Vol.2016-ICS-183 No.6
2016/3/16

記事関連度を利用した ブログ記事からの話題抽出

松本和也[†1]　芋野美紗子[†2]　土屋誠司[†3]　渡部広一[†3]

概要：本稿では，ブログ記事で扱われている主題を一言で表すものを「話題」と定義し，ブログ記事から話題を抽出することを目的とする．しかし，ブログ記事は自由な立場で書かれており複数の話題が含まれる記事が多い．そこで，1 件のブログ記事を同じ意味のまとまりであるブロックに分割し，各ブロックから話題を抽出する手法を提案した．提案手法により，ある話題に関して記述されている範囲を把握でき，1 件のブログ記事から複数の話題を抽出することができる．本研究では，記事間の関連の強さを判断できる記事関連度を利用しブログ記事をブロックに分割する．その後，ブロックを形態素解析し話題候補語を獲得する．そして，話題候補語間の関連性を判断する関連度を利用して重要度を付与し一番高いものを話題とする．今回，記事関連度を利用することで同じ意味の文のまとまりであるブロックに分割することができ，各ブロックから適した話題を抽出することができた．

Topic Extraction from Blog Articles using Degree of Association between Articles

KAZUYA MATSUMOTO[†1]　MISAKO IMONO[†2]
SEIJI TSUCHIYA[†2]　HIROKAZU WATABE[†2]

Abstract: In this paper, the one that represents a word the subject that has been dealt with in blog article is defined as a "topic", an object of the present invention is to extract the topic from the blog article. But, blog article there are many articles that contain more than one topic is written in a free position. So, to divide the 1 blog article to block a group of the same meaning, it was proposed a method to extract the topic from each block. The proposed method, it is possible to understand the range that has been described with respect to a certain topic, it is possible to extract a plurality of topic from 1 blog article. In this study, we use the degree of association between articles that can determine the strength of association between articles to divide the blog article to block. Then, to win the topic candidate word and morphological analysis of the block. And, by utilizing the degree of association to determine the relevance of the topic between the candidate words to grant the importance to talk about things most high. This time, can be divided into blocks which are group of sentences having the same meaning by utilizing degree of association between articles, it was possible to extract the topic that is suitable from each block.

1. はじめに

近年，パソコンや携帯電話などのコンピュータの普及と共に，ブログを始めとする CGM（Consumer Generated Media : 消費者生成メディア）が一般的に使われるようになった．現在のブログ記事は，Web 上での個人の日記という側面と特定のニュースや商品に対する個人の意見を表現するメディアという側面がある．実際に，ブログ記事は商品の口コミ情報として消費者に読まれており，ブログ上での商品の評判が実際の商品購買動向に影響を及ぼしている．このような個人による情報発信は，個人の意見や主観がその内容に含まれることが大きな特徴である．特にブログ記事はその傾向が強く，ニュースに対する世間の反応や，新商品の評判を得るためにブログサイトを巡回するという閲覧形式もある．しかし，ブログ記事は自由な立場で書かれているために，書かれている話題は人によって異なり，複数の話題が含まれている．そのため，話題を知る為には，ブログ記事内で欲しい話題について書かれた場所を探し出して目を通す必要がある．

本研究では，ブログ記事で扱われている主題を一言で表せるものを「話題」と定義する．今回，本研究で利用するブログ記事は，短く主観的に書かれていることが多い．また，発信者が好んで用いる略称，顔文字などの記号，特徴的な一人称などが存在し，それらが文中で繰り返し用いられる．その為，単にブログ記事中で多く出現している単語が「話題」とはならない．そこで，語概念連想システムの中にある概念ベース[1]を用いる．概念ベースを利用することで一つの単語からさまざまな単語を連想でき，ブログ記事内の単語間の関連の強さを表す関連度を利用できる．ブログ記事中には「見た映画の感想を書き，帰りに食べたラーメンの話で締める」というように複数の「話題」が含ま

†1 同志社大学大学院理工学研究科
Graduate School of Science and Engineering, Doshisha University
†2 同志社大学　研究開発推進機構
Doshisha University Organization for Research Initiatives and Development
†3 同志社大学　理工学部
Faculty of Science and Engineering, Doshisha University

れていることがある．このように複数の話題が含まれているブログ記事から話題を抽出する際，どの話題を抽出すればよいのかが問題となる．そこで，ブログ記事を内容的な文のまとまりの単位であるブロックに分割することができれば，ある話題に関して記述されている範囲を把握することが可能である．また，ブログ記事をブロックに分割することで 1 つのブログ記事から複数の話題を抽出することができる．そこで，記事と記事の関連性を表すことができる記事関連度計算方式[2]を用いてブログ記事をブロックに分割する．その後，各ブロックから「話題」を抽出する手法を提案する．

