文脈的特徴を用いたランキング学習によるブログからの主題抽出

川場 真理子† 平野 徹† 松尾 義博† 菊井 玄一郎†

†日本電信電話株式会社 NTT サイバースペース研究所
〒239-0847 神奈川県横須賀市光の丘１－１
E-mail: †{kawaba.mariko, hirano.tohru, matsuo.yoshihiro, kikui.genichiro}@lab.ntt.co.jp

あらまし 本研究は，人・物・場所などの具体物が主題となるか否かを判定し，具体物が主題となる場合は，それが何かを抽出する事を目的とする．ここで主題とは，テキストを「X について書かれたテキスト」と言い換えた際の X とし，テキストの内容を最も含意するような言葉とする．また，人・物・場所等の具体物を指す名詞句が主題となる場合，それを具体主題とする．本稿では，具体主題を抽出する課題を，【①具体主題が存在するテキストか否かの分類】と【②テキスト中の具体主題の抽出】の 2 つのステップに分け，【②テキスト中の具体主題の抽出】を行う手法を提案し，評価を行った．具体的には，テキスト中の名詞句をランキング学習によって並び変え，1 位のものを具体主題として抽出した．著者が強調している名詞句及び，テキスト中の他の語より，多くの情報を持つ名詞句は主題になりやすいと仮定し，日本語の語彙的特徴及び大規模テキストからの係り受け関係を学習の素性として利用した．その結果，これらの素性を加えなかった場合と比較して，性能が向上したことを確認した．

キーワード 情報抽出，ブログ，主題

Topic Term Extraction Based on Learning to Rank with Contextual Feature

Mariko KAWABA† Toru HIRANO† Yoshihiro MATSUO† and Genichiro KIKUI†

†NTT Cyber Space Laboratories, NTT Corporation
1-1 Hikarinooka Yokosuka-Shi Kanagawa 239-0847 Japanan
E-mail: †{kawaba.mariko, hirano.tohru, matsuo.yoshihiro, kikui.genichiro}@lab.ntt.co.jp

1. はじめに

ブログや SNS の普及により，多くの情報を手軽にウェブから得られるようになった．一方で，膨大な量のウェブページの中から自分の関心のある情報に到達することが困難になっている．そこで近年，ウェブ上のテキストから複数のキーワードを抽出する技術が重要視されている．抽出されたキーワードを使って目的とするテキストを検索することができ，さらに，テキストに書かれたおおまかな内容を理解する手助けをすることができる．しかし，キーワード抽出は複数のキーワードをテキストから抽出するため，特に『どのキーワードがテキストの核となるか』が分かり辛いという問題がある．そこで，本稿では，ウェブ上のテキストに「主題」を付与することで，各テキストが何について書かれているものかを明確にする．

主題とは，テキストが何について書かれたものかを表し，テキストの核となる言葉である．主題の種類は様々だが，本稿では主題を「具体物を表す具体主題」「動作や経験を表す動作主題」「思想などを表す抽象主題」の 3 つに分類した．具体主題は名詞句としてテキスト中に現れやすく，動作主題は，名詞句や動詞句としてテキスト中に現れる．また，抽象主題はテキスト中に言語表現が現れない場合がある．本稿では，主題抽出の最初のとりかかりとして具体主題を抽出する．

以下に本稿の構成を述べる．第 2 章では本研究の対象である主題について分析するとともに，課題について整理する．第 3 章では，第 2 章で述べた課題を解決する為のアプローチと提案手法について述べる．第 4 章では，提案手法の評価を行い，考察をまとめた．第 5 章では関連研究について述べ，第 6 章ではまとめと今後の展開について述べる．

2. 課題

2.1. 主題

主題とはテキストを「X について書かれたテキスト」と言い換えた際の X とし，テキストの核となる言葉である．

また，常に同じ言葉が主題になるわけではなく，テキストごとに，テキストで主張されている物を最も適切に表わす言葉を選択する．図 1 のテキストでは「チョコレート」そのものについて語られており，下線部の「トリュフ」や「パレファン」は「チョコレート」の中に含まれる情報である．そのためこの場合「チョコレート」が主題となる．一方，図 2 は「チョコレート」そのものについての記述はなく，「トリュフ」について記述されている．このような場合の主題は「トリュフ」

