

ニュース記事理解における発言情報の重要度算出手法

吉田 慶章† 柿崎 淑郎‡ 辻 秀一††

† 東海大学 工学研究科 259-1292 平塚市北金目 1117 番地
‡ 東京理科大学 工学部 102-0073 千代田区九段北 1-14-6

†† 東海大学 情報通信学部
259-1292 平塚市北金目 1117 番地

あらまし 政治家の発言情報を提供することで、ユーザのニュース記事理解を支援する方式を既に提案しているが、発言情報の収集を手で行っていた点と、発言情報の表示手法に課題が残されていた。本研究では、発言情報の抽出手法と重要度算出手法から構成される自動収集手法を提案する。さらに、発言情報の表示手法も提案する。最後に、試作システムを用いた実験より、抽出手法の精度や重要度算出手法の妥当性等に関して評価を行う。

A Method of Calculating Quote Importance for News Understanding

Yoshiaki YOSHIDA† Yoshio KAKIZAKI‡ Hidekazu TSUJI††

† School of Engineering
Tokai University
1117 Kitakaname, Hiratsuka,
Kanagawa, 259-1292 Japan

‡ Faculty of Engineering
Tokyo University of Science
1-14-6 Kudankita, Chiyoda-Ku, Tokyo, 102-0073 Japan

†† School of Information and Telecommunication Engineering
Tokai University
1117 Kitakaname, Hiratsuka,
Kanagawa, 259-1292 Japan

Abstract We proposed a method to support for news understanding using politician's quotes. However, the method has two issues that extracting quote by hand and presentation method of quote. In this paper, we propose automatic methods to extract quotes and to calculate quote importance. In extract method, quotes are extracted using representation dictionary of quote, surface pattern rule and news ontology. In calculate method, quotes are calculated by elements of quote. Finally, we evaluate accuracy of the extract method and adequacy of the calculate method.

1 はじめに

近年、WWW に日々膨大な情報が公開されており、これらの情報を知識資源と捉え、特定のコンテンツに関連する情報を提供するサービスが注目されている。

私たちは、ユーザのニュース記事の理解を支

援することを目的とし、ニュース記事のトピックに関連する政治家の発言情報を提供する方式を提案している [1]。しかしこの方式において、発言情報の収集を手で行っていた点に課題が残されていた。本研究では、この作業を自動化した発言情報の自動収集手法を提案する。

2 発言情報の自動収集手法

発言情報をニュース記事集合から人手で収集する作業は非常に高コストであるため、発言情報の自動収集手法を提案する。自動収集手法は抽出手法と重要度算出手法から構成される。

2.1 発言情報の抽出手法

発言情報を抽出するにあたり、評判情報抽出分野で提案されている意見抽出手法が関連している。この意見抽出手法は、主に評価表現辞書とパターンマッチングルールが用いられている[2]。評価表現辞書とは、「～は綺麗だ」などの主観的な表現と、その表現が肯定/否定のどちらに属するかを収集した辞書である。また、人手での評価表現辞書の作成は非常に高コストな作業であるため、評価表現を自動的に収集する研究も行われている[3, 4, 5]。

本研究では、発言表現辞書と表層パターンマッチングルール、そしてニュースオントロジーを用いた抽出手法を提案する。

まず、括弧表現の一種である鉤括弧表現「」に着目する。ニュース記事における鉤括弧表現は発言情報を表す一般的な表現である。しかし、実際に鉤括弧表現が使われる箇所は発言情報だけではなく、重要語や新語などにも使われるために、鉤括弧表現に囲まれた文字列が、発言情報なのかどうかを判別する必要がある。

2.1.1 表層パターンマッチングルール

本手法では、図1、図2、図3に示した表層パターンマッチングルールを用いる。まず、基本的な表層パターンである図1に関して述べる。

発言情報には、必ず発言者が存在するという特徴があるため、人名候補部を抽出する。

次に「について」などを手掛かり語としてサブコンテキストを抽出する。サブコンテキストは、発言が行われた場面や発言情報では暗黙的に省略されている対象を補足する情報である。しかし、サブコンテキストが出現しない発言情報も多いため、必須抽出項目とはしない。

