# 閲覧中のニュース記事に対するブログ記事から主張を抽出して提示するシステムの提案

# 閲覧中のニュース記事に対するブログ記事から主張を抽出して提示するシステムの提案

佐藤　大輔† 　中川　博之† 　田原　康之† 　大須賀昭彦†

# System Presenting Primary Opinions Extracted from Blogs about the News Article Being Browsed

Daisuke SATO†, Hiroyuki NAKAGAWA†, Yasuyuki TAHARA†, and Akihiko OHSUGA†

**あらまし**　本論文では，閲覧中の Web 上のニュース記事に対する意見を個人のブログから収集し，その本文中の主張部分を抽出して提示するシステムの提案を行う．現在ニュースサイトにコメント欄が用意されているところは少なく，検索エンジンを用いても個人の意見のみを収集するのは容易ではない．そこで個人の意見を述べやすい場であるブログに着目してニュース記事に関連した意見を集め，主張を抽出する．本研究では主張とは意見の中で筆者が強く述べている主観的な部分を指す．開発中の主張提示システムの中で，本論文では主張抽出に焦点を当てる．主張抽出には人手により主張であるとされた文章から形態素解析を利用して特徴的な抽出ルールを設定した．本システムによりユーザはニュースサイトを閲覧すると同時に意見の多角的な見方が可能になり，より深い洞察が得られるようになる．評価実験において人手による正解との適合率を求めたところ 70.0%となった．

**キーワード**　意見抽出，形態素解析，ブログ

## 1. まえがき

近年のインターネットの普及により，大多数の人がインターネットを利用するようになった．また，インターネットは情報を得るだけでなく，情報を発信する場になりつつある．ここ数年でブログや SNS (Social Networking Service) は多くのユーザが存在し，毎日多くの記事の投稿が行われている．それに伴い，自分が興味ある物事に対して，他の人がどう思っているかをインターネット上で調べ，それをもとにまた考え直すという利用方法が増えている．例えば自分が欲しいと思っている製品についての情報を事前に調べ，自分が重視している部分についての評価を比較し，購入するかどうかを決定するなどである．

このように Web 上で発言をすることや他人の意見を知ることがインターネットユーザにとって身近なことになりつつある中，ニュースサイトでニュースを閲覧する際に他人の意見も知りたいと思うことも考えられる．図 1 は，想定するニュース記事閲覧ユーザのニーズを示している．例えば，あるニュースを読んだときに，ある一方の立場での見方だけが書かれていて，別の立場からの見方が書かれていない場合がある．そのような場合に他のユーザがブログ等に他の立場からの意見を述べている可能性がある．このように，ニュースに対する素朴な疑問がある場合や，疑わしい記述がなされていると感じる場合に，同じように思っている人の意見を知りたいというニーズがあると考えている．しかし，現在ニュースサイトにはコメント欄が存在していないことが多い．ニュースに関するコメントを得るには掲示板が最も利用されていると考えられるが，掲示板に投稿される内容は多数派や過激派の意見が場を占めていることが多く [1]，掲示板やコメント欄があっても公平な立場で両者の意見を知ることは容易ではない．

一方，個人ブログは，現在の検索エンジンでは上位に現れない傾向にある．ブログ検索であっても，ニュースに関して検索を行うと，報道機関の記事が上位に挙がり，個人ブログは下位になる傾向がある．そのため

† 電気通信大学大学院情報システム学研究科，調布市
Graduate School of Information Systems, The University of Electro-Communications, 1–5–1 Chofugaoka, Chofu-shi, 182–8585 Japan


図 1　ニュースに対する意見へのニーズ
Fig. 1　Thinking for news.

意見を知るためには検索結果を深いところまで探さなければならない．

また，単にキーワードが含まれているだけで，あまり意見が述べられていないページも検索結果として現れるため，個人の意見を検索によって得ることは現状では困難なことが多い．しかし，個人ブログにはそれぞれが思うことが書き込まれているため，その中には強い主張や，ニュースへのストレートな批判などが含まれている．このようなブログ記事をユーザに提示できるようにすれば，それぞれのユーザはニュースについて様々な側面を知ることができ，より深い洞察を得られるようになる．

そこで，ブログから意見を集めて提示するシステムを提案する．ただし，ブログは膨大な量があるため，その全てを提示することは適切ではない．そのため，意見の中でも特に重要な部分のみを抽出することが求められる．本研究における意見の重要な部分とは，客観的な事実の記述ではなく主観的な記述の中の重要部分であり，筆者の気持ちや提案を述べている部分を指す．本研究ではこれを主張と捉え，ブログの文章からいかに主張を抽出するかに焦点を当てる．広く主張を抽出するために，本論文では筆者とはブログ投稿者のみを示すものではなく，ブログ投稿者が引用した文章の筆者も含むものとし，主張の筆者が誰であるのかについては考慮していない．文章中，どの部分が主張と思うかは人により多少異なる部分はあるが，多くの人が主張だと思う部分を出力することを目的とし，人手によって選ばれた主張部分を用いて検討を行った．

主張抽出は形態素解析を用い，主張抽出ルールを適用することで行われる．主張抽出ルールには，人手により選ばれた，筆者が意見を強く述べている部分に含まれる特徴的な形態素や表現を設定した．実験を行い精度を求めたところ，人手で作成した正解と主張抽出ルールによる出力との適合率は 70.0%となった．

