーネル空間と呼ばれる内積空間において行うことで,次元増加による計算量の増大を緩和する。学習結果として分離平面を定める少数のsupport vectorとその重みが得られる。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み合わせを探っていく必要がある。高次元の特徴空間に対して少ない事例数でも高い汎化能力を示すSVMは,われわれの研究に適している。

7　分類実験

7.1 実験設定

コンピュータ分野の箇条書きだけからなるデータセットを用意し,手順タイプと非手順タイプに自動分類し

た。評価は5foldの交差検定によって行った。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,手順タイプと非手順タイプに自動分類を行った。いずれの場合もrecall, precision, F-measure によって評価した。F-measureは式（2）によった。

$$F = \frac{2PR}{P+R} \qquad (2)$$

今回は,箇条書きされた部分のテキストだけを使用し,箇条書きの前後は利用しなかった。表 1 に示した箇条書きを分野毎に分け,表 2 のようなデータセットとした。分類器にはTiny-SVM[13]を使用し,PrefixSpanは同じく工藤拓氏による実装[14]を使用した。

箇条書き	総数	手順	非手順	項目数	文数
コンピュータ	1019	295	724	4710	6687
その他	540	64	476	2348	2937

表2　データセット

・用いた特徴量

1）文字 bi-gram

2）形態素 bi-gram

3）著者判定1

読点前の文字種,各文毎の各ひらがなの出現頻度,各品詞の出現頻度,文頭での各品詞の出現頻度,文末での各品詞の出現頻度,1文辺りの文字数／漢字数／読点数

4）著者判定2

2）+語の出現ﾊﾟﾀｰﾝ（文単位1000ﾊﾟﾀｰﾝ）

5）箇条書きパターン

2）+語の出現ﾊﾟﾀｰﾝ（箇条書き単位1000ﾊﾟﾀｰﾝ）

各項目の先頭1文目だけを分類に使用した。また,特徴量における頻度の影響についても調べるために,2値ベクトルと頻度ベクトルの両方について実験を行った。SVMのカーネル関数には2次の多項式関数を使用した。

7.2 実験結果

コンピュータ分野に関しては文字 bi-gram を除きいずれの記事セットでも９０％以上の分類精度が得られた（表3）。特に著者判定1は再現率で１００％となった。

一方,コンピュータ分野によって学習したモデルを使用してその他の分野の記事を分類した場合には,いずれの場合もF-measureで50%を下回っている。箇条書きの項目に横断的に現れるパターンを使用した場合のみ,再現率で50%を上回っている。

7.3 考察

コンピュータ分野に限れば,著者推定に用いられる特徴量は有効である。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各階層の箇条書きを1つの記事として切り出しているため,名詞や動詞などの内容語には箇条書きの間で共通するものが多いと考えられる。著者推定の特徴量は,このようなデータセットに対しても有効に働いている。異なる分野のデータセットに対しては,品詞や字種などの頻度情報だけを用いた著者判定1が有効に働いていないことから,単語の情報を特徴量に用いる必要があると考えられる。その他の分野の記事セットにおいて正例は１０％程度しか含

	コンピュータ(5-fold 交差検定)		コンピュータ/その他分野	
	2値	頻度	2値	頻度
文字　N=2	0.93	0.92	0.86	0.44
	0.88	0.83	0.19	0.23
	0.90	0.87	0.31	0.31
形態素 N=2	0.94	0.93	0.61	0.65
	0.89	0.91	0.30	0.38
	0.91	0.92	0.40	0.48
著者判定1	1.00	0.82	1.00	1.00
	0.89	1.00	0.03	0.09
	0.94	0.90	0.06	0.17
著者判定2	0.93	0.93	0.79	0.79
	0.91	0.91	0.17	0.17
	0.92	0.92	0.28	0.28
箇条書き	0.93	0.93	0.54	0.37
	0.91	0.91	0.31	0.50
	0.92	0.92	0.40	0.43

表3　分類結果　　上段 : precision　中段 : recall　下段 : F-measure

まれていないことから bi-gram をベースにした特徴量については箇条書きの抽出に有効に働いていることがわかる。

8．まとめと今後の課題

Web 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調べた。

キーワードとして「手順」を含むコンピュータ分野の Web ページについては,著者推定の特徴量を用いて高い精度で箇条書きの分類が可能であることを示した。またコンピュータ分野の箇条書きから得られた特徴量を用いて異なる分野の箇条書きを分類する場合には,bi-gram に加えて箇条書きの項目に横断的に現れるパターンを利用することで,分類精度を向上できる可能性を示した。

現在さらに実験の精度を向上させるため,評価用記事セットの記事数を４０００件程度にまで増やして実験を行っている。今後は,名詞タイプの箇条書きに絞った抽出精度の向上,箇条書きの前に現れる特徴を利用した分類を考えている。

謝辞

Tiny-SVM,PrefixSpan をご提供頂いた奈良先端大の工藤拓氏に深く感謝致します。東京工業大学大学院田中・徳永研究室及び奈良先端大の高橋哲朗氏ほか松本研究室の皆様のご協力に深く感謝致します。貴重なご助言とご協力を賜りました法務・知的財産権本部富士原裕文氏,富士通研究所颯々野学氏,松井くにお氏並びにドキュメント研究部の皆様, 内野寛治氏,菊田泰代氏並びにＤＢサービス部の皆様,富士通（株）の窪田伸一氏,窪山庄一氏,足立顕氏並びに情報メディアシステム事業部の皆様,に深く感謝致します。

参考文献

[1] Moldovan, Dan and Pasca, Marius and Harabagiu, Sanda and Surdeanu, Mihai Performance Issues and Error Analysis in an Open-Domain Question Answering System Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL),pp.33-40,2002

[2] 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,pp.181-198,1999

[3]Sadao Kurohashi and Wataru Higasa Dialogue Helpsystem based on Flexible Matching of User Query with Natural Language Knowledge Base Proceedings of 1st ACL SIGdial Workshop on Discourse and Dialogue,pp.141-149,2000

[4] Marti Hearst Multi-Paragraph Segmentation of Expository Texts Proc. of the 32nd Annual Meeting of the Association for Computational Linguistics,pp.9-16,1994

[5] 望月 源, 岩山 真, 奥村 学 語彙的連鎖に基づくパッセージ検索 自然言語処理,pp.101-126,1999

[6] 奥村学, 難波英嗣 テキスト自動要約に関する研究動向 自然言語処理 「テキスト要約のための言語処理」特集号 Vol.6, No.6,pp.9-16,1999

[7] 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

[8] Yuta Tsuboi, Yuji Matsumoto Authorship Identification for Heterogeneous Documents 情報処理学会研究会報告 NL-148-3,pp.17-24,2002

[9] Jian Pei, Jiawei Han, and et al. Prefixspan: Mining sequential patterns by prefix-projected growth In Proc. of International Conference of Data Engineering,pp.215-224,2001

[10] Nello Cristianini, John Shawe-Taylor An Introduction to Support Vector Machines, Cambridge Univ. Press, 2000

[11] 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 奈良先端科学技術大学院大学,1999

[12] Rakesh Agrawal and Ramakrishnan Srikant. Fast algorithms for mining association rules. Proc. 20th Int. Conf. Very Large Data Base (VLDB),pp.487-499,1995

[13]TakuKudo,Tiny-SVM:<http://cl.aist-nara.ac.jp/~taku-ku/software/TinySVM/>, 2001

[14]TakuKudo,PrefixSpan:<http://cl.aist-nara.ac.jp/~taku-ku/software/prefixspan/>, 2001