になる．

主題は，係り受け等で様々な補足や説明がなされる．たとえば図 3 のテキストの主題は「ギャラクシー」だが，「ドコモの」「マンガビューアーとして利用する点においては現時点で最強のマシン」のような補足・説明がなされる．

本稿では主題を整理し，以下の 3 つに分類した．

① 人や物，場所などの具体主題[図 3]
　例．木村拓哉，テレビ，NTT 横須賀通信研究所
② 体験，経験などの動作主題[図 4]
　例．お買い物，旅行
③ 思想的，宗教的な抽象主題[図 5]
　例．人生

これらの主題のうち,①は名詞句として表れやすい傾向がある．②はテキスト中に名詞句や動詞句として表れる場合もあれば，言語表現として表れないこともある．また，③はテキスト中に言語表現として，現れにくい傾向にある．本稿では，主題抽出の最初のステップとして，テキスト中に名詞句として現れやすい，具体主題を抽出する．

今日は疲れたー．こういうときはやっぱり**チョコレート**！ポリフェノールは美容にもいいしね．おまけに脳にもいいらしい．嬉しい成分盛りだくさん！カロリー高めなのがちょっと残念！
この時期はデパートにもたくさんの**チョコレート**達…トリュフ，パレファン，生チョコ…．太らないように気をつけないとね．

図 1：チョコレートが主題のテキスト

バレンタインなので，町はチョコレート一色～．今日はお父さんに**トリュフ**を買いました．チョコレートの定番だけど，でも，これはちょっと変り種．ピンクのキャンディでデコレーションされてて超可愛いの．口に入れるとぱちぱち弾けるよ．
おいしー．

図 2：トリュフが主題のテキスト

ついにドコモのギャラクシー発売．
主にマンガデータを読む用途限定で買いました．なので，それ以外のことにはほとんど利用していません．マンガビューアとして利用する点においては現時点で最強のマシンだと思います．

図 3：具体主題【ギャラクシー】

今日は銀座に行きました！久々な上にバーゲンだったから超盛り上がっちゃった♪プラダでカバンを買って，シャネルで新色のアイシャドーをゲット！ディオールでサングラスを買って，ついでにジルのクリスマスコフレ予約してきちゃった！

図 4 :動作主題【買い物】

今帰ってきました．ぃや
明日の準備時間がかかるのにまだ何もしてないし今からぉ風呂だし．なのに明日は 4 時起き．今日は一睡もできないな．とりぁぇず早くぉ風呂はいろ…

図 5：抽象主題【主題不明】

2.2. 主題抽出における課題

ブログテキストから具体主題を抽出するには,次の 2 つの課題を解決する必要がある．これら 2 つの課題を解決するものを抽出する．

① 名詞句を最適な範囲で抽出する必要がある
　主題として名詞句を抽出する際には,単純に名詞をつなぎ合わせてもうまくいかない場合が多いため,サ変名詞などを考慮する必要がある．
② 同じ名詞句であっても,テキストの書かれ方,他の候補次第で重要度が変わるため,必ず主題になる名詞句が存在しない．IDF の値や，頻度のみで主題を選ぶことが難しい．

3. 提案手法

3.1. アプローチ

2.2 節で述べたように主題を抽出する為には，各テキストから名詞句を適切な範囲で候補として切り出し,テキスト内で候補同士の比較などを行う必要がある．

本研究では，名詞句を適切な範囲で切り取る為に,名詞句抽出器を作成し，主題の候補とした．さらに,主題の候補をテキストごとに比較するため，テキスト内で候補をランキングし，上位 1 位を主題として取り出す，というアプローチを取った．