2. 関連技術

2.1 概念ベース

概念ベースとは複数の電子国語辞書から機械的に構築された大規模な知識ベースである．ある単語を概念と定義し，その意味特徴を表す語である属性と，属性の重要度を数値で表した重みの対の集合によって構成されている．概念ベースの例を表 1 に，概念ベースの構造を図 1 に示す．

表 1　概念ベースの例

概念	属性
医者	(医師, 0.34) (患者, 0.11) (病院, 0.08) …
病院	(医院, 0.25) (手術, 0.18) (施設, 0.04) …
治す	(治療, 0.43) (医療, 0.21) (病気, 0.13) …

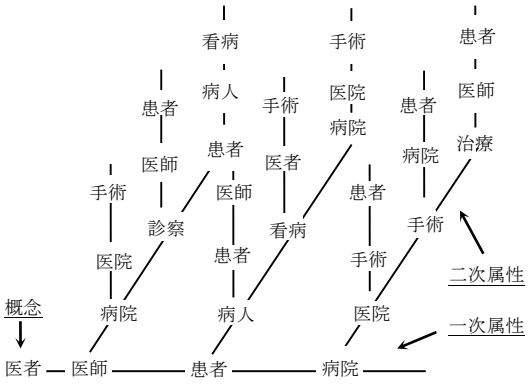

図 1　概念ベースの構造

ここで，概念自身が持つ属性を一次属性と呼ぶ．概念ベースの特徴として，全ての属性は概念としても存在している．そのため，一次属性からさらに属性を導くことができ，これを元の概念の二次属性と呼ぶ．同様に任意の次元までの属性（*n* 次属性）を導出することができる．つまり，概念ベースは，*n* 次属性の連鎖集合の構造になっている．なお，重みは概念と属性の関連性が強いほど大きな数値が付与されている．

2.2 関連度計算方式

関連度計算方式[3]とは，ある 2 つの概念間の関連の強さを定量的に表現する手法である．関連度は 0 から 1 までの実数値で表現され，概念間の関連が強いほど大きな値を示す．関連度を計算する際には，各概念の属性間にどれくらい一致する属性があるかを示した一致度を用いる．以下に，一致度の求め方と一致度を用いた関連度計算方式を示す．

概念 *A*，*B* の一次属性を a_i，b_j，重みを u_i，v_j とし，各概念が持つ属性の個数を *L* 個，*M* 個とすると，概念 *A*，*B* は，以下の式で表現される．

$$A = \{(a1,u1),\ (a2,u2),\ \cdots,\ (aL,uL)\} \quad (1)$$

$$B = \{(b1,v1),\ (b2,v2),\ \cdots,\ (bM,vM)\} \quad (2)$$

概念 *A*，*B* の一致度 *DoM(A,B)*を以下の式で定義する．

$$DoM(A,B) = \sum_{a_i=b_j} \min(u_i, v_j) \quad (3)$$

一致度では概念 *A* と概念 *B* で表記一致する属性における小さい方の重みの総和をとる．これは，両概念の属性に共通して存在する重み分は有効だと考えるためである．

また，概念 *A* の一次属性の並びを固定し，概念 *B* の各一次属性を，対応する概念 *A* の各一次属性との一致度の合計が大きい属性の順に以下の式並べ替える．

$$Bx = ((bx1, vx1),\ (bx2, vx2)\ ,\ \cdot\ ,\ \cdot\ ,\ (bxL, vxL)) \quad (4)$$

概念 *A*，*B* の関連度 *DoA(A,B)*は以下の式で表現される．

$$DoA(A,B) = \sum_i DoM(a_i, b_{Xi}) \times \frac{(u_i + v_{Xi})}{2} \times \frac{\min(u_i, v_{Xi})}{\max(u_i, v_{Xi})} \quad (5)$$

概念ベースに定義されている 2 つの概念間の関連の強さを定量的に表現する手法である．関連度は 0.0 から 1.0 の間の実数値で表され，概念間の関連が強いほど大きな数値となる．例えば概念「自動車」に対して，「車」，「自転車」，「馬」の関連の強さは表 4 に示す通りとなり，コンピュータは「自転車」と関連が強いのは 3 つの内，「車」であるということを判断できる．

2.3 記事関連度計算方式

関連度計算方式が語と語の関連の強さを表現するのに対して，記事関連度計算方式は記事と記事の関連の強さを定量的に表現することができる．定量化された値を記事関連度と呼ぶ．