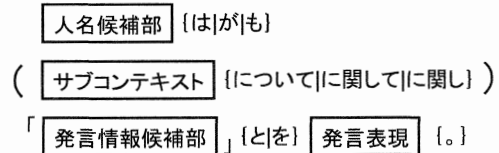


図 1: 表層パターンマッチングルール 1

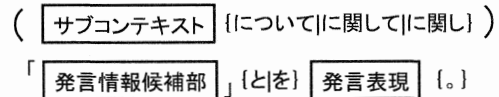


図 2: 表層パターンマッチングルール 2

最後に、発言表現の前方に出現する、鉤括弧表現に囲まれた文字列を発言情報候補部として抽出する。発言表現とは、表1のような発言情報の後方に出現する文末表現のことを指す。今回、この発言表現を収集した発言表現辞書を人手にて作成した。

図2に示した表層パターン2は、発言者が暗黙的に省略されたものであり、図3に示した表層パターン3は、1文で2者間の発言情報を対比させて表す場合に用いられる。

2.1.2 正規化

上記の表層パターンマッチングルールによって抽出された人名候補部は、文章の構造によって、表2に記したような多様性を持った文字列として抽出される。そこで、人名候補部に人名情報が含まれている場合 1) 2) 3) には、正しい人名情報に正規化をすることで、発言情報候補

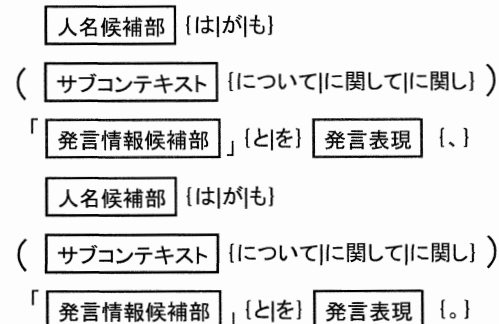


図 3: 表層パターンマッチングルール 3

表 1: 発言表現の例

発言表現例		
発言した	述べた	訴えた
表明した	発表した	批判した

表 2: 人名候補部の抽出例

人名候補部の例	判定
1) 福田康夫	正規化不要
2) 民主党の山岡賢次国対委員長	正規化必要
3) しかし、この日の首脳会談で首相	正規化必要
4) 与党と民主党が修正協議という歩み 寄りに至った背景に	除去

部を発言情報と見なし、人名候補部に人名情報が含まれていない場合 4) には、発言情報候補部を発言情報と見なさず、除去する。

正規化には、ニュースオントロジーの知識を利用する。人名情報の正規化に用いるニュースオントロジーの部分を図 4 に示す。今回作成した `aliasIs` プロパティによって人名の表記揺れ情報が各役職のインスタンスごとに体系化されている。この知識を利用することで、人名候補部を正しい人名情報に正規化をする。

次に、表 2 の 3) で例示した「首相」の同定に関して述べる。ニュース記事では、内閣総理大臣のことを表す際、その通称である首相が用いられることが多いという特徴があるため、そのニュース記事における首相は誰であるのかを同定する必要がある。

首相の同定にも、ニュースオントロジーの知識を利用する。今回作成した `assumptionDate` プロパティと `resignationDate` プロパティによって就任日情報と辞任日情報が各役職のインスタンスごとに体系化されている。これらの日時を比較することで首相を正しく同定する。

2.2 発言情報の重要度算出手法

抽出手法を適用することで、発言情報の抽出が容易になる一方、膨大な発言情報の中で、どの発言情報が重要であるのかを判断できないと

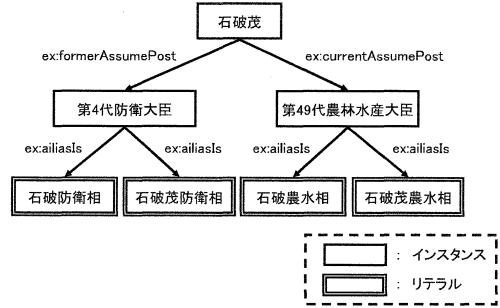


図 4: 正規化で用いるニュースオントロジー

いう問題点が存在する。そこで、各発言情報に重要度を算出することで、この問題を解決する。

重要度を算出するにあたり、文書要約分野における重要文抽出タスクが、ニュース記事中のテキストに重要度を与える点で類似している [6]。代表的な手法として、テキスト中のキーワードの出現頻度を利用する手法 [7] や、位置情報を利用する手法 [8]、そしてニュースドメインにおいては、見出し情報を重要視する手法などが挙げられる [9]。また、これらの要素を複合的に用いた手法もある [10]。

