電子掲示板の主要人物に関するストーリー要約

Story Extraction Related to Key Person in Bulletin Board System

田村　幸寛　　砂山　渡
Yukihiro Tamura　Wataru Sunayama

広島市立大学大学院情報科学研究科
Graduate School of Information Sciences,　Hiroshima City University

Recently, bulletin board systems are common in the internet. However, the written comments are so enormous that we cannot grasp the contents in a short time. In this paper, we propose a system that extracts story about a key person, who is the most frequently appeared in a bulletin board system. The system extracts comments written by the key person and others communicating with, in a bulletin board site “2-channel”.

1.　はじめに

近年，インターネットの普及により「2 ちゃんねる」や「Yahoo!掲示板」のような電子掲示板が注目を集めている．電子掲示板は匿名性が高く，PC 以外にも携帯電話などからも読み書き可能となっている．また，上記 2 つに代表されるような大型掲示板だけでなく，レンタル式の電子掲示板も存在し，個人で立ち上げているブログやホームページに組み込まれていることがある．このような背景から，多くの人々が電子掲示板を利用するため，電子掲示板では情報交換や議論などが盛んに行われるようになってきている．しかし，誰もが利用可能なため電子掲示板の数は膨大となり，1 つのスレッド（コメントとレスを一まとめにくくった時の名称）においても大量のレス（書き込み，返信）があることが多くなっている．そのため，閲覧しているスレッドにはどのように話が進んでいるかをすぐに把握することは難しいと考えられる．

本研究では，大量に書き込まれたスレッドの内容をすばやく理解するために，スレッドのストーリーを要約するシステムの構築を目指す．

以下，本論文では 2. 節では関連研究として現在行われている電子掲示板に関する研究について述べる．3. 章では提案システムの構成を述べる．4. 章では本システムの有効性を確認するための評価実験とその実験結果，および考察について述べる．最後に，5. 章で結論を述べて本論文を締めくくる．

2.　関連研究

本章では本研究の関連研究として電子掲示板を用いた研究，対話データの要約をまとめる．

2.1　電子掲示板を用いた研究

本節では，既存の電子掲示板を用いた研究について述べる．まず，どのような話題がトピックになっており，どのような会話が行われているかを掲示する手法 [1] がある．この研究は Yahoo!掲示板のスレッドのレスを列挙し，それぞれのレスでトピックになる部分をハイライトで表示するというものであった．また，電子掲示板の評判情報を抽出する手法 [2] もある．これは商品やサービスなどを評価する文章を肯定的なのか否定的なのかを自動で分類するものであった．本研究では，要約に必要なレスのみを取り出すといった点で前者とは異なり，話の流れのあるスレッドを使用するという点で後者とは異なる．

連絡先: 田村幸寛，広島市立大学大学院情報科学研究科，広島県広島市安佐南区大塚東三丁目 4 番 1 号，tamura@sys.info.hiroshima-cu.ac.jp

2.2　対話データの要約

本節では，既存の対話データを要約する研究について述べる．まず，電子掲示板での対話履歴から協調学習に有効な情報を抽出して視覚化する手法 [3] がある．また，コールセンターの対話データから重要文を抽出する手法 [4] もある．これらは，対話データ，つまり会話を要約するための研究だが，本研究では電子掲示板の中での会話を要約するという点で異なっている．

3.　提案システムの構成

本研究では掲示板のストーリーを主要書き込み人物（メイン）とその他の人（サブ）との会話と定義する．

これらのレスを抽出する提案システムの構成図を図 1 に示す．提案システムは，入力を 2 ちゃんねるのメインが存在するスレッドの html ファイルとし，「無条件で除外するレスの除去」，「メインとなる人物の特定」，「特定されたメインのレスの抽出」，「メインとの明示的なアンカーをもつサブのレスの抽出」，「メインとの暗示的なアンカーをもつサブのレスの抽出」を行い，絞り込まれたレスを html ファイルとして出力する．

3.1　入力：スレッドの html ファイル

提案システムの入力はメインが存在するスレッドの html ファイルとする．今回用いた html ファイルは Live2ch と呼ばれる 2 ちゃんねるのビューアから取得したものを用いている．

