社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2009－ＦＩ－ 93（18）
2009－ＮＬ－189（18）
2009／1／23

対話文生成のためのWebを用いた話題語の抽出

下川 尚亮　Rafal Rzepka　荒木 健治
北海道大学大学院情報科学研究科

本研究では，Internet Relay Chatの対話文の１文中で話題となっている話題語の抽出を行う．インターネット上の対話文は，blogやWebページに比べ，口語に近い表現が多数使用される．このことから，Web上の対話においてその入力文が意味するところを把握するには，ユーザの意図にそった入力文の解釈が必要となる．しかし，名詞だけを対象にユーザの入力文から，その文中で話題になっていることを把握ことは困難である．そこで本稿では，形容詞，動詞も考慮したWebからの話題語の抽出手法を提案する．実験の結果，名詞のみを対象にした場合よりも形容詞を考慮した場合の方が多様な表現の話題語を抽出できることが明らかとなった．

Extracting Topic Words from the Web for Dialogue Sentence Generation

Naoaki Shimokawa, Rafal Rzepka, Kenji Araki
Graduate School of Information Science and Technology Hokkaido University

In this paper we extract topic words from Internet Relay Chat utterances. In such dialogues there are many more spoken language expressions than in blogs or usual Web pages and we presume that the always changing topic is difficult to determine only by nouns which are usually used for topic recognition. In this paper we propose a method for determining a conversation topic considering also association adjectives and verbs retrieved from the Web. Our first experiments show that extracting association words using nouns and adjectives leads to determining topic labels of higher diversity.

１．はじめに

近年，インターネットの普及により，掲示板やチャットなどの対話文章に接する機会が多くなってきている．そして，対話文において，その対話文内で話題となっていることを把握することは重要である．しかし，対話文では現在話題となっていることを把握するために，その話題についての知識をあらかじめ持っていなければ内容を把握することが困難である．

また，各個人によって持っている知識には差があるので，まったく知らない話題について，適当な話題語を思いつけるとは限らない．たとえ，対話文中の知らない語をもとにWeb検索を行い，関連する語を調べたとしても，検索結果中のページ内から，対話文内に出現したキーワードを元に，重要単語を見つけ，キーワードに関する情報を収集する必要がある．

本研究では，これらの問題に対する１つのアプローチとして，対話文における話題語をWebから抽出することを考える．入力文には，本来複数の名詞，形容詞，動詞が含まれている．これらから，名詞のみを用いた場合，名詞と形容詞，名詞と動詞，名詞と形容詞と動詞を考慮した場合の４つのパターンについて，話題語の抽出を行う．文中の形容詞，動詞を考慮することにより，文の話題語として，より具体的な話題語の抽出を行うことができる．

本研究では，最終目標としてリアルタイム応答のための対話文生成を目標としている．そのため，少数のWebページから話題語の抽出を行う必要がある．そのため，話題語候補を取得する際のWebページ数は少なければ，少ないほどよい．このため，Web検索結果には，最小限のWebページを用いいることとした．

2．関連研究

話題となるキーワード抽出のための研究は多数提案されている．

Shinzato らの研究[1]では HTML 文書の構造を利用し，Table タグから下位概念の候補語の抽出を行い，それに対応する上位概念候補語を抽出し，上位概念と下位概念のペアから，上位概念と下位概念を抽出するものである．

Ando らの研究[2]では大規模コーパスから，下位概念を自動的に抽出するために，定型表現を用いて過去６年分の新聞記事から，下位概念を自動的に抽出し，連想辞書との比較を行い，ほぼ６割以上の正解率で下位概念の抽出を行っている．

山本らの研究[3]では辞書に記載されていない，語から新規に関連語を獲得し，辞書を用いないシスーラスの自動構築を行っている．

佐藤らの研究[4]では与えられた専門用語から，４種類のクエリを作成し，コーパスを作成する．そこから，重要文の抽出を行い，Web の Hit 数を用いてフィルタリングを行い，関連する専門用語の抽出行っている．

佐藤らの研究[5]では Web 上のニュース記事を対象に，情報の広がりと情報の伸びに注目し，文書をジャンルに分類し，コサイン類似度を用いて，文書間類似度の計算を行い，文書クラスタリングを行う．そして文書新鮮度と文書話題度の算出を行い話題語の抽出を行っている．