2.2.1 重要度算出に用いる要素

本研究では、これまで文書要約分野において利用されてきた表層的な要素と、発言情報における表層的な要素を複合的に用いた重要度算出手法を提案する。利用する要素は、1) 記事内順位 2) 記事構造 3) 数値文 4) 発言表現確信度 5) 時系列の計 5 要素である。

記事内順位スコア

記事内順位スコアは、抽出された発言情報がその抽出源となったニュース記事の中で、何番目に重要度の高い文に含まれているかという指標であり、他の記事と相対的に比較を行うため、記事内順位値の割合を算出する。記事内順位値を算出するために、文書内 TF-IDF を利用する。文書内 TF-IDF は、対象文書内の 1 文を 1 文書と見なし、全文数を全文書数と見なして TF-IDF を計算する手法である。ニュース記事の各文の

スコア $T(S)$ は、式 (1) によって文中に出現する単語の TF-IDF 値の総和を取ることによって算出する [11].

$$T(S) = \sum_{w \in S} TF(w, S) * \log \frac{totalsentence}{line(w)} \quad (1)$$

ここで、 $TF(w, S)$ は文 S での単語 w の出現頻度、 $totalsentence$ は文の総数、 $line(w)$ は単語 w を含む文の総数である。

式 (1) によって算出されたスコア $T(S)$ を元に、式 (2) によって、記事内順位スコア $R(S)$ を算出する。

$$R(S) = 1 - \frac{rank(S)}{totalsentence} \quad (2)$$

ここで、 $rank(S)$ は、 $T(S)$ を降順にソートしたときの順位、 $totalsentence$ は式 (1) と同様に文の総数である。

記事構造

記事構造には、2 種類の子要素がある。まず、発言情報がニュース記事の冒頭部分に位置しているかという指標である。これは今回対象とする政治ジャンルのニュース記事が、記事の冒頭に要点が書かれることが多いという特徴に基づいている。本手法では、記事の上位 50% に発言情報が位置している場合に重要視する。重要視する度合いは、上位 20% と上位 20-50% で異なる。

次に、発言情報が見出しの一部になっているかという指標である。見出しの一部に発言情報が含まれている場合、その発言情報が記事の中心論点である場合が多く、重要な発言情報である。

しかし、この場合の多くは、発言情報に対して要約や言い換えが行われているため、双方が同一の発言情報であるかを判定する必要がある。そこで、発言情報と見出しの一部である発言情報の特徴語を用いた判定を行う。

判定として、双方の文字列の特徴語を比較し、発言情報の特徴語が見出しの一部の発言情報の特徴語を閾値以上再現している場合、双方が同一の発言情報であると判定する。今回、閾値を 45% とした。特徴語として、名詞と動詞の原形

を用いる。動詞の原形を用いるのは、たとえば「そらすことなく」と「そらすず」のような言い換えがしばしば行われるためである。

数値文

数値文とは、数値情報を含んだ発言情報のことである。数値情報とは、「10 日」や「2 億円」などの数値と単位の組み合わせと定義する。政治ジャンルのニュース記事において、数値情報が報道の焦点になることがしばしばあるため、重要視する。ここで発言情報に含まれる数値情報のセット数を *NUMERAL* とする。

発言表現確信度

発言表現確信度とは、発言者がどの程度の意識で発言を行ったのかという指標を数値化した要素である。具体的には、「明言した」とあれば、方針を打ち出すような場合が多く、注目すべきだと考えられるが、「つぶやいた」とあれば、大きな影響を与えるような発言情報ではない場合が多い。各発言表現に、1-5 の確信度スコア *CONFIDENCE* を付与した。確信度スコアが高い程、重要な発言情報を示す発言表現である。

確信度スコアの付与は、発言表現とその発言表現が用いられた具体的な発言表現を提示し、「どの発言表現が用いられている発言情報が重要か？」という基準の基、全て人手にて行った。

時系列

ニュース記事は新しいほど重要であるという特徴を持っているため、ニュース記事から抽出した発言情報においても、同様に新しい発言情報ほど重要であると考えられる。そこで、特定のトピックにおける時系列係数 *TIMELINE(RT)* を式 3 より算出する。

$$TIMELINE(QT) = \frac{2}{newtime - oldtime} QT + 1 \quad (3)$$

ここで、 QT は特定のトピックに関する発言情報の発言時刻の UNIX 時間であり、 $newtime$ は特定のトピックに関する最新の発言情報の

発言時刻の UNIX 時間, *oldtime* は, 特定のトピックに関係する最古の発言情報の発言時刻の UNIX 時間である.