既存研究として評価表現抽出や自動要約があるが，ニュースに対するブログ記事には評価表現が少なく，主張の抽出は容易ではない．また自動要約は重要語句が含まれる文が抽出されるが，重要語句は客観的な記述に多く含まれるため，本研究で提示したい主観的な部分は出力されにくいと考えられる．そのため Word2007 による自動要約の結果との比較を行い，自動要約が主張を抽出することに適していないことを示す．

本論文の構成を述べる．**2.** では主張抽出に関して，本論文で扱う主張について説明し，他の意見抽出の研究との違いや問題点を述べる．**3.** で提案システムについて概要を説明する．**4.** で主張抽出ルールに関して述べ，**5.** で実験とその結果について述べる．**6.** で関連研究に触れ，最後にまとめと今後の課題を述べる．

## 2. 主張抽出

本論文における主張とは，意見の中でも筆者が強く述べている主観的な部分を指す．そのため，従来の意見抽出における意見とは異なる．ここでは，提案システムで提示する意見について，及び従来の意見抽出手法の問題点について述べる．

### 2.1 本論文で扱う主張

本論文で提案するシステムが扱う情報元はニュース記事に関するブログ記事である．例として，東京都青少年健全育成条例改正案に関するブログ記事の一例を以下に示す．

> 東京都青少年健全育成条例改正案が可決してしまいました．（中略）ここでこうして愚痴をかいていても負け犬の遠吠えでしかありません．条例施行の 7 月までに何らかの対抗処置はまだ出来うるでしょうし，やれることがあればやらねばなりません．蟷螂の斧だろうと，漫画やアニメが好きな方々が力を合わせて，一度は否決に持ち込んだのです．少しでも出版物検閲社会にならずにすむように，こちらが逆に漫画・アニメを不健全だとかなんとか言って選別する側を監視していくべきですね．条例は可決されてしまいました．けど私は引き続き，断固としてこの悪法に反対していきます．

提案システムでは，ブログ記事から筆者の主張部分と思われる文を出力する．出力する文の数は 3 文とし

た．これは一般の web 検索結果においてタイトルと同時に出力されるスニペット（検索クエリを含む前後の文章）が約 3 行で表示されているため，同程度の文量となる 3 文を出力することとした．

どの文が筆者が強く意見を述べているかについては，読者の感性によって多少の違いがあることが考えられる．この例文に対して 7 人の被験者に，どの文章が強く意見を述べている文であるのかを調査したところ，中略後の 2 番目の文章が 4 人，最後から 3 番目の文章と最後の文章がそれぞれ 6 人が強く意見を述べていると答えた．本研究では，主張として多くの人が強く意見を述べていると感じた 3 文の出力を目指す．

### 2.2 既存の意見抽出手法とその問題点

文章から意見を抽出する研究は多く行われており，意見マイニングと呼ばれる技術が主流である．これについては評価表現を用いた評判検索や評判分析の研究が多く行われている．これらの分析対象は製品やサービスのレビューであり，口コミサイトなどから多数のレビューを分析し，製品やサービスに対する情報を集約して提示する．小林ら [2] は意見を「< 対象 > の < 属性 > は < 評価値 > である」という形で記述できる文型で表せるとし，意見は対象，属性，評価値の三つの要素からなるとした．対象は対象となっている商品やサービス名であり，属性は意見の焦点となっている対象の性質，評価値は対象あるいは属性に関する値あるいは記述者の心的な態度であると定義されている．この定義のもとで，共起パターンに基づいて評価表現の収集を半自動で行い，人手で評価表現情報を付与したデータが公開されている [3]．

この評価極性辞書を用いてニュース記事について書かれているブログ記事にどの程度評価表現が含まれているかを調査した．楽天株式会社の英語の公用語化のニュースに関するブログ記事から，主張となる記述が含まれているブログ記事 30 件を収集した．各ブログ記事で評価表現が出現する頻度を計測した結果を表 1 に示す．評価表現が一つも含まれていなかったブログ記事は 8 件あり，二つ以下のブログ記事は半数を超えた．

表 1　ブログ記事の評価表現数
Table 1　Number of evaluative expressions per blog text.

| 評価表現数 | ブログ記事数 |
|---|---|
| 0 | 8 |
| 1〜2 | 10 |
| 3〜6 | 5 |
| 7〜10 | 4 |
| 11〜 | 2 |

次に，各評価表現の PN（P は Positive，N は Negative）の評価極性と，意見の賛成反対が一致しているかどうかについても調査した．意見の内容を人手で評価したものと評価表現の極性が一致したのは 9 件のみで，残り 21 件は一致しなかった．これにより，ニュース記事の意見に対しては評価極性による主張の抽出は困難であることがいえる．

## 3. 提案システム

本章ではニュース記事に関するブログ記事の主張を提示する提案システムについて述べる．

### 3.1 概　　要

提案システムの概要を図 2 に示す．ユーザがクエリを入力し検索を実行するとブログ検索を行う．検索結果から各ブログにアクセスし，記事本文を収集する．収集した本文を形態素解析し，主張と判定される文を抽出し，提示する．