小山らの研究[6]ではユーザからの質問に対し，それをより詳細化するための Web からの話題語の抽出を Web ページ中のタイトルにキーワードが現れる場合とそうでない場合に区別し，主題を詳細化するための，キーワードの抽出を行っている．

稲川らの研究[7]では人物や作品などのオブジェクトに対して，Web の Hit 数と語の距離を用いて，クエリから候補語集合の抽出を行い，また逆方向の候補語集合からクエリの特定を行っている．

これらの研究では，名詞のみを対象として語の抽出を行っている．しかし，対話を目的とした文生成では，不十分であると考え，本研究では形容詞，動詞を考慮し，多様な表現の話題語抽出を行う．

表１：各品詞パターン

パターン１：名詞
パターン２：名詞，形容詞
パターン３：名詞，動詞
パターン４：名詞，形容詞，動詞

3．提案手法

本研究では，対話ログの入力文１文をそれぞれユーザの入力として，その入力文１文において，話題となっている語の発見を目標としている．以下のような手順で抽出を試みた．

ステップ１：検索クエリ候補語の抽出

ユーザの入力文１文から検索クエリに用いる語の抽出を行う．まず，入力文に形態素解析を行い，入力文に含まれる語の抽出を行った．本研究では，形態素解析ツールとして MeCab[8]を用いた．

入力文の形態素解析結果から，表１の品詞の組み合わせのパターンを用いて，語の抽出を行った．これらの語を検索クエリ候補語とする．なお，本研究ではこれらの品詞の組み合わせを以後品詞パターンと呼ぶ．名詞が連続している場合には連結し，複合名詞化して得た語を抽出語として扱う．また，ノイズ除去として，あらかじめ検索クエリ候補語の扱う名詞から非自立，代名詞を除外した．

[1]]2]では，名詞のみを扱っているが，本研究では名詞の他に形容詞，動詞を考慮した，話題語の抽出を行う．対話文においては特に形容詞，動詞を考慮することにより，文の意図が大きく変わってしまうこともある．例えば，入力文として「青春の頃の苦い思い出」を入力し品詞パターン１，つまり名詞のみを対象として形態素解析を行った場合，抽出される語は「青春」「思い出」となり入力文がどういった「青春」を意図しているのか明確ではない．しかし，品詞パターン２を用いて，形容詞を考慮することにより，入力文からは「青春」「苦い」「思い出」が抽出される．このように本研究では，入力文として対話文を対象としているため，名詞だけを対象として，入力文で話題となっている語の抽出を行うよりも，

形容詞，動詞を考慮することにより，ユーザの意図にそった話題語を抽出できると考える．よって，本研究では名詞だけではなく形容詞，動詞を考慮した話題語の抽出を行う．

形態素解析の結果から，検索クエリ候補語を組み合わせて検索クエリを作成する．検索クエリ候補語が２語以上の場合，作成される検索クエリの数が重複を除いて，式（１）のようになる．

$$q_{num} = \frac{x(x-1)}{2} \quad (x = 2,3,4...) \qquad (1)$$

入力文が「規則正しい生活、睡眠を多くとる」の場合，品詞パターン２を適用すると，検索クエリ候補語は「規則正しい生活」「睡眠」「多く」の３語となる．これらから検索クエリ候補語２語のペアを作成し，検索クエリとする．また，式（１）によりペアの作成数を決定する．検索クエリは「"規則正しい生活" "睡眠"」「"規則正しい生活" "多く"」「"睡眠" "多く"」のようになる．これらの検索クエリを用いて Web 検索を行う．検索エンジンとして，Yahoo! JAPAN が API を提供している Yahoo!検索 Web サービス[9]用いた．

Web 検索結果から得られる URL から上位１０件の URL を取得した．今回はスニペットを用いずに，取得した URL から直接 Web ページの取得を行った．これは複数の語からなる検索クエリを投げた場合，スニペットには分散して，検索クエリ候補語が表示されるためである．

ステップ２：話題語候補語の抽出

取得した Web ページに対して，形態素解析を行い，Web ページに含まれる語の抽出を行う．抽出する語の対象は入力文から語の抽出を行ったときに用いた品詞パターンと同じ品詞パターンを適用する．抽出する文は Web ページ中で，検索クエリ候補語が含まれる文のみを対象とした．これは Web ページ全体を対象としてしまうと，関連がない語が多く抽出されノイズとなるためである．