3.2. 主題候補抽出器

2.2 節で述べたように，テキスト中の名詞を単純に連結して名詞句を作成する方法では適切な名詞句が抽出できない．そこで本研究では，名詞句の抽出ルールを作成し，形態素解析結果に適応した．また，具体的な名詞句は，人・組織・場所・もの，等のほかに名前のついたイベントも対象とする．抽出ルールを以下に述べる．

1 つ目のルールは,【名詞:形容】を品詞に持つ名詞句の扱いである．ルールを表 1 に示す．【名詞:形容】は『大好き』『最高』『ソフト』などの形容を表す名詞

である．単純に名詞につなげて取得してしまうと，『iPhone 最高です』という文から，『iPhone 最高』と抽出されてしまう．しかし，【名詞:形容】をすべて削除してしまうと，『パソコンソフトを買った』という文から『ソフト』が削除されてしまい，『パソコン』が抽出されてしまう．このような問題を解決する為，【名詞:形容】の品詞を持つ形態素の後に『が』もしくは『を』が現れれば，【名詞:形容】までを含めて 1 つの名詞句として取得し，それ以外の【名詞:形容】を品詞に持つ形態素は削除するようにした．

2 つ目のルールは【名詞：動作】を品詞に持つ形態素の扱いについてである．【名詞:動作】を品詞に持つ形態素は『マッサージ』『連絡』『購入』などがあり，動詞的に用いられる場合が多い為，『iPad 購入』や『連絡する』など名詞句として抽出しなくてよい場合が多い．しかし，『台湾式マッサージ来たよ』の『台湾式マッサージ』のように，【名詞:動作】の形態素まで含めてひとつの名詞句となる場合も存在する．そこで，【名詞:動作】の後に，【動詞語幹】が現れる場合のみ，【名詞:動作】まで取得する．

3 つ目のルールは，発話以外の括弧の内容を抽出するというものである．『東海道新幹線「のぞみ」に乗りたい』の文のように，特定の物や商品を指す用語が括弧の中に記述されることが多い．ただし，多くの場合，括弧の中の文字列は『カンヌ映画祭特別招待作品で「抱腹絶倒のユーモアと芸術性が共存」と高評価を得た』のように発話であるため，発話と用語を区別する必要がある．そこで，『括弧内の形態素情報』『括弧前後の 3 つの形態素情報』の 2 つの素性に基づき，括弧を発話か用語かに分類する分類器を作成した．この分類器は 2458 の学習データを 10 分割交差検定で評価した結果 F 値 0.95 という値を達成している．

表 1：ルール 1．名詞:形容の削除

ルール 1. 名詞:形容の削除
名詞:形容=(大好き，最高，ソフト)
・(<名詞>*)<名詞:形容> →名詞:形容前まで抽出 例：Iphone 最高です→Iphone
・(<名詞>*<名詞:形容>)(が|を) →が|をまでを抽出 例：パソコンソフトを買った →パソコンソフト

表 2：ルール 2．名詞：動作の削除

ルール 2. 名詞:動作の削除
名詞:動作=（マッサージ，連絡，購入）
・(<名詞>*)<名詞:動作>→名詞の連続まで抽出 例：Ipad 購入→Ipad
・(<名詞>*<名詞:動作>)<動詞語幹> →名詞:動作まで抽出 例：台湾式マッサージ来たよ →台湾式マッサージ
・(<名詞>*)<名詞:動作><動詞接尾辞> →名詞の連続まで抽出 例：連絡する→Null

表 3：ルール 3．発話以外の括弧の抽出

ルール 3. 発話以外の括弧『』「」の抽出
以下の素性に基づき括弧を発話か用語か SVM で分類した． ・括弧内の形態素情報 ・括弧前後 3 つの形態素情報 学習データ 2458 記事で学習 ※10 分割交差検定にて，F 値 0.95
発話=２００４年カンヌ映画祭特別招待作品で「抱腹絶倒のユーモアと芸術性が共存」と高評を得た 用語=東海道新幹線「のぞみ」に乗りたい