次に提案システムの利用イメージを図 3 に示す．提案システムはブラウザの下部にフッターとして現れる．フッターにはテキスト入力エリアと検索ボタンが設置されている．ユーザはニュースサイトを自由に閲覧し，他の人の意見が知りたいと思うようなニュースがあった場合に，ニュースの重要語句を入力し検索を行う．入力する語句は一つのみではニュースと関係ないブログが多くヒットしてしまうため，複数の語句を入力する．検索が実行されると，システムがバックグラウンドで主張を抽出し，フッター部分の上部の半透明のウィンドウに主張を提示する．主張の抽出は複数のブログ記事から行われ，主張は一つのブログごとにスクロールしながら提示される．ユーザは提示された主張を読み，全文を読みたいと思った場合は主張部分をクリックすることで元サイトのリンクをたどることができる．

### 3.2 ブログ記事の収集とブログ本文抽出

提案システムでは個人のブログの収集方法としてブログ検索エンジンによる検索を行う．ブログ検索エンジンには Yahoo!JAPAN のブログ検索 API [4] を用い，ユーザが入力したクエリを用いてブログ検索を行う．検索結果の各ブログの URL からブログ記事が記載されているページを読み込み，収集を行う．ただし，個人のブログを収集の対象とするため，検索結果に含まれるあらかじめ登録していたニュースサイトなどのドメインの記事からは収集を行わない．また，多

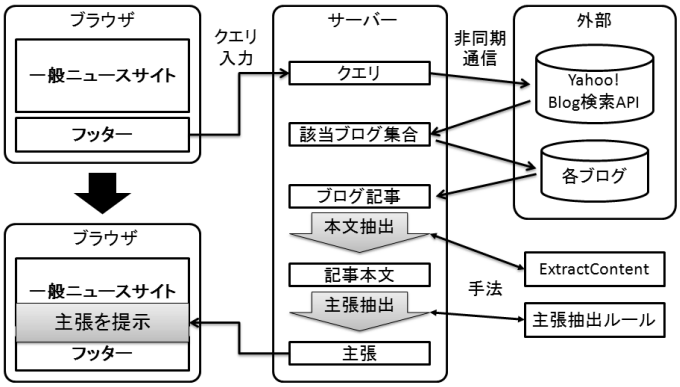



図 2　提案システムの概要
Fig. 2　Proposed system overview.

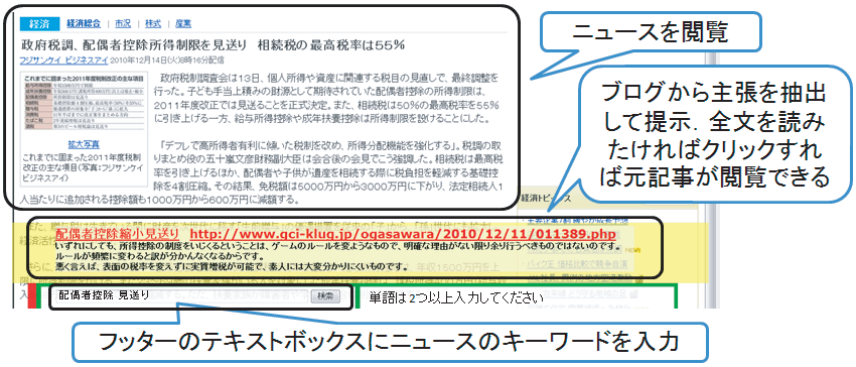



図 3　提案システムの利用イメージ
Fig. 3　Usage of the proposed system.

くのブログ記事の HTML にはコンテンツメニューやリンクや広告など記事本文以外の要素が多数含まれている．HTML タグの構成はブログサービス提供会社やデザイン構成によって異なっている．本研究では，ブログ記事 HTML から記事本文だけを抽出するために，ブログ本文抽出モジュール Extractcontent [5] を用いる．Extractcontent では，まず HTML タグの構成によりブロックに分け，それぞれのブロックの句読点の数をベーススコアとする．次にリンクやアフィリエイトなどの特有のキーワードが含まれていればそのブロックはブログ本文ではないと判定し，除外される．

### 3.3　主張抽出ルール

抽出した本文に対し，主張抽出ルールが適用できるかどうかを調べ，どの程度適用できたかによってその文の主張の度合を決定する．全ての文のルール適用部分を調べた後，ブログの各文の主張の度合を比較し，強かった 3 文を出力する．主張の度合とは，優先順位の高いルールから適用できるかを調べ，出現回数で差がつけば多いものを上位とし，出現回数が等しければ次に優先順位の高いルールを調べる，という手順を繰り返して主張の度合を相対的に測る指標を表す．主張抽出のためのルールを以下に示す．ルールに関する詳細は **4. 2** で述べる．

- 1 文ごとに (1)〜(6) の形態素が含まれているかどうかを調べ，主張の度合を測る
- ルール番号が若い方が主張の度合が強いと判定する
- 最終的に主張の度合が強い上位三つの文を出力する
- 文章の最終文は，適用できるルールがあれば度合が低くても出力する

（1）　助動詞「べし」
（2）　文末に「ば」+「なる」

（3） 思考の動詞（思う，考える）
（4） 文末に用言の評価表現
（5） 「私」などの一人称の代名詞や「個人的」などの語句
（6） 形容動詞語幹となる名詞

ただし，ここでの文末とは，文の最後の自立語以降の形態素の集合を指す．また補足として，各ルールに当てはまる形態素であっても，直後の形態素が非自立あるいは接尾語ではない名詞である場合にはルールは適用されない．

