文判別タスクのための特徴的形態素パターン自動抽出手法の提案

小林竜己

株式会社ＣＳＫ　情報システム本部
Tatsumi_Kobayashi@cii.csk.co.jp

1　はじめに

近年、電子メールや Web メールなどによる問合せが一般化し、大量の問合せを受ける企業や組織では、効率的対応や状況分析などのため、蓄積された問合せを精度よく分類したいというニーズが高まっている。しかし、電子メールのような比較的短い文章サイズの文書を適切に自動分類することは非常に難しい。例えば、ベクトル空間モデルに基づく文書分類では、文書内のすべての語句の特徴ベクトルが用いられるため、文章が長く、語句が多い場合には良好な分類結果が得られるが、文章が短く、語句が少なくなると分類を誤るケースが増えてくる。

一方、人が文書分類を行う場合を考えてみると、領域知識や文脈などを利用し、分類の対象とならない文を適切に除外し、残された重要そうな文やパッセージに基づき最終的な判断を下すという方略がとられることが多い。この事実に基づき、本研究では文書自動分類の基盤技術として、文に対する判別タスクを考える。特に、今まで見落とされがちだった質問や依頼、症状の説明などといった文の抽象的概念クラスの判別に対する解決を目指す。具体的には、今までは名詞（コンテンツワード）の影に隠れて見落とされがちだった文末述語に着目し、類似文集合から「名詞」と「文末述語」の特徴的形態素パターンの二種類の特徴量を自動抽出する手法を提案する。そして文末述語を名詞にブレンドした特徴パターンを用いることで、抽象度の異なるクラスの文判別に柔軟に対応できる可能性があることを示す。文判別技術が確立されることで、文属性を利用した新しい文書自動分類への展望が開けるのではないかと考える。

2　特徴的形態素パターンの自動抽出手法

2．1　文からの特徴自動抽出の研究

大量の言語データから使用頻度の高い表現や固定的な言い回しを自動抽出する研究として、任意の長さ以上で任意の出現回数以上の文字列を漏れなく自動抽出する、n-gram 統計に基づく方法が提案されている[1]。この文字列抽出法では、文全体で文字単位の組合せが行われるため、文法的・意味的に表現の単位をなさない断片的な文字列の抽出が大量に行われてしまうことが報告されている。また抽出される文字列の内、離散型共起表現は文全体での順序が保たれたものであるため、本研究で扱うような類似文集合内での共通の特徴とするには長すぎると考えられる。

一方、大規模な文書集合から、頻度の高い、係り受け関係にある動詞と名詞の組み合わせ（構造化相関パターン）を自動抽出するアルゴリズムの研究がある[2]。この手法は、本研究と同じく問合せメールなどの自動分類のために開発されたものであるが、動詞と名詞の組合せの抽出であるために、文末の助述表現に特徴を持った抽象的概念の判別が困難であると考えられる。また一旦抽出されたパターンはその文書集合内で特徴的である可能性が非常に高いが、単語の共起を扱うため、実際の文書分類においては、精度が高い反面、類似表現の拾い漏れが生じ、再現率が低くなる恐れがある。

2．2　提案手法の特徴

本研究では、文に含まれる特徴量を名詞および文末述語パターンの集合であると定義し、人が任意の観点で類似性を認めて収集した文集合から、高い頻度で出現する名詞と文末述語の形態素パターンの自動抽出を目指す。名詞の他に文末述語を使う理由には、文末述語に質問文、依頼文、事態の説明文などといった文の抽象的クラスを識別するための固定的な表現や言い回しが存在すると考えるからである。別の理由としては、文全体を対象とする文字列パターンでは長すぎ、また係り受け関係にある動詞と名詞では特徴の拾い漏れがあると考える。さらに形態素の利用により、文法的・意味的に正しくない位置での文字列分割が回避できると考える。

以下に、今回考案した文末述語部分からの特徴的形態素パターン抽出の考え方を説明する。名詞の抽出も類似の考え方を適用する。

まず、Ｎ個の類似した事例文（正事例文と呼ぶ）すべてを形態素解析し、文末述語部分集合Ｅが特定できたとする。集合Ｅに出現する全形態素ｋ個を集め、形態素集合Ｍ＝$\{m_1, m_2, \cdots, m_k\}$ および各形態素毎の出現頻度

$Mfreq(m_i)$, $(1 \leqq i \leqq k)$を 得る。集合Eの要素eは形態素のシーケンスパターンである。形態素の出現頻度は一つの文末述語e内に複数存在する場合、1つと数える。