3.3. ランキング学習

最近ハマってるドレッシングがコレっ．

伊豆カメヤさんのわさびドレッシング！

以前，軽井沢のスーパーで見て美味しそうだなぁと思ってたんだけど，都内のスーパ　ーでも発見．すぐにかごに入れました(笑)

図 6：表現の特徴の例

今日はピエールエルメの**トリュフ**を買ってきました！
さすが有名店．カリっとした外側のほろ苦いチョコレートの内側にはトロットしたクリーム．絶妙でした！珈琲にも合うし，形が可愛いからプレゼントにも最高です．

図 7：記事内での優位を利用した例

主題の候補は 1 つのテキストから複数現れる．また，同じ名詞句であってもテキストでの書かれ方によってあるテキストでは主題だが，あるテキストでは主題でないという場合がある．本稿では，同じテキスト内の主題の候補を比較し，主題として最も適切な物を 1 つ取得する．

学習に利用する特徴として，表現の特徴と記事内で

の候補同士の優位差などを利用した．ブログ等のテキストでは，著者が主張したい事柄は括弧囲みや，体言止めなどで表現される場合が多い．図 6 の下線部では，著者が主題である「わさびドレッシング」を目立たせるために体言止めを用いている．また，主題となる名詞句はテキスト内の他の候補と比較して，より多くの説明・補足がなされる．そのため，係り受け関係で係り先になるものは主題になりやすい．候補同士を比較した時に係り先になりやすい名詞句は他の候補と比較して優位であるとし，学習の特徴として利用した．図 7 の下線部では，「ピエールエルメ」が「トリュフ」に係っており，「トリュフ」の方がより主題らしいと言える．

学習の特徴として利用した素性を以下に示す[表 4].

A) 表現の特徴
著者が強調するような書き方をしている名詞句はテキスト内で重要な語句であり，主題になる可能性が高い．そこで，著者が文章を書く上で，特定の表現を目立たせるのに利用する，体言止め，括弧囲み，候補の名詞句だけで 1 文使っている，などの表現の有無を利用した．

B) 記事内での優位差
テキスト内の候補のペアを作成し，係り受け関係や共起関係などで，どちらが優位であるかを比較した．『*A* の *B*』という表現がある時に，係り先である B の方が主題になりやすい．そこで，大規模なブログデータから係り受け情報を取得した．さらに，各候補のペアの比較を行い，係り先になる頻度が多い方を勝者，頻度が少ない方を敗者とした．また，各テキスト内での係り受け関係のみでの比較も行った．また，『*A*』と『*B*』の 2 つの表現がある場合に，*A* が出現したときに B が出現する確率が高ければ，A の方が B よりも主題になりやすいと考えられる．そこで本稿では，それぞれ $P(A|B)$と $P(B|A)$を求め，$P(A|B)<P(B|A)$ならば，*A* の方が主題になりやすいので *A* の勝ちとした．例えば，『犬』と『動物』という名詞句があった場合，犬が出てくるテキストに動物が出てくる確率と，動物が出てくるテキストに犬が出てくる確率を比較したときに，前者の確率が高かったとする．その場合は，『犬』の方が主題になりやすいと考えられる．

C) 固有表現
主題となる名詞句は，固有表現である場合も多い．そこで名詞句が固有表現かどうかを考慮するため，Wikipedia の見出し語になっているかどうか，固有表現抽出器で取得できる固有表現かどうかを特徴として利用した．

D) TFIDF
他のテキストと比較して，対象とするテキストにより特徴的に表れる名詞句は，テキストの内容を表す語である場合が多く，主題となりやすい．そこで，特定のテキストに偏って多く表れているかどうかを判断する指標である TFIDF を特徴として利用した．

E) 名詞句のドメイン
本研究が対象とする主題は，物や人，場所などの具体的な物を表す名詞句である．具体的な物を表す名詞句が属するドメインは限られているため，ドメインを見ることで主題になりやすい名詞句かそうでない名詞句かを判断することが可能である．そこで，西田ら[7]の手法を用いて，テキスト中に現れる名詞句に『教えて goo！』のカテゴリ，及び日本語語彙大系のカテゴリを対応付けた．また，名詞句が固有表現であった場合は IREX の固有表現のクラスを対応づけた．