## 4. 主張抽出ルール

本章では **3.3** で述べた主張抽出ルールの詳細について述べる．ルールの設定方法，ルールの詳細，ルール設定のための予備実験について説明する．

### 4.1 抽出ルール設定手順

主張抽出ルールの設定については次のように行う．複数の文章において人手による正解データを用意し，正解の文章に表れる **3.3** で紹介した特徴的な形態素のパターンを抽出し，ルールとする．そのとき，人手により主張と判断された人数によって順序付けを行う．順序付けを行った結果，**3.3** で述べた順序を優先順位として設定した．次に，定めたルールで抽出できなかった文章や，正解ではないが抽出された文章を分析し，特徴的なパターンを抽出し，ルールとする．以上を繰り返し行うことで，ルールを洗練化した．なお，各ルール内での優先順位は設定してはいない．

### 4.2 抽出ルールの詳細

#### 4.2.1 形容動詞語幹

まずはじめに，主張文に多く出現する形態素を調査した．その結果，主張文中の形態素では形容動詞語幹が多く含まれていることが分かった．形容動詞語幹とは，後ろに「～な」などの形をとることで形容動詞となる名詞のことである．例として「必要」，「不可能」，「困難」が挙げられる．ただし，形容動詞語幹は出現頻度が他のルールの形態素に比べて多く，また，割合としては多くはないが，事実や解説をしている文章にも出現するため，主張の度合としては低く設定した．

#### 4.2.2 助動詞「べし」

次に，主張の度合を測るために文末表現に着目した．本研究において，文末とは文の最後に出現する自立語（用言や体言や副詞など）以降の形態素の集合を指す．設定手順において，助動詞の「べし」は今回調査したブログの文章には出現数自体はそれほど多くはないものの，含まれていた文章ではほぼ全てが被験者に強く意見を述べている文と選択されていた．そのため，抽出ルールに設定した．「べし」には，当然・義務・命令・可能・禁止の意味があり，「～すべきだ」など，「べし」を含む文末である場合，筆者が意見を強く訴えていることがうかがえるため，主張の度合は一番高く設定した．

#### 4.2.3 思考の動詞（思う・考える）

また同様に，多くの被験者に選ばれた文の文末には特定の動詞が含まれていることが多かった．「思う」「考える」といった思考の動詞である．これらは，筆者の気持ちや意見を述べるときによく使われる動詞であるが，1 文章でこの動詞を多用しているものがなかったため，特に用いられていれば，それは筆者の主張であるといえると考え，この二つの動詞をルールに設定した．

#### 4.2.4 文末の評価表現

ニュースに対するブログの文書に評価表現があまり使われていないことは **2.2** で述べた．しかし，少ないながらも評価表現が筆者の主張として表れる場合があった．それは文末に現れる場合である．「～すればよい」などが例として挙げられる．このように評価表現が文末にある場合を抽出ルールと設定した．

#### 4.2.5 私・個人的

ニュースに対して意見を述べるとき，主観的な部分と客観的な部分を組み合わせて述べている文章が多い．文章中に，「私」や「個人的」という単語が現れた場合，その文は主観的な部分であり，筆者が考えていることや思うことであることを示していると考えられる．そのため，これらの単語が含まれている文は予備実験にて被験者にも選ばれやすい結果となったと考えられる．特に，「私は～だと思う」などのように思考動詞とともに現れた場合，主張の度合は高くなる．

#### 4.2.6 文末の「ば」+「なる」

これは「～しなければならない」などの文末表現を表す．このような文末の場合，筆者が意見を強く訴えていることがうかがえる．システムの出力したい文章として適しており，また予備実験での被験者による選択にも現れていたため，ルールに設定した．この表現に関してはその強さから，主張の度合を高く設定した．

#### 4.2.7 最終文の評価

これは形態素に関わるルールではなく，出力する文

表 2　予備実験：被験者が選んだ主張文
Table 2　Pre-experiment: Opinion sentences selected by subjects.

| 文章 | A | B | C | D | E | F | G | H | I | J |
|---|---|---|---|---|---|---|---|---|---|---|
| 文数 | 24 | 32 | 16 | 15 | 22 | 12 | 21 | 13 | 43 | 28 |
| 上位 3 位 | 6,22,17,20 | 10,31,32 | 15,16,6 | 10,11,12 | 22,1,17 | 12,11,6 | 9,18,16 | 13,9,10 | 37,39,36 | 25,27,24 |
| 2 人以上 | 11,24 | 9,11,13,27,28 | 4 | 3,9,13,15 | 6,9,18 | なし | 7,15,5,10,19 | 1,2 | 34,38,42 | 4,21,6,20 |

に関わるルールである．「最終文に関しては，適用できるルールがあれば出力する」を設定する．これは，最終文において適用できるルールがある場合，最終文は結論を述べている可能性が高いからである．予備実験においても，最終文が結論であった場合，選ばれる可能性が高かったため，最終文は特別な評価を行う．

**4.2.8**　補足ルール

ルールの補足として，「各形態素の直後が接尾・非自立でない名詞である場合，ルールの適用はしない」を設定する．これは形態素解析が複合語に対応しきれないためである．複合語の中にルールに適応する形態素が含まれていた場合に抽出してしまうことを防ぐ．例えば「特別議決」という語句の場合，形容動詞語幹「特別」が含まれているが，直後が「議決」という名詞であるため，ルール適用外となる．

