Vol.2015-ICS-181 No.4
2015/12/22

議論掲示板におけるテンプレートを用いた見出し生成手法

渡辺 亮輔[1,a)]　藤田 桂英[1]

概要：近年，Web 上の議論掲示板などで様々なことを議論する機会が増えてきている．しかし，その議論を把握するために全ての文章を読むことは，投稿が増えるにつれて困難になる．また，議論の構造化により議論の理解を支援する研究が行われているが，各記事を一目で理解できる見出しを自動で生成することは重要である．そこで，本研究では議論掲示板を対象に記事の見出し抽出及び生成を自動で行う手法を提案する．提案する手法では，実際に議論掲示板で発言された内容から複数パターンの見出しテンプレートを作成し，マッチしたパターンに用意されたモデルに従って見出しを生成する．また，アンケートを用いて提案した手法の有効性を評価する．

1. はじめに

近年，Web 上で自分の意見を発信する手段が増え，様々な方法によって多くの人が意見の交換を行っている．特に，SNS や議論掲示板のような Web システム上ではこれまで考えられなかった量の意見が投稿され，場所や時間という制約なしに自由に投稿できる環境ができている．このような環境において膨大な意見投稿や議論が議論集約機構可能になったことにより，これまで考えられなかった大規模な議論や交渉を行える可能性が出てきており，今後，更なる発展が期待される．さらに，Web 上のテキスト情報が爆発的に増えたことにより，「言語資源」が豊かになり，言語処理の精度向上という変容をもたらした [1]．しかし，言語処理の研究が発展する一方で，要件に対する正確な情報を得ることが難しくなっている．これは，不必要な情報も増えてしまっていることに起因する．情報増加による弊害は，意見の集約を目指す議論掲示板でも考えることができる．情報量，すなわち投稿数が増えるにつれ議論の理解に必要な情報以外の情報も増え，理解しようと全ての文章を読むことは困難になる．

これまでに議論掲示板を論理的構造を重視して容易に理解するために，議論の構造化や可視化を自動で行うシステムが提案されている [2], [3]．しかし，意見の構造化や可視化を自動で行うためには，各投稿を一目で理解できる見出しを自動で生成することが重要と考えられる．これまでに，自動見出し生成に関する研究成果はいくつも存在するが ([4], [5], [6] etc.)，オンライン大規模議論を対象にし，議論の構造まで考慮した手法はあまり多くない．

本論文では，議論掲示板を対象として投稿の見出し抽出及び生成を自動で行う手法を提案する．提案する手法では，実際に議論掲示板で発言された内容から複数パターンの見出しテンプレートを作成し，マッチしたパターンに用意されたモデルに従って見出しを生成する．また，アンケートを用いて提案した手法の有効性を評価する．実験に使用するデータは名古屋市次期総合計画 [2] で実際に投稿され，収集された意見データを用いる．アンケートに基づく評価では，本研究の手法によって抽出された文および生成された見出しに対して，様々な既存手法と本研究の手法とを比較し，評価を行う．

以下に，本論文の構成を示す．第 2 章では自動見出し生成手法に関する既存研究を示す．第 3 章ではテンプレートを用いた自動見出し抽出，および生成手法を提案する．その後，第 4 章で，評価実験結果と議論を行い，第 5 章で本論文のまとめを示す．

2. 関連研究

本章では，文書に対する見出しを生成する手法に関する既存研究を示す．

文書を構成するすべての部分文字列の中から，適切なものを抽出して連結することにより見出しとする手法が提案されている．Filippova[4] らは関連した文書の集合を短文にまとめるマルチ文圧縮を扱っている．Filippova らは文書集合から得られる有向単語グラフの中で最短経路探索をすることで見出しを生成するアプローチをしている．

文書集合に含まれる単語の中から，単語の言語尤度と重

[1] 東京農工大学 工学部 情報工学科
Institute of Information and Computer Sciences, Tokyo University of Agriculture and Technology, Koganei, Tokyo,184-8588, Japan
[a)] watanabe@katfuji.lab.tuat.ac.jp

要度が最も高くなるように抽出することで見出しを生成する手法が提案されている．廣嶋 [5] らは，統計学習により Web ページのヘッドラインを生成している．文生成モデル学習用コーパスから単語の言語尤度を，重要語選択モデル学習用コーパスから単語の重要度を求めるための重要語選択モデルをそれぞれ学習する．そして，言語尤度と重要度をもとに単語をつなぎ合わせてヘッドラインを生成している．