2.4 AF（オートフィードバック）

AF[4]とは，未定義語（概念ベースに定義されていない概念）の意味的特徴をあらわす単語（属性）とその重要性を表す重みの組を Web を用いて自動的に構築する手法である．まず，ロボット型検索エンジンを用いて未定義語の検索を行う．そして，獲得した検索結果ページから形態素解析を行い，自立語を概念ベースに存在する語に限定して抽出する．その後，獲得した検索結果ページ内での自立語の出現頻度と *Web-idf* を用いて，*tf・Web-idf* 重みづけを行う．*Web-idf* とは Web にある文書のみを用いて索引語の出現頻度を考慮する手法である．*Web-idf* では式(6)の *N* を Google が保有している日本語のページ数，*df(t)*を索引語 *t* を Google で検索を行ったときのヒット件数とする．なお，Google は全言語において保有しているページ数は公開されているが，日本語のページとして保有している数は公開されていない

ため，日本語の文書として最も使われている「は」で検索を行ったヒット件数（1,260.000,000 件（2015 年 1 月現在））を Google が保有している日本語の前ページ数としている．

$$Web\text{-}idf\ (t) = \log_2 \frac{N}{df(t)} + 1 \qquad (6)$$

そして，重み順に並べ替えた自立語とその重みの対の集合を X の属性とする．この手法を用いて未定義語 X の属性 x_i とその重み w_i の組を構築する．未定義語 X の属性 x_i とその重み w_i の組を以下に示す．

$$X=\{(a1,w1),\ (a2,w2),\ \cdots,\ (aL,wn)\} \qquad (7)$$

例として，未定義語「同志社大学」を入力した時の結果を表 2 に示す．

表 2　入力「同志社大学」における処理結果

属性	重み
大学	141.592
研究	130.532
:	:

3. 提案手法

本研究では，まず図 2 のようなブログ記事を句点によって各文に分割する．その後，内容的な文のまとまりの単位であるブロックに統合する．そして，1 つのブログ記事から分割された各ブロックで扱われている主題を一言で表しているものを「話題」として 1 つ抽出する．

すごーい！サンフレッチェはやっぱりさいこー！(≧∇≦)
最初に失点したときはもうダメ・・これだから広島は・って思ってチャンネルをそっと消した・・のは事実だけど・。信じてたよ広島・・！それにしてもチャンピオンシップ、結構面白いかも(^_-)☆。地上波放送局が広島じゃなくてガンバ贔屓過ぎた点が気になったけど(笑)消してごめんなさい o(_ _*)o。後半４６分で青山さんがクロスを上げて、反応した佐々木さんがヘディングシュートで後半ロスタイムの土壇場での同点の展開なんてドラマでも出来すぎた展開。あっ・・一応、カープのブログなので少し野球の話も。カープは投手を補強したほうがいいかも。理由はイニングイーターのマエケンが抜けたら絶対投手不足。それより、サンフレッチェ!おめでと～！広島に感動をありがとう(*^^*)。

図 2　ブログ記事の例

ブログ記事をブロックに分割した後，各ブロックから「話題」を抽出する．「話題」となる主題を一言で表しているものは名詞で書かれているものが多い．そこで，文章を単語ごとに分割し品詞情報を得ることができる形態素解析を各ブロックに対して行う．この形態素解析には形態素解析ソフト茶筌[5]を用いる．しかし，名詞の中でも茶筌の辞書に登録されていない名詞は形態素解析の結果，未知語と判断されてしまう．そこで，形態素解析の結果，名詞と未知語に分割される単語を用いて，それ以外の品詞の単語は用いない．

提案手法の概要を図 3 に示す．以下のそれぞれの節において，図 2 のブログ記事をブロックに分割する方法，ブロックから話題を抽出する方法についての詳細を述べる．ブログ記事をブロックに分割する手法では，まず，句点によりブログ記事を文単位に分割する．文単位に分割した後，記事関連度によって，各文をブロックに統合する．ブロックから話題を抽出する手法では，まず，ブロックから形態素解析により話題候補語を抽出する．そして，抽出した話題候補語に重要度を設定し，その重要度が一番高かった話題候補語を「話題」として抽出する．

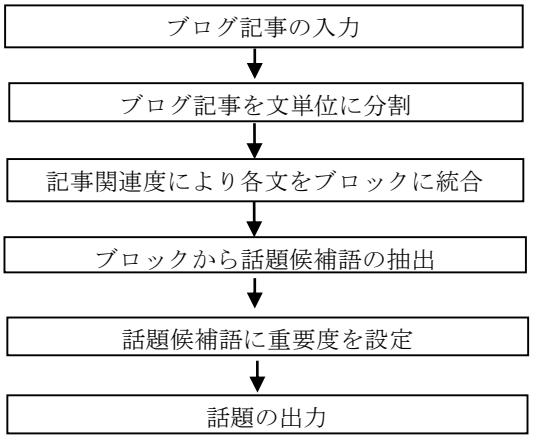

図 3　提案システムの流れ

3.1 文単位に分割

与えられたブログ記事を文単位に分割し N 個の文を得る．ここで，文単位とは句点で区切られる一文を指す．ただし，会話文（カギ括弧で囲まれた文）中に現れる句点では，会話内容を途中で切るため分割を行わない．