### 4.3　予 備 実 験

抽出ルールの設定のために行った予備実験について述べる．それまでに設定していたルールを用いて主張抽出を行い，正解の文章が抽出できるか適合率を求めた．適合率は式 (1) で求める．

$$\text{適合率} = \frac{\text{抽出ルールによる出力集合}}{\text{人手による正解集合}} \qquad (1)$$

再現率に関しては，人手による選択が 3 文と決められているため，適合率と同値になる．

正解の文章は 9 人の被験者により人手で作成したものである．被験者として大学院生 8 名，助教 1 名の計 9 名に協力いただいた．被験者にはそれぞれ 10 の文章を読み，主張部分として筆者が強く意見を述べていると感じた文 ID（文番号）を選択してもらった．選んでもらう箇所は，システムの出力と同じ 3 文とした．被験者が選んだ文を集計し，選んだ人数が多かった上位 3 位の文のみを正解とした場合の適合率と，2 人以上が選んだ文を正解とした場合の適合率との 2 種類を求めた．

まずそれまでにルールとして設定していた形容動詞語幹，助動詞「べし」，思考の動詞の三つの形態素を用いたルールを文章に適用させ，出力を求めた．

表 2 に被験者が選んだ文 ID を示す．上位 3 位は，被験者 9 人のうち選んだ人数が多かった上位 3 位までの文 ID を表し，2 人以上は上位 3 位以内ではないが，被験者 2 人以上が選んだ文 ID を表している．上位 3 位で 4 文含まれているのは，3 位が同じ人数で 2 文あったためである．表 3 に初期ルールによってシステムが選んだ 3 文を示す．網掛けが濃いセルが上位 3 位に入った文 ID で薄いセルが 2 人以上が選んだ文 ID を表す．これより，上位 3 位の文のみを正解とする場合，システムの適合率は 33.3%，2 人以上が選んだ文も正解とすると適合率は 60.0%となった．

表 3　予備実験：抽出結果
Table 3　Pre-experiment: Results of rule-based extraction.

| 文章 | A | B | C | D | E | F | G | H | I | J |
|---|---|---|---|---|---|---|---|---|---|---|
| 出力 1 | 22 | 11 | 15 | 3 | 6 | 6 | 15 | 13 | 37 | 28 |
| 出力 2 | 24 | 10 | 2 | 5 | 22 | 8 | 19 | 8 | 39 | 9 |
| 出力 3 | 6 | 28 | 8 | 7 | 1 | 10 | 21 | - | - | 27 |

また，抽出に失敗した文章から新たにルールを決定した．例えば，最終文に関しては文章 A，B，C，D，E，F，H で被験者 2 人以上が選択しているが，抽出できているのは A，E，H のみである．しかし，出力されなかった最終文それぞれにはルールが適用できる箇所が存在した．これより，最終文に関しては特別に優先順位を高くすべきと考え，ルールを設定した．このようにして，文末の評価表現，「私」「個人的」といった単語，文末の「ば」+「なる」，最終文の評価，補足ルールが設定された．

## 5. 実　　験

### 5.1　実 験 概 要

**4.2** で紹介した抽出ルールの評価を行うため，実験を行った．実験により抽出ルールによる結果と人手による正解との適合率を求めた．**4.3** の予備実験と同様に，大学院生 6 名と助教 1 名の計 7 名の被験者に協力して頂き，2010 年 12 月のブログから収集した新しい文章 K〜T の 10 文章 [17] を読んでもらい，筆者が意見を強く述べていると感じた文を選択してもらった．選択する文の数についても同様にシステムの出力と同

表 4　実験：被験者が選んだ主張文
Table 4　Experiment: Opinion sentences selected by subjects.

| 文章 | K | L | M | N | O | P | Q | R | S | T |
|---|---|---|---|---|---|---|---|---|---|---|
| 文数 | 33 | 14 | 18 | 16 | 16 | 11 | 27 | 22 | 18 | 26 |
| 上位 3 位 | 33,23,7 | 14,5,13 | 16,18,14 | 7,4,10 | 16,4,7 | 10,4,11 | 6,22,24 | 8,10,17,19 | 7,17,16,18 | 7,20,22 |
| 2 人以上 | 10,19,29,32 | 3,7 | 12 | 8,11,16 | 8,13,14 | 6 | 7,21,27 | 22 | なし | 4,6,12,26 |

表 5　実験：抽出結果
Table 5　Experiment: Results of rule-based extraction.

| 文章 | K | L | M | N | O | P | Q | R | S | T |
|---|---|---|---|---|---|---|---|---|---|---|
| 出力 1 | 26 | 14 | 16 | 7 | 9 | 2 | 24 | 17 | 18 | 6 |
| 出力 2 | 33 | 5 | 18 | 10 | 14 | 6 | 7 | 10 | 16 | 26 |
| 出力 3 | 10 | 10 | 4 | 16 | 15 | - | 22 | 14 | 17 | 22 |

じ 3 文とした．文章のテーマについてはテレビや新聞で大きく取り上げられたものを中心に，また文章の長さは短すぎる場合十分な評価が行えないため 10 文以上述べられているものを選択した．表 4 に被験者が選んだ文 ID を示す．表の読み方は **4. 3** を参照されたし．