各発言情報における時系列係数は 1-3 の値となり, 古い発言情報であるほど時系列係数は 1 に近似され, 新しい発言情報であるほど時系列係数は 3 に近似される.

2.2.2 重要度の算出

各発言の重要度は, 上記の 5 要素を複合的に用いることで算出する. 要素とその得点の関係を表 3 に示す. そして, 各発言情報の重要度 $IM(Q)$ は, 式 4 によって算出される. 各要素に対する得点は, 人手による試行錯誤によって決定した.

$$IM(Q) = (E1 + E2_1 + E2_2 + E3 + E4) \times E5(4)$$

2.3 発言情報の表示手法

ニュース記事理解支援方式は政治家の発言情報を提供することで, ニュース記事の理解を支援する方式であるが, 発言情報の表示手法もユーザの理解度を変化させる要因であると考えられるため, 以下の 2 種類の表示手法を提案する.

表示手法 1 重要度が高い発言情報から提供する表示手法

表示手法 2 時系列順に提供し, その重要度に応じてテキスト装飾を行う可視化手法を適用した表示手法

表示手法 1 は, その発言情報が行われた日時よりも重要度を重要視する表示手法である. 表示手法 2 は, 時系列順に提供する表示手法である. さらに, 重要度に応じてテキスト装飾を行う可視化手法を適用する. この可視化手法として, タグクラウド手法を用いる. タグクラウド手法は, 散在したタグの頻出度や重要度によって強調し表現する可視化手法であり, 私たちはこのタグクラウド手法をニュース記事に適用し, 重要な特徴語を視覚的に得る方式を提案した [12]. 表示手法 2 の例を図 5 に示す. 今回は, 色調 (黒と青) と文字サイズの 2 要素を用いてテキストの装飾を行った.

石破氏の発言(時系列順でタグクラウド表示)	重要度	日時
ご家族、関係者、国民に深くおわびしたい	22.7604790419	2008-08-04
今まで日本の中でずっと真っ正義から議論されてこなかった自衛隊について、方向性をはっきりつけたい。ずっと思い通りにしてきたことにひとつの道筋をつけたい	10.3592814371	2008-03-03
救助に全力を尽くせ。事案が深い次第発着する	10.2395209581	2008-03-01
大混乱しているときに発着するよりも状況をきちんと整理することが大事だった。議員側面は防衛省まで車で 5 分くらいで（別棟で居るので）特に問題があったわけではない。（防衛省に）いけばよかったが、離れた所においても（状況は）分かる	15.1796407186	2008-03-01
今の時点で、軽々にここで諦めることが一番つらい思いをしている 地元の方々の期待に応えることは、いささかも思っていない	22.2634730539	2008-02-29
事実の確認と公表の可否に時間を要した	48.75	2008-02-26
情報が断片的だったので再度イージス艦の乗員に確認を行い、最終的に判明したのが 翌 20 日午前 2 時 47 分だった	71.25	2008-02-26
危機管理意識が欠如していたと言わざるを得ない	16.9760479042	2008-02-21

図 5: 表示手法 2 の例

3 実験と評価

3.1 発言情報の抽出手法の評価

2.1 節で提案した抽出手法の精度を検証するため実験を行った. 本手法を適用するコーパスは, 無作為に抽出したニュース記事 100 件である. ニュース記事は, MSN 産経ニュース¹ の政治分野の記事を利用した.

実験内容としては, 本手法により抽出された発言情報が, 予め人手で抽出した発言情報 (正解集合) をどの程度再現しているかを検証し, そして, 本手法により抽出された発言情報にどの程度ノイズが含まれているのかを検証する. さらに, 抽出することのできなかった発言情報から, 本手法の課題点を明らかにする.

実験結果を表 4 に記す. 抽出手法に対する再現率と適合率を求めるにあたり, 再現率 *recall*, 適合率 *precision* および *F* 値 *F-measure* を式 5, 式 6 および式 7 と定義する.

$$recall = \frac{R}{N} \quad (5)$$

$$precision = \frac{R}{C} \quad (6)$$

$$F\text{-measure} = \frac{R}{\frac{1}{2}(N+C)} \quad (7)$$

ここで R を抽出手法により抽出された正解集合, N を全正解集合, C を抽出手法により抽出された全発言情報集合である.