文書の構造に着目し，あらかじめ用意したパターンにマッチする部分を抜き出すことで見出しを生成する手法がある．議事録の構造に着目した手法 [6] では，質問答弁の「〜について質問させていただきます。」や「次に、〜です。」といった表現を構造化して見出しのテンプレートを作成し，見出しを生成している．

このように，見出し生成にはいくつものアプローチが存在するが，議論の構造や見出しの読みやすさを考慮すると抽出パターンをあらかじめ決めておき，それらに基づいて自動抽出するのが有効と考えられる．また，本テンプレートは，見出し生成や議論の構造化の際に活用できる可能性もあるから，本論文では，あらかじめ用意したパターンにマッチする部分を抜き出すことで見出しを生成する手法を参考にして，新たな手法を提案することとする．

3. テンプレートを用いた見出しの自動生成手法

本提案手法では，入力を投稿された１発言とし，出力を投稿された発言の見出しとする．本提案手法は，前処理，見出し抽出，見出し生成，見出し補完，後処理という手順からなる．

3.1 前処理

入力された文章を文単位に分ける処理や，CaboCha[7] を用いて形態素解析，文節分解および係り受け解析を行い，タグなどを付与する処理を行う．文は記号「.。！？」や改行を境界として分割する．この際，記号や URL を削除しておく．見出しに強調表現である括弧や感嘆符，疑問符などは不要であり，URL は見出しとして有用な情報を含まないため削除する必要がある．これにより記号や URL しか存在しない発言がされた場合は見出しが存在しなくなるため，見出しとしては “URL or Symbol ” と出力することにする．本研究では，CaboCha による形態素解析における品詞細分類の「非自立」「代名詞」のいずれかに分類されたものを除いた名詞，および「接頭詞」「接尾辞」も名詞とみなして連接する名詞をすべて連結した連接名詞を名詞として扱う．

3.2 見出し抽出

文単位に分けた発言の中から，その発言の見出しとなりうる文を抜き出す．既存の議事録に焦点をあてた研究では「〜について」のような表現に着目することで見出しを生成していた [6] が，議論掲示板ではこのように明確な構造は存在しない．しかし，発言者の意図を汲むと思われる文を分析すると，いくつかの表現がパターンとして利用できる．小泉 [6] らの手法に倣い，議論のパターンとして表 1 のテンプレートを作成し，３つのパターンを提案する．実際に投稿される発言は「〜だと思います。」のように丁寧語を用いた表現であるが，テンプレートでは，パターン１，パターン２で示されているパターンは単語の原形を使用する．CaboCha による形態素解析から各単語の原形を得られるため，それを用いてマッチングを行う．

表 1 見出しのテンプレート

優先度	適用するパターン
１ (パターン１)	＊＋ “必要”，＊＋ “重要”
２ (パターン２)	＊＋ “思う”，＊＋ “考える”
３ (パターン３)	返信先への賛成・反対をもとに 賛成→ポジティブワード 反対→ネガティブワード

表 1 は見出し抽出のためのパターン表を示している．以下に各パターンの詳細を示す．

パターン１： 必要，重要を含む発言
「〜が必要」や「〜が重要」という表現で示された部分は，発言者が発言の中で特に重視している場合が多い．したがって，テンプレートで最も優先度の高いパターンとする．

パターン２： 思う，考えるを含む発言
議論掲示板は参加者が意見を投稿する場であるため，「〜だと思う」,「〜と考える」といった，自らの意思を示す表現に意見が反映されている可能性が高い場合が多い．したがって，重点を置く表現であるパターン１の次に優先度の高いパターンとする．

パターン３： 返信先への賛成・反対
議論掲示板において返信により議論を進める場合，投稿者による返信先への賛成，もしくは反対の意見の表明は見出しとして重要である．そこで，発言中の各文に含まれるポジティブワードを＋１，ネガティブワードを－１として足し合わせ，その合計値の絶対値が大きい文をパターン３として取得する．このポジティブワード，ネガティブワードの判定には小林 [8] ら，東山 [9] らの感情極性辞書を使用した．

表 1 のテンプレートにおいて同じ文が複数のパターンにマッチした場合は，優先度の高い順に採用する．例えば，「〜は必要だと思う。」という発言ではパターン１の「〜が必要」とパターン２の「〜と思う」にマッチするが，優先度の高いパターン１として採用する．また，同じパターンに複数の文がマッチした場合は，返信先とのコサイン類似度の値が大きい方を採用する．返信先に含まれる m 種類の名詞の集合を $P = \{p_1, p_2, ..., p_m\}$，マッチした文に含まれる

n 種類の名詞の集合を $Q = \{q_1, q_2, ..., q_n\}$ としてベクトルとみなす．一般的には各名詞に対応する *tf-idf* の値が集合それぞれの要素になるが，本研究ではその文における名詞の出現数を要素として扱っている．コサイン類似度は以下の式で求められる．

$$cos(P, Q) = \frac{P \cdot Q}{|P||Q|} \quad (1)$$

この値が最大となる文を採用する．そのため，返信先に含まれる単語と同じ単語を多く含む文が採用されやすい．また，表 1 のテンプレートに当てはまらない場合は「パターンなし」とする．