ここで出現頻度 $Mfreq(m_i) \geqq q$ の形態素集合M_qのみで形成される2つ以上の隣接した形態素列（パターン）集合P_qを集合Eから抽出する。M_qに含まれる形態素の列がM_q以外の形態素の列を左右から取り囲んでいる場合、囲まれた部分をアスタリスクで置きかえる（離散型共起表現の取得）。出現頻度qで抽出されるパターン集合P_qは最大でもたかだかN個である。出現頻度をNから1になるまで小さくしつつパターンを取得した時、取得される全パターン$\{P_q, P_{q-1}, \cdots, P_1\}$の個数は最大でもN×Nである。複数の同一パターンは1つと数える。通常、自然文では類似の文を集めてもqの最大値はNより小さい。

この抽出手法の基本的考え方は、

①頻出パターンは必ず頻出する形態素を構成要素として持たなければならない。

②あるパターンに含まれる部分パターンは必ずその上位のパターンに含まれるため、抽出しても意味がない。ただし、その部分パターンが単独で他の事例文に存在する場合は抽出される。

の二つである。上の①、②は高頻出パターンであるための必要条件であるが十分条件ではない。そのため各パターンが事例全体からどれだけサポートを得られているか（パターンの出現頻度）を調べる必要がある。

この手法には、文末述語の集合全体で出現頻度が同一の形態素で構成されるパターン（ABC）があった時、実際は[（AB）の頻度]>[（ABC）の頻度]であるパターン（AB）を取得できない欠点があるが、このケースは事例数が多くなればなるほど確率的に減少する。また（ABC）抽出後に、人手により（AB）を分離することも可能である。さらに、この手法はシーケンス（文末述語の長さ）が短く、アイテム（形態素）の種類が多く、1シーケンスに同一アイテムを極力含まない傾向のあるデータ集合に適用するのがよいと考えられる。DNAのようなアイテムが4種類で、シーケンスが極端に長い対象には適用できない。

以上で説明した手法はアイテムの出現頻度に基づきパターン抽出を行うことから、**FPEM** (Frequency-based Pattern Extraction Method)と呼ぶことにする。

2．3 文判別のためのパターン抽出の手順

（1）事前準備

文判別によって文に付与されるカテゴリーの集合を定義する。次に、定義した各カテゴリーに該当すると思われる事例文を収集する。前後でカテゴリーが異なる複文は、接続助詞（または読点）の後ろで分割する。

（2）文末述語からのパターン抽出フェーズ

【ステップ1】

収集した文すべてを形態素解析を行う。そして、文末から文頭方向に探索し、格助詞、係助詞、副助詞、読点が発見された箇所の後ろ以降を文末述語の開始とみなす。このルールにより、接続助詞で結合された複数の述語や副詞は文末述語に含められる。

【ステップ2】

文末述語に出現する形態素をすべて収集する。なお、句読点は残すが、括弧などの記号、数値、アルファベット1文字、人名は除外する。

【ステップ3】

ステップ2で収集した各形態素の出現頻度を求める。同一の文に複数回出現する場合は1回とカウントする。

【ステップ4】

抽出された中で最も高い出現頻度の形態素集合を用いて、一つの事例文の文末述語内で2つ以上連続し、かつ、全角5文字以上18文字以下の形態素パターンを取得する。もし同一の文末述語中に最大全角5文字以内で離れて存在するパターン（離散型共起表現）があれば、その間をアスタリスクで埋めて取得する。

【ステップ5】

頻度を1つづつ減少させて1になるまで、ステップ4の処理を繰り返す。これにより得られるパターン集合を「全パターン」と呼ぶ。

【ステップ6】

収集されたパターンすべての出現頻度を事例文の文末述語集合を探索して求める。独立パターンを取得する場合のみステップ7へ進む。

【ステップ7】

ステップ6で得られた出現頻度に基づき、あるパターンの部分パターンで、かつ上位パターンと比べ出現頻度がそれ以下のパターンをすべて削除する。これにより得られるパターン集合を「独立パターン」と呼ぶ。

（3）名詞のパターン抽出フェーズ

名詞を「非自立、接尾、代名詞、副詞可能、形容動詞語幹でない名詞および未知語」と定義する。名詞の抽出は（2）のステップに以下の変更を加えたものである。ステップ5以降は（2）と同様である。

・名詞は、文末述語を除いた文部分から抽出する。

・連続する形態素パターン（複合名詞）は全角2文字以上10文字以下とする。

・離散型の共起表現は取得しない。

（4）パターン選択フェーズ

パターン選択フェーズでは、（2）、（3）で抽出したパターン集合から特徴的なパターンを選択する。

あるカテゴリーc での正事例文集合におけるパターンiの出現確率を$p_{OK}(c,i)$、負事例文集合（カテゴリーc 以外のすべて）における出現確率を$p_{NG}(c,i)$とする。また、パターンiがn個のカテゴリーの特徴パターン集合に含まれる時、Cfreq(i)＝nと記述する。今回、次の3つの異なるパターン選択手法を試み、評価を行う。