実験結果は表 5 のようになった．結果の読み方は **4. 3** と同様であり，網掛けの部分が正解となる部分である．適合率は上位 3 位のみを正解とする場合は 50.0%，2 人以上を正解とする場合は 70.0%となった

### 5. 2　考　　察

#### 5. 2. 1　自動要約との比較

文章から文を抽出することは要約ともいえるため，自動要約との比較を行った．Microsoft 社の Word2007 で利用できる自動要約機能を用いて同じ文章の要約文を取得し，主張が抽出できたかどうかを比較した．結果を表 6，表 7 に示す．適合率を表 8 に示す．予備実験に対応する文章 A～J の適合率は上位 3 位のみを正解とする場合は 10.0%，2 人以上を正解とする場合は 23.3%となった．実験に対応する文章 K～T の適合率は上位 3 位のみを正解とする場合は 6.7%，2 人以上を正解とする場合は 13.3%となった．いずれの場合も，提案手法と比べて下回っている．この結果から自動要約を用いても主張の抽出は容易ではないことが考えられる．

また参考として，今回予備実験及び実験で用いた文章にどの程度評価表現が含まれているかを表 9 に示す．実験の文章でも評価表現の出現数は少なく，既存研究の技術も主張の抽出には適していないといえる．

#### 5. 2. 2　抽出できなかった文

予備実験をもとにルールを追加したことで，文章を変更しても適合率の向上をみることができた．しかし，

表 6　Word 自動要約による結果　文章 A～J
Table 6　Result from automatic summarization of texts A through J by Word2007.

| 文章 | A | B | C | D | E | F | G | H | I | J |
|---|---|---|---|---|---|---|---|---|---|---|
| 出力 1 | 8 | 16 | 1 | 3 | 4 | 2 | 13 | 4 | 5 | 7 |
| 出力 2 | 10 | 21 | 3 | 4 | 5 | 7 | 17 | 10 | 14 | 13 |
| 出力 3 | 18 | 25 | 16 | 10 | 18 | 10 | 20 | 13 | 15 | 27 |

表 7　Word 自動要約による結果　文章 K～T
Table 7　Result from automatic summarization of texts K through T by Word2007.

| 文章 | K | L | M | N | O | P | Q | R | S | T |
|---|---|---|---|---|---|---|---|---|---|---|
| 出力 1 | 2 | 3 | 3 | 10 | 1 | 8 | 3 | 8 | 6 | 1 |
| 出力 2 | 6 | 6 | 10 | 12 | 10 | 9 | 8 | 12 | 9 | 2 |
| 出力 3 | 8 | 8 | 17 | 14 | 13 | 11 | 15 | 18 | 15 | 7 |

表 8　文章 K～T の適合率比較
Table 8　Comparing precision of texts K through T.

| 手法 | 提案手法 | Word2007 |
|---|---|---|
| 上位 3 位 | 50.0% | 6.7% |
| 2 人以上 | 70.0% | 13.3% |

文章 O，P に関しては上位 3 位の文章が一つも抽出できなかった．そのうちの 2 文は被験者全員が選んだ文章であった．その文を紹介する．

> 彼の人間性を云々という話だが，それは彼が「英雄」としてマスメディアに出現するのを阻止しようとする勢力の意図的な報道だ．

> 国税で賄いきれないなら子ども手当てなんてやめちゃえよ．

この二つの文には，抽出ルールが適用できる部分が一つもなかったため，抽出できなかった．現状，この二つの文章から特別な要素を見出すことができていない．文末の助動詞「だ」には断定の意味があり，また終助詞の「よ」には強意の意味があるため，この部分から強く述べているという判断も可能であるが，助詞・助動詞に対して主張の度合を高く設定すると多くの別の文が主張として判定され，結局適合率が下がってしまうため，ルールに設定しなかった．

まだ発見していない効果的なルールがあるものと考

表 9　文章に含まれていた評価表現数
Table 9　Number of evaluative expressions from text.

| 文章 | A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 評価表現数 | 5 | 1 | 4 | 3 | 9 | 2 | 1 | 1 | 2 | 5 | 3 | 1 | 1 | 1 | 2 | 2 | 0 | 2 | 6 | 1 |

表 10　実験：被験者の選択の一致度
Table 10　Experiment: Kendall's coefficient of concordance.

| 文章 | K | L | M | N | O |
|---|---|---|---|---|---|
| 一致度 | 0.3661 | 0.5428 | 0.5159 | 0.1977 | 0.5493 |
| p 値 | $2.86^{-6}$ | $8.83^{-6}$ | $6.18^{-7}$ | 0.1448 | $6.29^{-7}$ |

| 文章 | P | Q | R | S | T |
|---|---|---|---|---|---|
| 一致度 | 0.6173 | 0.2203 | 0.2609 | 0.4970 | 0.2329 |
| p 値 | $4.56^{-6}$ | $3.825^{-2}$ | $1.172^{-2}$ | $1.45^{-6}$ | $2.432^{-2}$ |

えられるが，1 文だけで判断する方法のみではなく，文の前後関係や接続詞による文章の構造を利用した判断，重要語抽出を利用するなど文章全体を使った主張の抽出，助詞・助動詞の周辺の語句を含めた文末表現について細かく調査するなど，新たなルール考えることで適合率の向上を図りたいと考えている．