3.3 見出し生成

3.2 において抽出した文が 10-20 文字以内に収まる場合は，抽出文を見出しとして出力する．それ以外の場合は，抽出した際にマッチしたテンプレートに対応した出力モデルに従って見出しの自動生成を行う．これらの見出し生成モデルは実際に，複数名の議論掲示板の発言から見出しを生成した結果に基づいている．

以下が提案するモデルの詳細である．

パターン１：必要，重要を含む発言

「名詞＋が，(((名詞 or 動詞) ＋ (必要だと or 重要だと)) or (動詞，名詞＋が))」
「名詞＋で，名詞＋が」
「名詞＋として，名詞＋が」
「名詞＋の，名詞＋ (が or(に，名詞＋が)or の)」
「動詞＋ような，名詞＋も」
「名詞＋を，(名詞 or 動詞) ＋ (が or べきだと or(であれば，名詞＋も))」

パターン２：思う，考えるを含む発言

「名詞＋が，名詞＋ (に or も or と)」
「名詞＋という，名詞＋は」
「名詞＋な，(名詞 or 動詞) ＋だと」
「名詞＋なのは，名詞」
「名詞＋に，(名詞 or 動詞) ＋ (が or と or とは)」
「名詞＋の，名詞＋ (が or に or(に，名詞＋が)or を)」
「名詞＋は，(名詞 or 動詞) ＋ (が or の or と)」
「名詞＋を，(名詞 or 動詞)」

パターン３：返信先への賛成・反対

該当するポジティブワード，ネガティブワードのうち BM25 の重みの大きさ上位２語を含む文節

これらのモデルとのマッチングは CaboCha で解析した文節ごとに行う．例えば，「(例文 *1*) ここまでの都市で田園風景が残るのって日本ではとても奇跡的だと思いますよ」という文がパターン２で抽出された場合は次のように分解される．

ここまでの｜都市で｜田園風景が｜残るのって｜
日本では｜とても｜奇跡的だと｜思いますよ

これに対してモデルのマッチングを行うと，「ここまでの，田園風景が」「日本では，奇跡的だと」がマッチする．このように複数マッチングした場合は，BM25([10],[11]) により文節に含まれる名詞に付与した重みの平均が大きい方を採用する．BM25 で重み付けを行う単語は，3.1 で名詞としてタグ付けしたものである．n 個の文書集合 $D = \{d_1, d_2, ..., d_n\}$ に対する単語 w の BM25 は次式で与えられる．

$$score(w, D) = \sum_{i=1}^{n} idf(w) \cdot \frac{f(w, d_i) \cdot (k_1 + 1)}{f(w, d_i) + k_1 \cdot (1 - b + b \cdot \frac{|d_i|}{avgdl})} \quad (2)$$

$$idf(w) = log\frac{N - df(w) + 0.5}{df(w) + 0.5} \quad (3)$$

$f(w, d_i)$ は単語 w が文書 d_i に出現した回数，$|d_i|$ は文書 d_i に含まれる単語数，$avgdl$ は文書集合 D の平均単語数を示す．k_1 と b は事前に決定する変数で，本研究では BM25 で一般的に用いられる $k_1 = 2.0, b = 0.75$ を採用している．idf は式 (3) を用いた．N は全文書数，$df(w)$ は全文書中で単語 w を含む文書数である *．BM25 を利用する理由は，その掲示板全体で注目されている単語により大きな値を与える手法だからである．例えば，例文 1 に対して BM25 で重み付けを行うと「ここまでの，田園風景が」は 9.68358380258，「日本では，奇跡的だと」は 46.29105633183 となり，「日本では，奇跡的だと」が採用される．

3.4 見出しの補完

3.3 の見出し生成だけでは内容が不十分になる場合がある．上記の理由として，抽出した文節を修飾する文節が抽出されていないため，抽出した文節の説明がないからである．そこで，CaboCha の係り受け解析によって得られる「係る」と「受ける」の関係にある文節を補完する．特に，抽出した文節を説明するために，対応する文節に係っている文節を補完する．具体的には，3.3 で抽出された文節に係る文節が一つだけならそれを，複数存在する場合は BM25 の重みが大きいものを採用している．