手法1：Cfreq(i)＝1 (1)

手法2：$p_{OK}(c,i) / Cfreq(i) \geqq \gamma$ (2)

手法3：$p_{OK}(c,i) \geqq \alpha$、かつ $p_{NG}(c,i) \leqq \beta$ (3)

カテゴリー毎に抽出された各パターンiに対して、上の手法に示す式を満足する場合のみ、特徴パターンであると見なし、満足しない場合はそのパターンを抽出対象から除外する。なお、定数α、β、γは、パターン抽出に用いた事例文集合に対して良好な再現率と精度が得られる数値を試行錯誤的に求める。

3 実験

3．1 実験環境

実験システムはWindows98 上でjperl により構築し、形態素解析ツールは茶筅(ChaSen) version 2.1 for Windows（奈良先端大 松本研究室）を用いた。

また社内システムなどに関する社員用問合せ窓口への問合せ電子メールの過去2年のログの中から、パターン抽出用の事例文書を６２６文書、オープン試験用検証文書６２８文書をランダム収集した。

表1 目的別カテゴリーとラベル付けした文の数

	事例文書	検証文書
質問	267	284
症状・問合せ理由	509	537
操作・設定	180	163
作業依頼	674	692
お礼・謝罪	44	55

表2 対象別カテゴリーとラベル付けした文の数

	事例文書	検証文書
システムA	84	73
システムB	33	41
システムC	34	17
システムD	420	310
システムE	252	314
システムF	148	163
プリンタ	33	63
ウィルス	287	212
停電対応	19	28

※「システムA」等は固有のシステム名称を伏せたものである。

そして、文判別のカテゴリー体系として、抽象的文クラスである「目的別」と、名詞に特徴があると考えられる「対象別」という性質の異なる2つを作成し、事例文書、検証文書に含まれるすべての文に対してカテゴリーのラベル付けを行った。その結果を表1、表2に示す。

3．2 特徴パターンのマニュアル抽出による文判別

文判別の自動化の実験に先立ち、人が事例文書から抽出した特徴パターンを用いて、検証文書の文に対する判別の程度を調べてみた。再現率は目的別、対象別カテゴリーのどちらも６０％程で、一方、精度は目的別が５４％、対象別が７１．５％だった。抽出パターン数はカテゴリー平均で、目的別が７７個、対象別が２１個だった。

3．3 パターン自動抽出による文判別

提案手法に基づき、事例文書から特徴パターンを自動抽出し、検証文書の文判別（オープン試験）を行った。抽出した「全パターン」集合に2．3（4）で示した異なる3つのパターン選択の手法を適用して特徴パターンを取得した時のオープン試験結果を図1に示す。

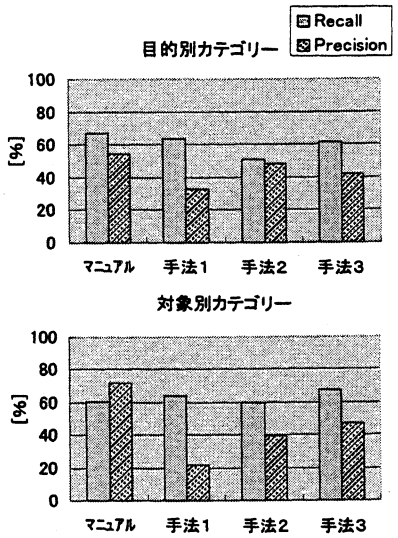

図1 各手法別の再現率と精度 (オープン試験)

また、パターン選択ー手法3を異なるパターン集合に適用した場合の再現率、精度を図2に示す（なお、マニュアル：マニュアル抽出パターン、全パ：全パターン、独立パ：独立パターン、名詞：全パターン内の名詞のみ、文末述語：全パターン内の文末述語のみ）。そのときの最終的に得られた特徴パターンの名詞と文末述語の含有率（全カテゴリー平均）を表3に示す。また抽出した「全パターン」の数が選択フェーズにより減少する程度（全カテゴリー平均）を表4に、取得した特徴パターンの一例を表5に示す。