**5. 2. 3**　被験者間の一致度

実験において，被験者の解答の一致度をケンドールの一致係数によって求めた．表 10 に結果を示す．被験者の解答の一致度は文章によって中程度の一致が見られるものや，あまり一致していないものなど，結果にばらつきがあった．程度の差はあるが被験者によって主張と思う部分が異なっていることがうかがえる．これをある程度吸収して抽出を試みたのが今回の実験である．今回の結果からは一致度の違いによる抽出精度への影響はあまりないのではないかと考えている．

**5. 2. 4**　ルールの貢献度

実験結果から，ルールの適用傾向と貢献について考察した．抽出結果のうち，正解と判定した 22 文のうち，最終文によるものが 6 文，形態素パターンによるものが 16 文となった．主張と判定される文はほとんどがルールを複数適用したものとなった．最終文の適用に関しては，ルール (1)〜(6) を伴って出力された．形態素パターンのみの場合のルールの適用傾向としては，ルール (1)〜(3) が含まれる場合に有力な候補となり，更なる主張度合を高めるものとしてルール (4)〜(6) の適用がみられた．よって今回の実験では，最終文，ルール (1)〜(3)，ルール (4)〜(6) がそれぞれ異なった役割で適用に貢献していたといえる．最終文については今回は日本語のみを対象にしたが，他言語でも日本語と同様な特徴があれば有効であると考えている．

## 6. 関 連 研 究

WebAlert [6] は，閲覧中の Web ページの内容に誤りの可能性がある場合に警告及び反証を提示するシステムである．あるトピックに関する閲覧中の Web ページのポジティブまたはネガティブな内容に対して，そのトピックに関する他の記事の多くがネガティブまたはポジティブな内容で書かれていた場合，閲覧中のページ内容が信頼できない可能性があるとして，ユーザに警告を発し，同時にその内容を提示するシステムである．情報の信頼性が確保されていないページに対し，情報の内容を鵜呑みにしてしまうことは，今後不都合が生じる可能性がある．そのような事態を防ぐ目的で作られたシステムである．

Googlesidewiki [7] は，Web 上のあらゆるページについてコメントを表示，投稿，共有できる Web ブラウザのサイドバーとして提供されている．閲覧中の Web ページあるいはページ内の文章に関して，各ユーザがコメントを投稿し，他のユーザはこれを読むことができる．これにより，今後そのページを閲覧するユーザに対し情報を提供することができる．共有される情報の内容は，専門家の洞察の紹介や，背景知識の説明，内容の補足説明や個人の感想などといったものである．Googlesidewiki の欠点としては，情報を積極的に投稿するユーザが必要であることと，その 1 ページに対して情報を投稿するため，閲覧する人が限られるということである．同じような内容のページが他にもあった場合，そちらにも投稿しなければ片方にしか情報がない状態になる．

WISDOM [8] (Web Information Sensibly and Discreetly Ordered and Marshaled) は，独立行政法人情報通信研究機構 (NICT) が研究開発を進めている情報分析システムであり，2010 年 8 月 9 日に正式公開された．Web にある情報を様々な観点から分析することによってユーザが情報を多角的に捉えながら情報の信頼性を判断するのを支援するためのシステムである．自然言語によるクエリを入力すると，検索結果に

関して情報外観，情報発信者，主要名詞句，主要・対立文，意見分析，要約の六つの分析結果が得られる．情報を多角的に捉えるという点で本研究と類似しているが，WISDOM は信頼度の高いと思われる結果を優先して提示するため，報道機関や営利・非営利団体が公表しているものが主に提示される．そのため個人の意見は結果にあまり含まれていない．実際に WISDOM を用いてニュース記事について重要語を入力して調べて見たところ，ニュース記事そのものに対する分析が結果の 6 割以上を占めていた．

その他，意見抽出に関する関連研究としては，SVM (Support Vector Machine) と新聞記事を用いて Weblog の記事をレビュー記事と非レビュー記事に分類し，意見を抽出する研究 [10]，賛否両論が対立する構図を論点に基づいて可視化するシステム OpinionReader [11]，ブログ記事からの多視点からのトピック抽出 [12]，両方の極性に含まれる単語の極性の判定精度の向上の研究 [13]，構文解析木を利用した要望意見の抽出の研究 [14]，係り受け解析を利用して態度・理由・主体・対象を抽出する研究 [15]，掲示板からの重要議論抽出と特徴表現抽出の研究 [16] などがある．

様々な既存のシステムや研究を紹介したが，主張の抽出に特化したものは確認できなかった．今後は，主張の抽出精度を向上させるとともに，抽出した主張の分類や提示の仕方などを従来の意見抽出手法やシステムを参考にしつつ，新たなルールの模索を行っていきたい．

## 7. むすび

閲覧中のニュース記事に対するブログの主張を提示するシステムの提案を行った．提案システムの重要な部分である主張抽出方法を形態素情報を利用することで判別し，その適合率を求めた．今後はより多くの文章に対して実験を行い，ルールの改善を行うことで精度の向上を図りたい．

提案システムの今後としては，今回はユーザに検索クエリを入力する形をとったが，重要語抽出技術等を用いてクエリを自動で生成するようにしたいと考えている．更に，各ブログの抽出結果によってリランキングを行うことにより，より強い主張や重要な意見を出力できるようにする予定である．これにより，主張の弱いブログ記事の順位が下がり，システムの出力が目的の達成により近づくものとなると考えている．出力する文についても今回は 3 行と限定していたが，今後は特に重要な文が多く抽出できたならば限定を解除するなども考慮したい．また，掲示板も抽出対象にしてブログとの比較を行うことや抽出した主張については観点別にクラスタリングを行うなどによって，システムのユーザが更なる洞察を得られるように工夫するなど，独自のシステムを構築したい．