3.2 記事関連度を用いてブログ記事をブロックに分割

3.1 節で述べた通りにブログ記事を文単位に分割した後，この各文を 2.3 節で述べた記事関連度を利用してブロックに統合する．まず，図 2 のブログ記事の 1 文目を文番号 1 とする．このように各文に対して文番号をつける．そして，文番号 1 を基準に各文との記事関連度を計算する．その記事関連度が閾値以上であれば 2 つの文は同じ意味を持つ文であると判断しこの 2 つの文をブロックとする．ここで，閾値は実験的に求めた値として 0.07 を利用する．ブロックとなった場合は，文番号 1 を含むブロックと各文で記事関連度を算出し，閾値以上であった文をブロックの後ろに追加していき，この処理を最後の文番号まで続ける．そして，ブロックに追加されなかった文のうち文番号が一番小さい文を新たに基準とし，上記の手順を繰り返す．こうすることでブログ記事の各文をブロックに統合する．図 2 を例としてブロックに分割する方法について説明する．まず，文番号 1 と文番号 2 の記事関連度を計算する．記事関連度を計算する際，文番号 1 からは「失点」，「広島」などの索引語を抽出する．また，文番号 2 からは「広島」，「チャンピオンシップ」などの索引語を抽出し，記事関連度を求める．

文番号 1 と文番号 2 の記事関連度は 0.093 であり，閾値以上であるため文番号 1 と文番号 2 は同じ意味を持つ文であると判断でき，図 4 のようにブロックとする．

すごーい！サンフレッチェはやっぱりさいこー！(≧∇≦)最初に失点したときはもうダメ・・これだから広島は･って思ってチャンネルをそっと消した･･のは事実だけど･。信じてたよ広島・・！それにしてもチャンピオンシップ、結構面白いかも(^_-)☆。

図 4　文番号 1 と文番号 2 で構成されたブロック

次に，図 4 のブロックと文番号 3 で記事関連度を算出する．ブロックと文番号 3 の記事関連度は 0.198 であり，閾値以上であるため図 4 のブロックと文番号 3 は同じ意味を持つと判断でき，図 4 のブロックの最後に文番号 3 を追加する．その結果を図 5 に示す．

すごーい！サンフレッチェはやっぱりさいこー！(≧∇≦)最初に失点したときはもうダメ・・これだから広島は･って思ってチャンネルをそっと消した･･のは事実だけど･。信じてたよ広島・・！それにしてもチャンピオンシップ、結構面白いかも(^_-)☆。地上波放送局が広島じゃなくてガンバ贔屓過ぎた点が気になったけど(笑)消してごめんなさい o(_ _*)o。

図 5　文番号 3 を後ろに追加したブロック

このように，文番号 1 を基準に記事関連度で閾値以上の文をブロックの後ろに追加していくと最終的に図 6 のようなブロック 1 に分割される．

すごーい！サンフレッチェはやっぱりさいこー！(≧∇≦)最初に失点したときはもうダメ･･これだから広島は･って思ってチャンネルをそっと消した･･のは事実だけど･。信じてたよ広島・・！それにしてもチャンピオンシップ、結構面白いかも(^_-)☆。地上波放送局が広島じゃなくてガンバ贔屓過ぎた点が気になったけど(笑)消してごめんなさい o(_ _*)o。後半４６分で青山さんがクロスを上げて、反応した佐々木さんがヘディングシュートで後半ロスタイムの土壇場での同点の展開なんてドラマでも出来すぎた展開。それより、サンフレッチェ!おめでと～！広島に感動をありがとう(*^^*)。

図 6　ブロック 1

図 6 のブロック 1 ができたが，まだ記事関連度で閾値以下であった文がある．そのため，次はブロックに分割されていない文番号 5 を基準に文番号 6，文番号 7 との記事関連度を算出する．文番号 5 と文番号 6 の記事関連度は 0.088 であり閾値以上であるため同じ意味を持つ文であると判断でき文番号 5 と文番号 6 は図 7 のようにブロックとなる．