ニュース記事 100 件に含まれる発言情報の正解集合 N は計 85 セットであり, 本手法が抽出

¹<http://sankei.jp.msn.com/>

表 3: 重要度算出のための要素と得点の一覧

要素番号	要素名	得点
$E1$	記事内順位	$R(S) \times 15$
$E2_1$	記事構造 1: 上位 20% (上位 20-50%)	10(3)
$E2_2$	記事構造 2: 見出しの部分一致	20
$E3$	数値文	$NUMERAL \times 5$
$E4$	発言表現確信度	$CONFIDENCE \times 5$
$E5$	時系列	$TIMELINE(QT)$

した発言情報の集合 C は計 63 セットであった。そのうち、抽出手法により抽出された正解集合 R が計 59 セットであり、本手法が抽出した不正解集合が 4 セットあった。その結果、再現率は 69.41%, 適合率は 93.65%, F 値は 79.73% という結果であり、再現率の向上に課題が残ることがわかった。

表 5 に、本手法が抽出することができなかった発言情報の特徴を記す。まず、人名表記揺れの収集漏れが 9 件あった。実験時で計 83 人の政治家の表記揺れ情報を計 226 件保持しているが、保持できていない人名表記揺れが存在した。次に、発言表現の収集漏れが 2 件あった。しかし、実験時で計 122 種類の発言情報を保持しており、またニュース記事で用いられる文末表現はある程度決まっているという特徴から、発言表現に関してはほぼ網羅できていると言える。さらに、表層パターンの不一致が 10 件あった。本手法では 3 種類の表層パターンマッチングルールを作成したが、1) 1 文に 1 人の発言情報が 2 つ出現する表層パターンと、2) 発言者名が鉤括弧の後の丸括弧で表される表層パターンに対応できていなかった。

再現率をより向上させるためには、特に影響の大きかった人名表記揺れの収集漏れの解決と表層パターンマッチングルールの追加が効果的であると考えられる。

3.2 重要度算出手法の評価

プロトタイプシステムを試作し実験を行った (2009.01.16~2009.01.31)。被験者 63 人の学生

表 4: 発言情報の抽出実験の結果

人手で収集した正解集合	85
本手法が収集した正解集合	59
本手法が収集した不正解集合	4
再現率	69.41%
適合率	93.65%
F 値	79.73%

表 5: 抽出できなかった発言情報の原因内訳

抽出できなかった原因	件数	割合
人名表記揺れの収集漏れ	9	34.62%
発言表現の収集漏れ	2	7.69%
表層パターンの不一致	10	38.46%
その他	5	49.23%
合計	26	100.00%

である。実験で用いたトピックは、イージス艦衝突事故、ガソリン税の暫定税率延長問題、中国製ギョーザで中毒症状、事故米転売問題である。

評価を行うにあたり、ニュース記事理解支援方式を、発言情報を全て人手で収集した方式 (以下、提案方式 A) と、自動収集手法を適用した方式 (以下、提案方式 B) に分類する。

まず、提案方式 B の重要度算出手法の妥当性を評価するため、ある特定のニュース記事に対して、提案方式 A と提案方式 B による発言情報をユーザに提供し、どちらがより理解を深めたかを集計した。集計結果を表 6 に記す。ここで、

表 6: 設問 2-6 の集計結果

項目	設問 2-6
1. 提案方式 A	28 (44.44%)
2. 提案方式 B	20 (31.75%)
3. どちらも同じ	15 (23.81%)

重要度算出手法の妥当性とは、重要度算出手法の精度が人手の作業と比較してどの程度同等であるかという指標である。

表 6 より、人手で抽出を行った提案方式 A の方が多くユーザに選択されていることがわかるが、この差が統計的に有意の差であるのかを検証するため、重複部分である「3. どちらも同じ」の 15 票を双方に加算することで二項分布に帰着させ、二項検定を行った。二項分布の一般式を式 8 に示す。

$$p(x) = {}_n C_x p^x (1-p)^{n-x} \quad (8)$$

ここで $p(x)$ は、 n 人の被験者の中で x 人が提案方式 A を選択する確率である。 n は全被験者数、 x は提案方式 A を選択する人数、 p は提案方式 A が理論上選択される確率であり、今回 0.5 である。

そして、帰無仮説を「提案方式 A と提案方式 B の有用度に偏りが無い」と定義し、対立仮説を「提案方式 A と提案方式 B の有用度に偏りがある」と定義する。有意水準 $\alpha = 0.01$ とした。