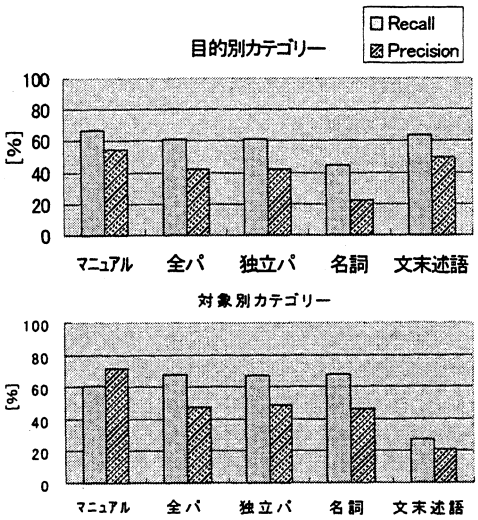

図2 各パターン集合別の再現率と精度 (オープン試験)

表3 特徴パターンの名詞と文末述語の含有率 (手法3)

	名詞 (%)	文末述語 (%)
目的別	43.6	56.4
対象別	77.7	22.3

表4 全パターン抽出後と選択後の平均パターン数変化

	全パターン抽出	手法1	手法2	手法3
目的別	831.0	402.8	6.6	20.2
対象別	407.4	222.9	7.6	42.9

表5 特徴パターンの一部例 (全パターン抽出、手法3)

質問	質問, 対処, 方法, バージョンアップ, 回答, 手順, Windows, でしょうか。, でしょうか?, ですか?, のでしょうか?, 教え下さい。, 教えて下さい。, 教えてください。, すれば*でしょうか?, あります。, 可能でしょうか?, ないでしょうか?
プリンタ	プリンタ, プリンター, トナー, 印刷, 紙詰まり, 汚れ, 支障, 出力, 部分, ピン, ERROR, ロール部分, エプソンLP, 故障してしまいました。, 出しております。, ておりません。, かすれる。

3．4 考察

マニュアル抽出した特徴パターンでの文判別は、かなりよい結果が得られた。人の場合、知識に基づく不要語排除や類義語追加が有効に働くためと考えられる。

一方、提案手法による文判別結果を見ると、手法3の成績がやや良かった。しかし再現率はマニュアル抽出のレベルに近いが、精度が低い。これは試行錯誤的に(3)式のα、βを求める際、まず事例文集合に対する再現率が（可能であれば）５０％以上のα、βの組み合わせを見つけ、その後、精度が高く、パターン数が多くなるように調整したためである。パターン自動抽出の研究では、取得するパターンの最低出現頻度が重要なパラメータであるが、良好な値の発見は人の手にゆだねられている。このことは手法3も同様であり、自動的な最適パラメータ調整の考案が望まれる。なおα、β値で高い再現率、精度を探索すると、異なるパターン数で良好な解が存在するが、なるべくパターン数を多くすべきである。

さて、表3を見ると、目的別カテゴリーでは名詞よりも文末述語が数多く抽出され、対象別カテゴリーではその逆となっている。これは提案手法によって、文末の述語部分に存在する抽象的な文クラスの特徴量が抽出されたことを示している。図2で目的別カテゴリーを見ると、文末述語のみの文判別でマニュアル抽出と同程度の成績が得られた。しかし適用する領域の性質によっては文末述語のみの特徴が適切であるとは限らないので、名詞と文末述語をブレンドした状態（全パターン）が、任意の文書集合やアドホックなカテゴリーに対して高い特徴量を保てるのではないかと考える。そのためには、表5に示すように現状では「出力」や「部分」などの一般語が特徴パターンに混在するため、不要語リストなどを用意するなどして特徴パターンの精度を高める必要がある。

なお図2から、独立パターン、全パターンを用いた文判別結果は、ほぼ同じであることが分かった。

4 おわりに

本研究では、類似文集合から特徴的な名詞と文末述語パターンの集合を抽出する手法を提案し、実験を行った。今後は、抽出パターンを絞り込むための効果的手法の考案と同時に、不要語リストの利用や異なる抽出パターン形式（“N＋の＋N”や他の文法構造）の可能性を検証し、さらに精度の良い特徴パターン抽出を目指す。そして文判別を利用した文書自動分類に着手する予定である。

謝辞

本研究を進めるにあたり、東京工業大学 徳永健伸助教授、ならびに研究室の方々に有益なコメントを頂きました。深く感謝申し上げます。

参考文献

[1] 池原 悟, 白井 諭, 河岡 司: 大規模日本語コーパスからの連鎖型および離散型の共起表現の自動抽出法, 情報処理学会論文誌, Vol. 36, No. 11, pp. 2584-2596 (1995).

[2] 松澤裕史: 大規模データベースからの頻出構造化パターンの抽出, 情報処理学会論文誌：データベース, Vol. 42, No. SIG 8 (TOD 10), pp. 21-35 (2001).