あっ・・一応、カープのブログなので少し野球の話も。カープは投手を補強したほうがいいかも。

図 7　文番号 5 と文番号 6 で構成されたブロック

最後に，図 7 のブロックと文番号 7 との記事関連度を算出する．図 7 のブロックと文番号 7 の記事関連度は 0.224 となり，閾値以上であるため図 7 のブロックと文番号 7 は同じ意味を持つと判断でき，図 7 のブロックの最後に文番号 7 を追加する．その結果を図 8 に示す．

あっ・・一応、カープのブログなので少し野球の話も。カープは投手を補強したほうがいいかも。理由はイニングイーターのマエケンが抜けたら絶対投手不足。

図 8　文番号 7 を後ろに追加したブロック

最終的に，図 2 のブログ記事は図 6 のブロック 1 と図 8 のブロック 2 の 2 つのブロックに分割される．

3.3 各ブロックから話題候補語の抽出

話題候補語を抽出する際に，各ブロックを形態素解析する．3 章で述べた通り今回，「話題」となる語は主題を一言で表しているものであるため，茶筌を用いて名詞と未知語を抽出する．しかし，形態素解析の際，話題候補語は形態素といわれる最小の単位になっているため複合語を考慮できていない．例えば，図 6 のブロック 1 の話題候補語である「地上波放送局」を茶筌にかけると「地上派」，「放送」，「局」の 3 つの形態素に分割されてしまう．そこで，本稿では茶筌の出力結果の前後関係にある語が「名詞＋名詞」，「名詞＋未知語」，「未知語＋名詞」という前後関係のある語を複合語として獲得する．また，3 語以上で構成される複合語についても同様に前後関係を見て複合語として獲得する．しかし，名詞の中で「こと」や「もの」を表す非自立語は話題候補語として不適切であると判断し話題候補語から削除する．図 6 のブロック 1 からは「サンフレッチェ」，「最初」，「失点」など 23 個の話題候補語が抽出される．

ブロック 2 の 2 つのブロックに分割される．

3.4 話題候補語に重要度を設定

本節では，3.3 節で抽出した話題候補語に重要度を付与する．そして，その重要度が最も大きくなった単語を「話題」として抽出する．以下の項で話題候補語に *tf* による重要度を付与する方法，話題候補語の出現順序により重要度を付与する方法，話題候補語の文法により重要度を付与する方法，話題候補語間の関連度により重要度を付与する方法について述べる．

3.4.1 TF による重要度の設定

TF[6]とは，索引語頻度を意味し，索引語がどれだけ多く，文書中に出現するかを示している．今回，索引語は話題候補語とし，ブロック d 中に出現する話題候補語 t の頻度をブロック d 中のすべての話題候補語の総数で割った $tf(t,d)$ を利用する．何度も繰り返し使われる語は，重要であると考えられる．そのため，各ブロック内の話題候補語の *tf* を計算し，それを話題候補語の重要度とする．

3.4.2 出現順序による重要度の設定

ブログ記事において，「話題」となる単語がブログ記事中

のどの場所に出現しているのかを調査するためブログ記事100件を対象に人手で「話題」を抽出した．その結果を表3に示す．

表3　ブログ記事100件における「話題」の存在位置

一文目	二文目	三文目	四文目	五文目以降
83件	6件	8件	2件	1件

表3より，ブログ記事の出だしの一文目に「話題」が多く存在していたため，ブログ記事を書く際，一文目に「話題」を述べる傾向にあることがわかった．この結果を利用して，出だしの一文目に出現した話題候補語の重要度を大きくする．重要度の倍率は実験的に決定する．まず *tf* のみでブログ記事100件を対象に「話題」を抽出し，評価を行った．評価の際，抽出した話題が理想の話題と一致した場合は○，理想の話題とは一致しないが意味が近いものは△，明らかに違うものは×とした．*tf* のみの評価結果を表4に示す．この表4の評価結果を基準として，一文目に出現する話題候補語の重要度を2倍，3倍，4倍，5倍したときの評価結果を表5に示す．

表4　tfのみの評価結果

	tfのみ
○	30件
△	11件
×	59件

表4の評価結果を基準として，一文目に出現する話題候補語の重要度を2倍，3倍，4倍，5倍したときの評価結果を表5に示す．

表5　一文目の話題候補語の倍率を変えた評価結果

	一文目を2倍	一文目を3倍	一文目を4倍	一文目を5倍
○	34件	36件	32件	32件
△	14件	14件	13件	13件
×	52件	50件	55件	55件

表5の評価結果より，一文目の話題候補語の重要度を3倍にしたときが一番○と△の割合が高かった．そこで，一文目の話題候補語の重要度を3倍にする．

3.4.3 出現順序による重要度の設定

文の要素には「主語」，「動詞」，「目的語」，「補語」があり，この4つは文を構成する重要な部分である．そこで，3.3.2項で述べたブログ記事100件を対象に人手で抽出した「話題」がブログ記事の文においてどうような要素になっているのかを調べた結果を表6に示す．