抽出した話題語候補を頻度の多い順から上位５０語を話題語候補とする．また，検索クエリが複数ある場合は，１つの検索クエリに対して得られる話題語候補の上位５０語同士を合計し再度，頻度の多い順から上位５０語を話題語候補とした．

表２：比較正規表現パターン

.*など
.*などの
.*に似た
.*のような
.*以外の
.*という
.*と呼ばれる

ステップ３：関連度計算

話題語の抽出においては，語の関連度をどのように定義するかが重要である．本研究では，Web ページから得られた話題語候補 T と検索クエリ候補語 Q の関連度を，Web 検索から得られる Web の Hit 数を用いて，式（２）のように定義する．また，Web ページに含まれる語の頻度は関連度を計算する上で重要であるため，語の頻度 tf を考慮したものを式（３）に示す．そして，取得した話題語候補語の中で，特徴的な語を抽出ために，tf・idf を考慮したものを式（４）に示す．以上の３式を用いて，語の関連度の計算を行った．

$$Jaccard = \frac{Q \cap T}{Q \cup T} \qquad (2)$$

$$Jaccard\_tf = tf \times \frac{Q \cap T}{Q \cup T} \qquad (3)$$

$$Jaccard\_tfidf = tf \times idf \times \frac{Q \cap T}{Q \cup T} \qquad (4)$$

ただし

$$idf = \log_2 \frac{N}{1+df}$$

とした．

ここで tf は語の頻度，df は Web の Hit 数を用いた．Web の Hit 数が 0 件のときを考慮し，ここでは 1+df とした．入力文ごとに取得する Web ページの数を合計し，全 Web ページ数を N として用いた．このようにして，入力文が

与えられるごとに tf・idf の計算を行った．

これらの式によって，計算された値の大きい方から順に上位５０語を話題語候補とした．

４．実験

本実験では，IRC の対話ログ(874865 文，約 58MB)から無作為に抜き出した 20 文を使用した．この 20 文を入力文として，語の関連度を計算し，ランキングされた上位 3 語を話題語とした．

４．１　比較方法

他研究との比較として，Ando らが用いた手法との比較を行った．Ando らは構文パターンを用いて，語の抽出を行っているが，本研究では正規表現パターンを用いて，文の表層的な情報だけを利用した．入力文に対して Web から取得した Web ページの HTML 文書に対して，正規表現パターンを適用し，語の抽出を行い，頻度順に並び替え上位 3 語を話題語とした．本研究で用いた正規表現パターンを表２に示す．

４．２　実験結果及び考察

主観評価のための被験者は１０代男性３名，１０代女性７名，２０代男性７名，２０代女性３名の大学生に対して行った．評価には入力文と，取得できた話題語を提示し，入力文と話題語の関連の強さを５段階の評価で行った．

評価には式（２）～式（４）に加え，本手法と比較のために Ando らの手法を加えた，４つの場合においての各品詞パターンに対して行った．２０人中１５人が５段階評価で３以上をつけた話題語を正解とした．これを以後関連語と呼ぶ．Ando らの手法を Alter として，２０人中１５人が関連があると判断した結果を図１に示す．図１から，品詞パターン１において，Jaccard_tf を使用した場合，約８８％の精度で話題語の抽出が行うことができた．すべての品詞パターンにおいて，Jaccard_tf は他手法と同等，または高い精度が得られた．このことから Jaccard_tf は，品詞の種類に関係なく，有効に作用することがわかる．これより，対話文における話題語の抽出において，Jaccard 係数に語の頻度 tf を

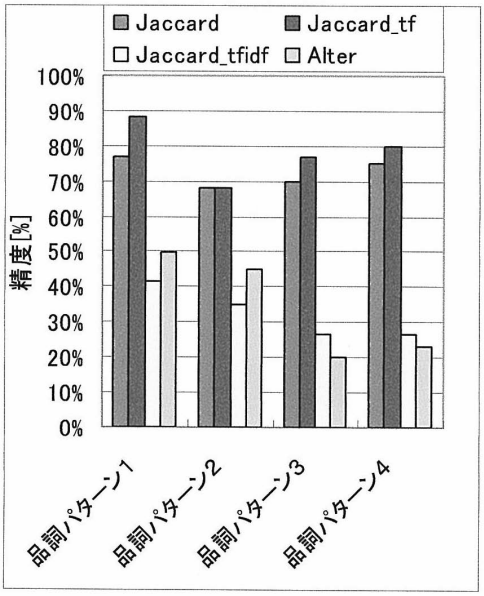

図１：各手法における精度

用いることは有効であることが確認された．

形容詞を考慮した場合は約６８％，動詞を考慮した場合は約７６％の精度で話題語の抽出が行えた．品詞パターン２，３，４は，名詞のみを考慮した品詞パターン１よりも２０％低い精度となった．