例文 1 において，以下の文を見出しとして抽出する．ある一定の文字数を超えない場合はさらに補完する．本研究では 20 文字を上限とし，補完した時に超えなければ採用を繰り返す．上限を超えたら不採用とし，補完を終了する．

残るのって日本では奇跡的だと

上記の，抽出文を補完する際に，「日本では，奇跡的だと」

*本研究では全文書が与えられた状態で行っているが，本来なら見出し生成は新しい発言が投稿されると随時行うようにするため，k_i, b を除く変数は投稿されるたびに変化する

の文節には「奇跡的だと」に係る「残るのって」「とても」があるが，どちらも名詞ではないため重み付けがされず同じ値になる．したがって，53.6159597342 を持つ「都市で」が係っている「残るのって」を採用する．以上の手順で補完を行うと次のようになる．

都市で残るのって日本では奇跡的だと

以上のような手順で見出しの補完を行い，最終的に決められた文字数内かつ発言を表現している見出しを生成できる．

3.5 後処理

生成した可読性を高くするために，生成した見出しの整形を行う．見出しの抽出は文節ごとのマッチングにより行うため，文末に助詞が残っている場合が多い．そこで，助詞等の不要な文節を置換，削除する．見出し中に出現したら削除する対象として，以下のような削除ワードリスト 1 を用意した．

削除ワードリスト 1

でしょう，です，でした

また，見出し中に出現したら置換する対象として，以下のような置換ワードを用意した．

置換ワードリスト

ありま す→ある，しなければならない→する必要，必要性がある→必要性，おります→いる，しました→した

さらに，見出しの末尾に出現する場合のみ削除する対象として，以下のような削除ワード 2 を用意した．

削除ワードリスト 2

かもしれない，考えられます，なっています，思われます，しれません，思います，感じます，考える，ように，思う，なの，かも，んだ，いう，かと，を，も，と，の，に，は，だ，ね，よ，が

削除ワード 1 および置換ワードは 1 度のマッチングにより，削除と置換を行う．一方，削除ワード 2 はマッチングと削除を繰り返し行い，見出しに変化がなくなるまで繰り返す．

例えば，例文 1 で抽出した「都市で残るのって日本では奇跡的だと」に対して後処理をを行うと以下のようになる．

都市で残るのって日本では奇跡的

以上の一連の操作により，最終的な見出しを生成する．

4. 評価実験

提案した手法を評価するため，対象文の抽出部分に対して比較実験を行う．データは名古屋市次期総合計画において実際に投稿された発言 [12] を対象として見出しを生成し，アンケートによる評価を行う．ただし，本研究の手法では，ファシリテータ（議論への介入と促進を行う者）の発言は別の構造モデルを持っていると判断したため対象としない．

見出し自動抽出部分の評価のために，正解データを作成する必要がある．そこで，名古屋市次期総合計画で実際に投稿されたすべての発言から，複数名に対して，適切だと思われる文を 1 つ選択し，正解データを作成した．正解データの作成は 1 人あたり 10 発言を対象に行い，評価者人数は 10 名であった．

その後，正解データをもとに正解率を評価する．見出し自動抽出部分に関する評価において，以下の 3 手法を比較する．

手法 1 本論文の抽出手法

手法 2 議論掲示板 COLLAGREE のシステムで現在採用されている，発言の先頭 1 文を抽出する手法である．これは，新聞など重要な内容を文章の前半に配置する構造を取る文章に有用であるとされている手法 [13] である．

手法 3 ランダムに一文を抽出した手法

3 つの手法について，選択してもらった正解データをもとにした正解率を表 2 に示す． 提案手法 (手法 1) が最も良好な結果を得られた．

手法 1 が手法 2 と比較して優れていた理由として，新聞に見られる重要文が先頭に配置されやすい構造は，議論掲示板では当てはまらないことが考えられる．実際に，発言を見ると，先頭一文は呼びかけや質問，「はい」「そうですね」と言った簡単な返事が多く，発言者の意図を汲む重要な文とは判断されなかったと考えられる．さらに，提案手法 (手法 1) と手法 3 を比較すると，提案手法が偶然正解したものではなく，提案手法が議論掲示板に対して有効であることが分かる．

また，提案手法よる正解率の詳細を表 3 に示す．パターン 1 は抽出数は少なく，正解率が高くなった．「必要」「重要」といった表現は，使用数は少ないが使用されると発言者の意図が最も反映される表現であると考えられる．パ