表6　ブログ記事100件における「話題」の存在位置

主語	目的語	サ変接続	その他
42件	37件	9件	12件

表6より，「話題」には「主語」，「目的語」が多いことがわかった．この結果を利用して，話題候補語がこれらの2つに該当する場合，重要度を大きくする．主語とは「～は」，「～が」に当たる文節のことで，文の主体のことであり，目的語には「～を」のような動作の直接の対象となるものである直接目的語と「～に」のような直接の対象とはならず，間接的に動作と関係を表す間接目的語がある．そこで，話題候補語の後ろに助詞の「は」，「が」があればその話題候補語は主語と判断し，「を」，「に」の場合はその話題候補語は目的語として判断する．まず，話題候補語が主語である場合の重要度の倍率を決めるため評価を行った．話題候補語が主語である場合の重要度を2倍，3倍，4倍，5倍したときの評価結果を表7に示す．

表7　話題候補語が主語のときの倍率を変えた評価結果

	主語を2倍	主語を3倍	主語を4倍	主語を5倍
○	45件	42件	42件	40件
△	21件	20件	19件	18件
×	34件	38件	39件	42件

表7の評価結果より，話題候補語が主語の場合，重要度を2倍にしたときが一番○と△の割合が高かった．そこで，話題候補語が主語の場合，重要度を2倍にする．次に，話題候補語が目的語である場合の重要度の倍率を決めるため評価を行った．話題候補語が目的語である場合の重要度を2倍，3倍，4倍，5倍したときの評価結果を表8に示す．このとき，話題候補語が主語の場合は2倍している．

表8　話題候補語が目的語のとき倍率を変えた評価結果

	目的語を2倍	目的語を3倍	目的語を4倍	目的語を5倍
○	45件	42件	42件	40件
△	21件	20件	19件	18件
×	34件	38件	39件	42件

表8の評価結果より話題候補語が目的語の場合，重要度を2倍にしたときが一番○と△の割合が高かった．そこで，話題候補語が目的語の場合，重要度を2倍にする．

3.4.4 関連度による重要度の設定

関連度は，値が高いほど語と語の関連が深いことを意味する．そのため，この関連度を用いることでブログ記事内での各話題候補語間の関連性を重要度に付与することができる．ブロックは，同じ意味を持つ文で統合されているため，ブロックの中で「話題」となりうる話題候補語は他の話題候補語との関連が高くなるはずである．そこで，ブロック中の話題候補語と同じブロック中の他の話題候補語すべてを関連度計算し，その総和を求める．ここで，話題候補語が未知語の場合，2.4節で説明したAFを利用して属性と重みを取得し，関連度を算出する．例えば，図6のブロック1の話題候補語「サンフレッチェ」の属性には「サッカー」，「チーム」，「ゴール」など他の話題候補語と関連がある属性を取得することができる．ここで，図6のブロック1の話題候補語「サンフレッチェ」では，図6の他の話

題候補語との関連度を算出していき，関連度の総和を求めこれまでに付与された重要度に足し合わせる．

3.5 話題の出力

3.3.1 項から 3.3.4 項までで付与した重要度の中で一番重要度が高い話題候補語を「話題」として出力する．図 6 のブロック 1 からは「サンフレッチェ」が話題となる．

4. 評価方法

評価には，ブログ記事を適したブロックに分割できているのかとブロックから適した「話題」が抽出できているのかという 2 つの評価を行う．評価セットには，実際のブログサイトから無作為に抽出したブログ記事 150 件を用いる．今回, 評価に用いたブログ記事 150 件の詳細を表 9 に示す．

表 9　評価セットの詳細

合計文字数	合計文数	平均文字数	平均文数
79472	2045	529.81	13.63

4.1 ブロック分割の評価と考察

最初に，評価者に評価セットのブログ記事をブロックに分割してもらいそれを正解ブロックとする．そして，提案手法で抽出されたブロックと正解ブロックを用いて適合率, 再現率，F 値で評価する．適合率と再現率の値がともに大きいときに F 値は大きい値をとる．評価における正解と不正解の判断は次のとおりである．正解ブロックに対し小さいブロックは正解とし，大きいブロックは不正解とする．一つの正解ブロックに対して複数のブロックが抽出された場合，一つのブロックだけを正解として扱い，残りのブロックは不正解として扱う．まず，一つ目の基準は，過大なブロックは本来の分割箇所で分割できなかったため不正解と扱う設定である．二つ目の基準は，過剰に分割したときの分割箇所が正解ブロックに比べて不要な分割が増えることから，その増加分は不正解と扱う設定である．評価結果を表 10 に示す．

$$\text{適合率} = \frac{\text{正しく検出されたブロックの総数}}{\text{手法で検出したブロックの総数}} \quad (8)$$

$$\text{再現率} = \frac{\text{正しく検出されたブロックの総数}}{\text{正解ブロックの総数}} \quad (9)$$

$$\text{F 値} = \frac{2\times\text{適合率}\times\text{再現率}}{\text{適合率}+\text{再現率}} \quad (10)$$