上位３語の話題語のうち，検索クエリ候補語を含まない語の割合を図２に示す．

図２に，話題語上位３語のうち，検索クエリ候補語ではない語が，含まれている割合を示す．図２より，Jaccard_tf の品詞パターン２において，約７０％の精度で検索クエリ候補語以外の語が話題語として抽出されていることがわかる．このことから，対話文において形容詞を考慮することにより，名詞のみを対象とするときよりも，多様な表現の抽出が行えていることがわかる．特に，式（２），式（３）を用いて関連度の計算を行ったときに，品詞パターン２において，他の品詞パターンにくらべて，多くの割合で検索クエリ候補語以外の語の抽出が行えていることがわかる．また，式（４）を用いた場合において，各品詞パターンにおける差がほとんどなく，どの品詞パターンにおいても，高い割合で検索クエリ候補語以外の語の抽出が行われている．これは，式（４）により，話題語候補語の中から，特徴的な語の影響が強く働いたためだと考えられる．

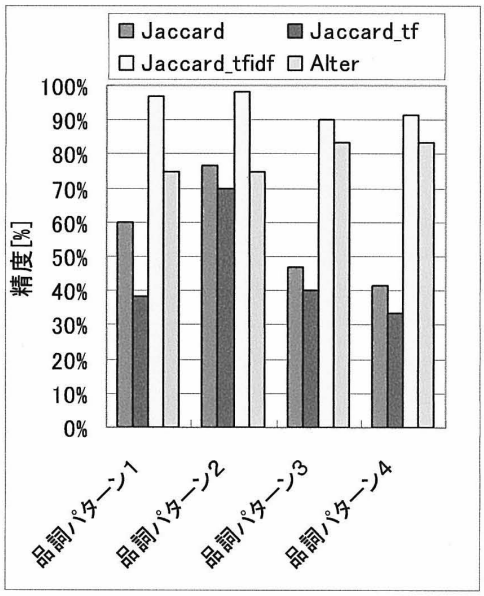

図２：話題語中の検索クエリ候補語以外の語の割合

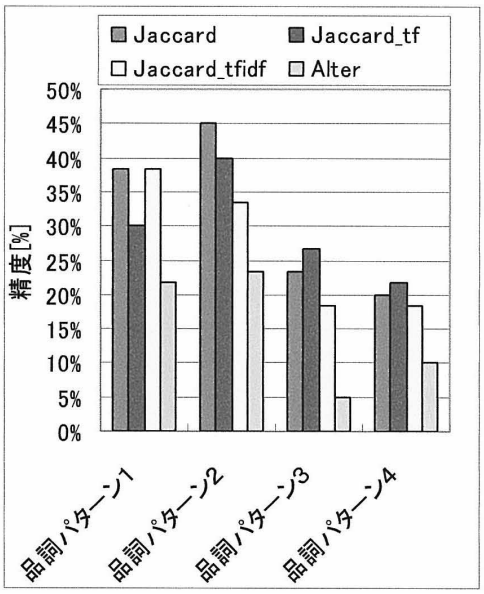

図３：入力文に含まれない話題語の正解の語の割合

図３に，検索クエリ候補語以外の話題語中に含まれる正解と判定された語の割合を示す．

図３より，形容詞を考慮した品詞パターン２の Jaccard_tf において，４０％の精度で話題語の抽出が行われている．したがって，形容詞を考慮することにより，多様な話題語の抽出が行えることが明らかとなった．

しかし，動詞を考慮した場合，抽出される語の精度が低下している．これは，動詞を考慮した場合，入力文とは関連のない文も対象として，話題語の抽出が行われてしまうため，精度が低下したものと考えられる．

図３より，品詞パターン２と品詞パターン４を見ると，形容詞と動詞を考慮した品詞パターン４に比べ，形容詞のみを考慮した品詞パターン２の方が約１８％も高い精度で話題語の抽出が行われている．