表 4 :素性一覧

素性	種類
体言止めで書かれている	A
1 文が 1 名詞句で終わっている	A
括弧で囲まれている	A
TFIDF1~N 位までの候補に対する係り受け関係の勝敗	B
TFIDF1~N 位までの候補に対する出現確率の勝敗	B
他の候補に係られているか否か	B
他の候補に係っているか否か	B
Wikipedia の見出し語になっているか否か	C
固有表現か否か	C
TFIDF	D
教えて goo!のカテゴリ	E
語彙体系の意味カテゴリ	E
固有表現の意味クラス	E

4. 評価と考察

評価には，収集したブログ記事の内，人手で具体主題ありと判定された 396 記事を利用した．あらかじめ抽出した主題の候補の名詞句を記事ごとにランキングし 1 位の名詞句を主題として抽出した．ランキングには SVM Light を用いた．また，ランキング 1 位として抽出された名詞句のうち，人手でつけた正解と一致している割合を正解率とし，評価尺度とした(1).

評価は，本稿で提案する，表現の特徴を用いた素性と記事内での優位性を用いた素性を除去した学習についてそれぞれ評価を行った．評価の一覧を表 5 に示す．

$$\text{正解率} = \frac{\text{1位にランクされ人手で主題と判定されたもの}}{\text{1 位として取得された名詞句の総計}} \quad (1)$$

表現の特徴や，記事内の優位差などを利用した場合は，利用しなかった場合と比較して，約 9%の性能向上が見られた．例えば，図 8 のテキストの主題はダノン BIO 野菜であるが,書き方などを考慮しない場合は,『ダノン』の方が出現するテキスト数も多く，かつ特定のテキストに多く出現するため，TFIDF などが高くなる．また Wikipedia 等の見出し語にもなりやすく，用語として適切であるため，主題として選ばれやすくなってしまう．しかし，書き方などの素性を加えることで,このテキストでより強調されている名詞句は『ダノン』ではなく『ダノン BIO 野菜』であるということが分かる．また，多くみられた誤りとして，共参照の問題がある．共参照とは異なる表現で同じ事柄を指す事である．図 9 では『僕の初恋をキミに捧ぐ』という候補と『僕キミ』が共参照の関係にあり，このテキストの主題は『僕の初恋をキミに捧ぐ』である．提案手法では，括弧囲みなどの強調されている候補の方が主題になりやすい．また,『僕の初恋をキミに捧ぐ』と『僕キミ』それぞれに特徴が付与されるため，主題になりやすい特徴が分散されてしまい，正確に主題を抽出できなくなるという問題点がある．この問題を解決する為には共参照の問題を解き，同じ事柄を指しているものの特徴をまとめるという処理が必要である．

表 5：主題抽出の評価

		利用した素性	正解率
1	提案手法	A, B, C, D, E	59%
2	比較手法	B, C, D, E	53%
3	比較手法	A, C, D, E	56%
4	比較手法	C, D, E	50%

健介ファミリーがただ今凝りまくってる食べ物はこれ．ダノンの『ダノン BIO 野菜』！なんではまったかというと…
ダノンもたくさん種類があるけど，野菜だよ．まずはあたしみたいに一週間分は買いだめ．でまた 1 週間分．これがベストだぜ．

候補：健介ファミリー/ダノン BIO 野菜/ダノン/…

正解：ダノン BIO 野菜
手法 1（提案手法）の出力：ダノン BIO 野菜
手法 4（比較手法）の出力：ダノン

図 8：提案手法のランキング学習結果

昨日，僕の初恋をキミに捧ぐの CM 見て，帰りに本屋さんをのぞいたら,『僕キミ』がずらっと並んでて．初めて僕キミが少女漫画だって知ったんだ．
携帯小説かと思ってた．