表 10　ブロック分割の評価結果

閾値	適合率	再現率	F 値
0.08	40.5%	71.8%	51.8%
0.07	42.1%	70.7%	52.8%
0.06	42.9%	67.7%	52.5%

評価セット 150 件をブロックに分割した際，1 文のみで構成されたブロックは 572 個のブロック中 253 個もあった．例えば, 図 9 のブログ記事をブロックに分割すると図 10 から図 13 のブロックに分割される．

> 日本でもフォルクスワーゲンのディーゼルエンジンに関するニュースは大きく報じられていますよね。せっかく、長年日本の輸入車ナンバーワンの地位にいたのに、このニュースでしばらくは他社が優位に立つのは間違いなさそう。それにしても、今回のことでディーゼルエンジン全般に対してのイメージが悪くってしまったとしたら、その責任は単に 1 メーカーの問題だけでは済まないですよね。ただでさえ日本のユーザーはディーゼルエンジンにあまりいいイメージを持っていないし、それでも、各メーカーの努力によってだんだんと普及してきた矢先のことなので残念ですね。個人的にフォルクスワーゲンを所有していることもあり、今回のことはすごく気になります。一番かわいそうなのは、販売店のセールスさんかな。常に矢面に立たされているわけですから。先日販売店からこんな手紙が届きました。一刻も早くこの不正問題が解決することを願います。

図 9　評価セットのブログ記事の例

> 日本でもフォルクスワーゲンのディーゼルエンジンに関するニュースは大きく報じられていますよね。せっかく、長年日本の輸入車ナンバーワンの地位にいたのに、このニュースでしばらくは他社が優位に立つのは間違いなさそう。それにしても、今回のことでディーゼルエンジン全般に対してのイメージが悪くってしまったとしたら、その責任は単に 1 メーカーの問題だけでは済まないですよね。ただでさえ日本のユーザーはディーゼルエンジンにあまりいいイメージを持っていないし、それでも、各メーカーの努力によってだんだんと普及してきた矢先のことなので残念ですね。個人的にフォルクスワーゲンを所有していることもあり、今回のことはすごく気になります。

図 10　図 9 のブログ記事から分割したブロック 1

> 一番かわいそうなのは、販売店のセールスさんかな。先日販売店からこんな手紙が届きました。

図 11　図 9 のブログ記事から分割したブロック 2

> 常に矢面に立たされているわけですから。

図 12　図 9 のブログ記事から分割したブロック 3

> 一刻も早くこの不正問題が解決することを願います。

図 13　図 9 のブログ記事から分割したブロック 4

図 12，図 13 のように 1 文が短い文で記事関連度を算出すると低い値になってしまい 1 文のみのブロックになってしまう．そこで，1 文のブロックを少なくするために，文字数が少ない文は前の文と統合させる方法が考えられる．

4.2 話題抽出の評価と考察

ブログ記事 150 件からブロックに分割した 2 文以上のブロック 319 個より「話題」を抽出し，評価者 3 人により目

視で評価をする．図 12，図 13 のように 1 文のみのブロックには話題が存在しないものも多いため今回 1 文のみのブロックは評価には用いない．理想の話題と一致するものは○，理想の話題と一致しないが意味的に近いものは△，明らかに違うものは×の三段階で○が 2 点，△が 1 点，×が 0 点となる．3 人の評価の合計が 5 点以上ならば○，4 点と 3 点ならば△，2 点以下ならば×とする．

表 11　話題抽出の評価結果

○	×	△
50.8%	16%	33.2%

今回，話題抽出の評価に利用したブロックの例を以下の図 14 から図 16 に示す．またそのブロックから提案手法によって出力された「話題」とその「話題」の正解・不正解(○，△，×)を表 12 に示す．