表 2 各手法の正解率

	正解率
手法 1	0.64
手法 2	0.37
手法 3	0.32

表 3 提案手法における各パターンごとの正解率

パターン	1	2	3	なし	合計
抽出数	22	48	30	0	100
正解数	19	29	16	0	64
割合	0.86	0.60	0.53	-	0.64

ターン２は抽出数が多く，正解率が低くなった．これは，「思う」という表現が安易に使える表現であり，発言者の主張などの重要部分以外にも多く現れたためであると考えられる．パターン３では，ポジティブ要素の多い文が抽出されることが多く，その中に重要部分が含まれていると判断されることが多い．また，今回の評価実験の対象とした発言において，パターンなしに該当する発言はなかった．これは，パターンなしに当てはまることが多いファシリテータの発言を対象外としたためである．

5. まとめ

本研究では，議論掲示板に投稿された発言を入力として，テンプレートマッチングによる自動見出し抽出および生成手法を提案した．提案する手法では，実際に議論掲示板で発言された内容から複数パターンの見出しテンプレートを生成し，マッチしたパターンに用意されたモデルに従って見出しを生成している．また，アンケートを用いた提案した手法の有効性の評価を行った．提案手法では，決まった構造を持たない発言に対し，出現する特徴表現を見つけテンプレートにすることで，見出しに利用する文の抽出では良好な結果が得られた．

今後の課題として，見出し自動生成部分に関する生成手法の評価が必要である．本論文の提案手法によって生成された見出しに対して，可読性と内容に関する評価を行う必要がある．可読性と内容に関する評価項目は西川らの評価 [14] を参考に決定してしていく予定である．さらに，見出しモデルに不十分と思われるパターン３の拡充，もしくは変更が考えられる．

謝辞

本研究は，JST，CREST の支援を受けたものである．

参考文献

[1] 喜連川優：情報爆発のこれまでとこれから，電子情報通信学会誌， Vol. 94, No. 8 (2011).
[2] 伊藤孝行， 奥村命，伊藤孝紀，秀島栄三：多人数ワークショップのための意見集約支援システム Collagree の試作と評価実験 ~ 議論プロセスの弱い構造化による意見集約支援 ~，日本経営工学会論文誌，Vol. 66, No. 2, pp. 83–108 (2015).
[3] Gürkan, A., Iandoli, L., Klein, M. and Zollo, G.: Mediating Debate Through On-line Large-scale Argumentation: Evidence from the Field, *Inf. Sci.*, Vol. 180, No. 19, pp. 3686–3702 (2010).
[4] Filippova, K.: Multi-sentence compression: finding shortest paths in word graphs, *Proceedings of the 23rd International Conference on Computational Linguistics*, Association for Computational Linguistics, pp. 322–330 (2010).
[5] 廣嶋伸章，長谷川隆明， 奥雅博：Web ページのヘッドライン生成のための統計的要約，自然言語処理， Vol. 12, No. 6, p. 113 (2005).
[6] 小泉元範，新谷虎松，大囿忠親， 白松俊：発言内容の関連性を用いた質問答弁の構造化に基づく議事録閲覧支援システム，全国大会講演論文集，Vol. 2012, No. 1, pp. 657–659 (2012).
[7] 工藤拓，松本裕治：チャンキングの段階適用による日本語係り受け解析， Vol. 43, No. 6, pp. 1834–1842 (2002).
[8] 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).
[9] 東山昌彦，乾健太郎，松本裕治：述語の選択選好性に着目した名詞評価極性の獲得，言語処理学会第 14 回年次大会論文集，pp. 584–587 (2008).
[10] Robertson, S. E., Walker, S., Jones, S., Hancock-Beaulieu, M. M., Gatford, M. et al.: Okapi at TREC-3, *NIST SPECIAL PUBLICATION SP*, pp. 109–109 (1995).
[11] Robertson, S. and Zaragoza, H.: *The probabilistic relevance framework: BM25 and beyond*, Now Publishers Inc (2009).
[12] 伊美裕麻，伊藤孝行，伊藤孝紀，秀島栄三：オンラインファシリテーション支援機構に基づく大規模意見集約システム COLLAGREE―名古屋市次期総合計画のための市民議論に向けた社会実装，情報処理学会論文誌，Vol. 56, No. 10, pp. 1996–2010 (2015).
[13] Brandow, R., Mitze, K. and Rau, L. F.: Automatic condensation of electronic publications by sentence selection, *Information Processing & Management*, Vol. 31, No. 5, pp. 675–685 (1995).
[14] 西川仁，今村賢治，別所克人，牧野俊朗，松尾義博：クエリ依存文短縮と見出し生成への応用，情報処理学会研究報告. 自然言語処理研究会報告， Vol. 2013, No. 2, pp. 1–7 (2013).