したがって，名詞以外に形容詞と動詞の２つを考慮した場合においても，動詞の影響により精度が低下したものと考えられる．したがって，なんらかのフィルタリングを適用することにより，形容詞，動詞を考慮した場合においても，名詞のみを用いた場合より，高い精度で話題語を抽出できるのではないかと考えられる．

図１より，Jaccard，Jaccard_tf において，入力文に関連がある語が抽出できていることが明らかとなった．また，図２より Jaccard_tfidf において，特徴的な語の抽出が行われていることがわかる．したがって，Jaccard，Jaccard_tf の方が，入力文に適した語の抽出が行われていると考えられる．その例として，入力文「規則正しい生活、睡眠を多く取る」に品詞パターン１を適用して抽出した話題語を表３に示す．

図２より，形容詞を考慮した場合の方が，多く検索クエリ候補語以外の語の抽出を行っていることがわかる．その例として，入力文「仙台ではライブドアの方が支持集めてたんですかね」を用いて，式（３）で関連度の計算を行い抽出した各品詞パターンの話題語を表４に示す．

結論として，名詞のみではなく，形容詞を考慮し Web から話題語の抽出を行った方が，多様な表現の話題語を抽出できることが明らかとなった．

表３：品詞パターン１で抽出した話題語

Jaccard	Jaccard_tf	Jaccrad_tfidf	Alter
睡眠	睡眠	睡眠空間	睡眠
睡眠時間	生活	睡眠単位	生活
睡眠不足	快眠	睡眠用加圧ソックス	眠り

表４：Jaccard_tf の各品詞パターンにおける抽出語

品詞パターン１	品詞パターン２	品詞パターン３	品詞パターン４
ライブドア	仙台ライブドアフェニックス	ライブドア	ライブドア
支持	球団名	仙台ライブドアフェニックス	仙台ライブドアフェニックス
仙台ライブドアフェニックス	成績	支持集め	支持集め

５．まとめと今後の課題

本稿では対話応答のための文生成を最終目標とし，そのために用いる話題語を発見するために Web ページ中の語の頻度，tf・idf を考慮し，また名詞の他に形容詞，動詞を考慮した対話ログからの話題語の抽出について述べた．

実験の結果から，本稿で試みた形容詞，動詞から話題語を抽出する際，形容詞考慮することにより，名詞のみを用いた場合より，多様な表現の話題語を抽出できることが明らかとなった．また，対話文から話題語の抽出を行う際，Jaccard 係数に tf を考慮するとすべての品詞パターンにおいて，高い精度で話題語の抽出が行えることが明らかとなった．

今後の課題として，話題語抽出の精度向上のための形容詞，動詞を考慮した，話題語に対するフィルタリングの開発と本研究の最終目標である対話応答文生成に向けての諸手法について検討を行っていく予定である．

参考文献

[1] Keiji Shinzato and Kentaro Torisawa,Acquiring Hyponymy Relations from Web Decuments, HLT-NAACL, 2004.

[2] Maya Ando, Satoshi Sekine, and Shun Ishizaki, Automatic Extraction of Hyponyms from Newspaper Using Lexicosyntactic Patterns, LREC 2004.

[3] 山本英子，梅村恭司，辞書を用いない関連語リストの構築方法，情報処理学会研究報告，2000-NL-148-12, pp. 81-88, (2000)

[4] 佐藤理史，佐々木靖弘，ウェブを利用した関連用語の自動収集，情報処理学会研究報告，2003-NL-153-8, pp. 57-64, (2003)

[5] 佐藤吉秀，川島晴美，佐々木努，大久保雅且,情報処理研究報告, 2003-NL-165-5, pp. 29-35, (2005)

[6] 小山聡，田中克己，文書構造を利用した Web からの話題発見，電子情報通信学会第 14 回データ工学ワークショップ (DEWS2003), 2-P-06, (2003)

[7] 稲川雅之，大島裕明，小山聡，田中克己，Web からの語集合間の特定関係の抽出とその可視化，電子情報通信学会第 19 回データ工学ワークショップ (DEWS2008), A7-4, (2008)

[8] 工藤拓，MeCab: Yet Another Part-of-Speech and Morphological Analyzer, http://mecab.sourceforge.jp/

[9] Yahoo!デベロッパーネットワーク，http://developer.yahoo.co.jp/