**謝辞**　本研究を遂行するにあたり，研究の機会と議論・研鑽の場を提供して頂き，御指導頂いた国立情報学研究所/東京大学本位田真一教授をはじめ，活発な議論と貴重な御意見を頂いた研究グループの皆様に感謝致します．

## 文　献

[1] S. Faridani, E. Bitton, K. Ryokai, and K. Goldberg, “Opinion space: A scalable tool for browsing online comments,” 28th ACM Conference on Human factors in Computing Systems (CHI), pp.1175–1184, 2010.

[2] 小林のぞみ，乾健太郎，松本裕治，立石健二，福島俊一，“意見抽出のための評価表現の収集，” 自然言語処理，vol.12, no.2, pp.203–222, 2005.

[3] 日本語評価極性辞書：http://cl.naist.jp/~inui/research/EM/sentiment-lexicon.html

[4] Yahoo!ブログ検索：http://developer.yahoo.co.jp/webapi/search/blogsearch/v1/blogsearch.html

[5] Web ページの本文抽出：http://labs.cybozu.co.jp/blog/nakatani/2007/09/web_1.html

[6] 山本祐輔，手塚太郎，アダム　ヤトフト，田中克己，“Web-Alert: Web 情報の印象集約を利用した閲覧ページ内容に対する反証提示，” 第 19 回データ工学ワークショップ (DEWS)，B10-3, 2008.

[7] Googlesidewiki：http://www.google.com/sidewikiGooglesidewiki

[8] NICT 情報分析システム WISDOM：http://wisdom-nict.jp

[9] 乾健太郎，“意見マイニング　ー Web からの意見情報の抽出と要約ー，” 電気四学会関西支部専門講習会「分析・可視化による情報の価値化」，2006.

[10] 川口敏宏，松井藤五郎，大和田勇人，“SVM と新聞記事を用いた Weblog からの意見文抽出，” 人工知能学会全国大会（第 20 回）(JSAI)，1A3-03, 2006.

[11] 藤井　敦，“OpinionReader: 意思決定支援を目的とした主観情報の集約・可視化システム，” 信学論（D），vol.J91-D, no.2, pp.459–470, Feb. 2008.

[12] 戸田智子，黒田晋矢，福田直樹，石川　博，“ブログにおける多視点からのトピック抽出手法の提案，” 電子情報通信学会第 19 回データ工学ワークショップ (DEWS)，B4-2, 2008.

[13] X. Ding, B. Liu, and P.S. Yu, “A holistic lexicon-based approach to opinion mining,” Proc. International Conference on Web Search and Web Data Mining (WSDM), pp.231–240, 2008.

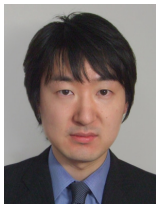
[14] H. Kanayama and T. Nasukawa, “Textual demand analysis: detection of users’ wants and needs from

opinions," International Conference on Computational Linguistics (COLING), pp.409–416, 2008.
[15] 小林大祐，井上　潮，"Web 上のレビュー情報からユーザが重要視する製品の特徴を抽出する手法の提案，" 第 1 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)，C6-4, 2009.
[16] 桜井茂明，折原良平，"掲示板サイト分析における重要議論抽出と特徴表現抽出，" 知能と情報，vol.19, no.1, pp.13–21, 2007.
[17] ブログの引用元について：http://www.ohsuga.is.uec.ac.jp/member/sato_ref.html

（平成 23 年 1 月 13 日受付，5 月 25 日再受付）

**佐藤　大輔**

2009 電通大・情報通信卒．現在同大大学院情報システム学研究科に在籍．



**中川　博之**　（正員）

1997 阪大・基礎工・情報工学卒．同年鹿島建設（株）に入社．2007 東京大学大学院情報理工学系研究科修士課程了，2008 同大学院博士課程中退．同年より電気通信大学助教，現在に至る．エージェント及び自己適応システム開発手法の研究に従事．情報処理学会，IEEE CS 各会員．

**田原　康之**

1991 東京大学大学院理学系研究科数学専攻修士課程了．同年（株）東芝入社．1993〜1996 情報処理振興事業協会に出向．1996〜1997 英国 City 大学客員研究員．1997〜1998 英国 Imperial College 客員研究員．2003 国立情報学研究所入所．2008 より電気通信大学准教授．博士（情報科学）（早稲田大学）．エージェント技術，及びソフトウェア工学などの研究に従事．情報処理学会，日本ソフトウェア科学会各会員．

**大須賀昭彦**　（正員）

1981 上智大・理工・数学卒．同年（株）東芝入社．同社研究開発センター，ソフトウェア技術センターなどに所属．1985〜1989（財）新世代コンピュータ技術開発機構 (ICOT) 出向．2007 より，電気通信大学大学院情報システム学研究科教授．工博（早稲田大学）．主としてソフトウェアのためのフォーマルメソッド，エージェント技術の研究に従事．1986 年度情報処理学会論文賞受賞．情報処理学会，日本ソフトウェア科学会，人工知能学会，IEEE CS 各会員．