今日のデザートは、「苺ソースのクレームダンジュ」。苺ソースがハート状になってかわいい。味は、濃厚なクリームチーズと苺ソースの酸味がマッチしておいしい。

図 14　評価に利用したブロックの例 1

某（それがし）、去る H27 年の年末に、『スーパーマリオメーカー』の WiiU セット買っちゃいましてね・・。それ以来今日まで、このスーパーマリオメーカーにハマっております昨日なんて夜 2 時にコース完成させるまでずっとやってたし・・・健康と生活習慣上良くないですね。だけどあれはハマる！ハマる要素がたくさんある！そもそもオリジナルのコース作れるというのがゲームコンセプトとして魅力的で、しかもそれをネットワークに繋げて投稿して世界中の人に遊んでもらえるってんですから、腕によりをかけて作り込みたくなるってモンです。勿論、作るだけでなくて、他の人が作ったオリジナルコースを遊ぶことも、Wi-Fi などのアクセスポイントに接続すれば可能。今まで幾つものオリジナルコースをプレイしましたが、世の中には創作の上手い人がたくさんいるみたい。それこそ公式スタッフ顔負けってぐらいの！もっとたくさんの人に遊んでもらえるように、もっと凝った、だけど遊びやすいってなコースを作っていかねばなぁ。

図 15　評価に利用したブロックの例 2

ヤクルトの山田哲人内野手が、9 月 6 日(日）に広島戦で今季 30 盗塁を決め、同一シーズン打率 3 割、30 本塁打、30 盗塁をマークする「トリプルスリー」の条件を現時点で満たしました。打率はシーズン終了まで判りませんが、現在.333 で規定打席にも達しているので、史上 9 人目の達成はほぼ確実ではないかと思います。それにしても凄い！特に昨季は、.324、29 本塁打に対し、盗塁が 15 個と一番難関と思われた部門なので、最後まで挑戦の気持ちでやりきって欲しいですね！

図 16　評価に利用したブロックの例 3

表 12　話題抽出の評価結果

ブロック	出力した話題	正解・不正解
図 14	苺ソース	△
図 15	スーパーマリオメーカー	○
図 16	盗塁	×

提案手法では関連度が計算できない話題候補語にも AF を利用し，web から属性と重みを取得したことで関連度計算を行うことができたため適した話題を抽出できたと考えられる．図 14 のブロックは「苺ソースのクレームダンジュ」が話題と考えられる．しかし，提案手法では「苺ソース」が話題として出力される．このような「の」という前置詞を含む話題には対応できない．そこで，名詞＋「の」＋名詞は一つの単語とし，話題候補語として抽出するといったことが必要であると考えられる．図 15 のブロックからは「スーパーマリオメーカー」が話題として抽出される．話題候補語「スーパーマリオメーカー」の属性には「ゲーム」，「コース」，「ネット」など他の話題候補語と関連がある属性が取得でき，関連度が高く「スーパーマリオメーカー」という適した話題を抽出することができた．図 16 のブロックは，「山田哲人」が話題と考えられる．しかし，提案手法では「盗塁」が話題として出力される．これは「盗塁」が一番多くブロック内に出現しており，他の話題候補語とも関連が強いため話題として出力されたと考えられる．

5. おわりに

本研究では，ブログ記事を同じ意味を持つ文のまとまりであるブロックに分割し，各ブロックから話題を抽出する手法を提案した．ブロックに分割する手法では，EMD を用いた記事関連度計算方式を利用することで，記事関連度を算出する際の文の長さや自立語の数を考慮でき，適合率 42.1%，再現率 70.7%，F 値 52.8%の結果を得た．また，ブロックごとで話題抽出を行うことで 1 件のブログ記事から複数の話題を抽出することができた．話題抽出の際，AF を利用して未知語に対して web から属性と重みを取得することで関連度により重要度を付与することができ 66.8%の精度を得られた．しかし，記事関連度では，文字数が少ない文に対応できず，一文で構成されるブロックが多く存在する．話題抽出では英字や前置詞を含んだ「話題」を抽出することが今後の課題である．

謝辞　本研究の一部は，科学研究費補助金（若手研究（B）24700215）の補助を受けて行った．

参考文献

[1] 笠原要，松澤和光，石川勉，“国語辞書を利用した日常語の類似性判別”，情報処理学会論文誌，Vol.38，No.7，pp.1272–1283，1997.
[2] 藤江悠五，渡部広一，河岡司，“概念ベースと Earth Mover’s Distance を用いた文書検索”，情報科学技術フォーラム，FIT2002，pp.159-160，2002.

[3] 荒木孝允，奥村紀之，渡部広一，河岡司，“比較対象概念の共通属性を重視する動的関連度計算方式”，同志社大学理工学研究報告，Vol.48，No.3，pp.14–24，2007.
[4] 徳永健伸，情報検索と言語処理，東京大学出版会，1999.
[5] ChaSen – 形態素解析器，奈良先端科学技術大学院大学情報科学研究科自然言語処理学講座(松本研究室) .
[6] 徳永健伸，情報検索と言語処理，東京大学出版会，1999.