その結果、 $p(x) = 0.060019$ となり、 $p(x) < \alpha$ が成り立たないことから、帰無仮説が採択される。よって、提案方式 A と提案方式 B の有用度には偏りがなく、双方は同等の精度を有していることが確認できた。

3.3 発言情報の表示手法に関する評価

最後に発言情報の表示手法に関する評価を行う。ユーザには、2.3 節で提案した 2 種類の表示手法を適用した発言情報を提供し、それぞれがどの程度適切であったかを集計した。集計結果を表 7 に記す。

表 7: 表示手法評価

項目	表示手法 1	表示手法 2
1. 適切だった	10	29
2. ある程度適切だった	32	27
3. 少し不適切だった	16	7
4. 不適切だった	5	0

表示手法 1 と表示手法 2 の集計結果に有意の差があるのかを検証するため χ^2 検定を行った。帰無仮説を「表示手法 1 と表示手法 2 の有用度に偏りが無い」と定義し、対立仮説を「表示手法 1 と表示手法 2 の有用度に偏りがある」と定義した。有意水準 $\alpha = 0.01$ とした。その結果、 $P = 0.000399$ となり、 $P < \alpha$ より帰無仮説が棄却され、双方の表示手法には有用度の偏りがあることが確認できた。

4 まとめ

本研究では、ニュース記事理解支援方式における、発言情報の収集コストを削減するための抽出手法と重要度算出手法を提案した。さらに、発言情報の表示手法も提案した。

評価より、抽出手法の再現率が 69.41% であり、課題を残す結果となった。そして重要度算出手法に関しては、人手での収集と同等の精度を有していることが確認できた。また発言情報の表示手法に関しては、時系列順に提供し、重要度に応じてタグクラウド手法を適用する表示手法が効果的であると確認できた。

今後の検討課題としては、たとえば、発言 A に賛成した発言 B や、発言 C を非難した発言 D など、発言情報間の関係を獲得することや、獲得した関係を適切に表現する新たな表示手法の提案が考えられる。発言間の関係と共に発言情報を提供することで、より理解の支援に有用な情報になると考えられる。

参考文献

- [1] 吉田慶章, 柿崎淑郎, 辻秀一. オントロジーを用いたニュース理解支援方式. 情報処理学会研究報告, 2008-FI-90, pp. 23–29, 2008.
- [2] 立石健二, 石黒義英, 福島俊一. インターネットからの評判情報検索. 人工知能学会誌, Vol. 19, No. 3, pp. 317–323, 2004.
- [3] 小林のぞみ, 乾健太郎, 松本裕治, 立石健二, 福島俊一. テキストマイニングによる評価表現の収集. 情報処理学会研究報告, 2005-NL-154, pp. 77–84, 2003.
- [4] 藤村滋, 豊田正史, 喜連川優. Web からの評判および評価表現抽出に関する一考察. 情報処理学会研究報告, 2004-DBS-134, pp. 461–468, 2004.
- [5] P. D. Turney. Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pp. 417–424, 2002.
- [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, Vol. 6, No. 6, pp. 1–26, 1999.
- [7] 砂山渡, 谷内田正彦. 文章の特徴を表すキーワードを発見して重要文を抽出する展望台システム. 電子情報通信学会論文誌, Vol. J84-D-1, No. 2, pp. 146–154, 2001.
- [8] M. Wasson. Using Leading Text for News Summaries: Evaluation Results and Implications for Commercial Summarization Applications. In *Proceedings of the 17th International Conference on Computational Linguistics and 36th Annual Meeting of the Association for Computational Linguistics*, pp. 1364–1368, 1998.
- [9] 仲尾由雄. 見出しを利用した新聞・レポートからのダイジェスト情報の抽出. 情報処理学会研究報告, 1997-NL-117, pp. 121–128, 1997.
- [10] 大竹清敬, 岡本大吾, 児玉充, 増山繁. 重要文抽出, 自由作成要約に対応した新聞記事要約システム YELLOW. 情報処理学会論文誌: データベース, Vol. 43, No. 2, pp. 37–47, 2002.
- [11] K. Zechner. Fast Generation of Abstracts from General Domain Text Corpora by Extracting Relevant Sentences. In *Proceedings of the 16th International Conference on Computational Linguistics*, pp. 986–989, 1996.
- [12] 吉田慶章, 柿崎淑郎, 辻秀一. タグクラウドを用いた注目情報提示方式. 情報処理学会第 69 回全国大会, 6S-2, 2007.