図 9：誤り事例

5. 関連研究

テキストに書かれている内容を抽出する研究は主に，トピック抽出と headline 生成がある．トピック抽出の研究は主に複数のテキストから複数のキーワードを抽出し，キーワードの集合をトピックとしている．主に，類似するテキストを集め，テキスト集合に偏って表れる複数の単語をトピックとするアプローチ[1][2][3]などが用いられる．しかし，これらの研究は検索の際の検索質問拡張や，クラスタリングのラベルとして用いることを目的としており，テキストに端的に何を書いてあるかを示すことを目的としていない.また，headline 生成は TFIDF などの値を利用して，入力したテキストに偏って表れる語句を重要語として抽出しそれを利用して headline を生成するアプローチ[5][6]や，テキスト内の文章を木構造で表し，不要な部分を削ることで簡潔な headline を生成するアプローチ[4]などがある．これらのアプローチで選ばれる重要語は，テキスト中で重要な語ではあるが，そのテキストに偏って表れる語が選ばれやすい為,『服』『スイーツ』などの多くのテキストに頻出する一般名詞を抽出する用途には不向きである.また,名詞句を抽出する研究として，専門用語を専門分野コーパスから自動抽出する研究[8]がある．これは，多くの言葉と連節するものを専門用語として抽出するため，我々の研究とは抽出の対象が異なっている．たとえば,「チョコレートパフェ」や「チョコレートアイス」という名詞句があるとき,「チョコレート」は「パフェ」や「アイス」と連節することができる為,「チョコレート」を専門用語として抽出する．我々の研究ではテキストの核となり，かつ具体性の高いものを主題として抽出するため「チョコレート」よりも,「チョコレートパフェ」や「チョコレートアイス」を主題として抽出する．

6. まとめと今後の予定

本稿では，ブログから名詞句を具体主題として抽出する課題を，ランキングを利用して解決するアプローチについて述べた．ランキングの素性として，テキストの表現の特徴や候補同士の比較を加えることを提案し，これらの素性を用いない場合と比較して，性能が向上することを確認した．

今後の課題は 3 つある.1 つめは主題の数を推定し，複数個の主題の取得を可能にすることである．また，2 つめは，経験などの行動に関する主題を取得することである.3 つめは，主題の有無を推定することである．本稿では具体主題が存在するテキストに限って評価を行ったが，具体主題が出現する記事としない記事を区別できれば，どのようなテキストに対してでも主題抽

出をすることが可能になる.

参 考 文 献

[1] 若木裕美, 正田備也, 高須淳宏, 安達淳, “検索語の曖昧性解消のためのトピック指向単語抽出および単語クラスタリング”, 情報処理学会論文誌データベース（TOD）, Vol49, No. SIG19(TOD32), pp. 72-85, 2006

[2] 橋本泰一, 村上浩司, 乾孝司, 内海和夫, 石川正道. “テキストクラスタリングによるトピック抽出および課題発見.”, 社会技術研究論文集, Vol.5, pp.216--226, 2008.

[3] 濱本雅史, 北川博之, Jia-Yu Pan, and Christos Faloutsos,” 独立成分分析を用いたテキストデータからのトピック検出”, 電子情報通信学会第 15 回データ工学ワークショップ(DEWS2004), 2004

[4] B. Dorr, D. Zajic, and R. Schwartz. Hedge: A parse-and-trim approach to headline generation.In Proceedings of the HLT-NAACL Text Summarization Workshop and DocumentUnderstanding Conference (DUC), 2003.

[5] 長安義夫, 山本和英. “タイトルパタンによるテキストの一文概要生成”. 言語処理学会第 13 回年次大会,pp.684-687 (2007).

[6] 廣島伸彰, 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,NL149-7,pp45-50,2002.

[7] 西田京介, 藤村考, "階層的オートタギングによるQ＆Aコミュニティの知識整理", 第 2 回データ工学と情報マネジメントに関するフォーラム(DEIM), D3-4, 2010

[8] 中川裕志,森辰則,湯本紘彰:出現頻度と連接頻度に基づく専門用語抽出, 自然言語処理 (2003)