3.2　除外対象レスの除去

本節では，入力されたスレッドから除外対象のレスを除去する方法を説明する．2 ちゃんねるのスレッドにはアスキーアート（以下 AA）のような不必要なレスが存在することが多い．AA に対しては「茶筌」[5] が記号と判別する文字が 1 レス中に 11 個以上含まれていれば除去する．また，ストップワードリストをあらかじめ作成しておき，その中の単語が 1 つでも含まれているレスを除去する．

3.3　メインとなる人物の特定

本節では，スレッドの中から主要書き込み人物であるメインのレスを特定する方法を説明する．メインとなる人物はサブの人物たちと区別が付くようにするため，独自の名前を自分で設定することが多い．したがって，スレッドの全レスの名前を抽

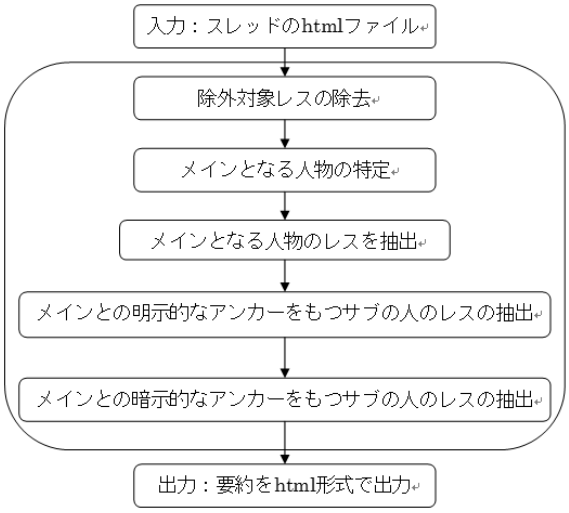

図 1: 提案システムの構成図

出し名前を取得して判断すればよい．このとき，電子掲示板は名前を入力せずに書き込まれたレスに対してはデフォルトとなる名前を付ける．つまり，全レスの名前を比較するときには 1 番多く使われている名前ではなく，2 番目に多く使われている名前で判断すればよい．また，メインが固有の名前を入力するのを忘れたり，名前を変更したりすることがあるため，掲示板の種類によっては存在する ID と関連付けるようにする．そして取られた ID が用いられているレスの名前が 1 番多いものでなければその名前も候補に入れる．これらの動作を 2 回行うことにより，メインのレスを特定する．

3.4 メインとなる人物のレスを抽出

3.3 節で特定されたレスの抽出する．このとき，3.2 節で除外対象レスとなっていた場合でも抽出する．

3.5 メインとの明示的なアンカーを持つサブの人のレスの抽出

本節では，明示的なアンカーの意味と，メインと明示的なアンカーを持つサブのレスを抽出する方法について説明する．アンカーとは今回用いた 2 ちゃんねるの場合「>>」という記号を用いた元発言のレス番号に対してリンク設定を行うことにより，そのレス番号の発言や人物に対しての返信であるかを明確にしているものとなっている．「>>」は，この記号の後に発言番号を入力することで使用する．例えば，「>>20」と行えば，発言番号が 20 のレスに対しての返信ということになる．

これが 3.4 節で抽出されたメインのレスに対して行われている場合は，3.2 節で除去されたレスでなければ抽出するようにする．また，メインのレスが使用している場合は，ストーリーが成り立たなくなってしまうため，無条件で返信先のレスを抽出する．

3.6 メインとの暗示的なアンカーを持つサブの人のレスの抽出

本節では，暗示的なアンカーの意味と，メインと暗示的なアンカーを持つサブのレスを抽出する方法について説明する．暗示的なアンカーとは 3.5 節で説明した明示的なアンカーとは違い，「>>」を使用していないにもかかわらず，返信を行っていることとなっている．話の流れで上記の記号を使用していない場合があることや，メインが使用し忘れることなどがあり，それを抽出するために行う．

これは，サブのレスに含まれる名詞がメインのレスに 1 つでも含まれていた場合に暗示的なアンカーとして抽出する．また，未知語と判断された 3 文字以上のカタカナにも名詞が多くあったため，名詞として加えることにした．ただし，いずれかのメインのレスの 10 分前から 15 分後までの直近のサブのレスのみを比較対象とする．この時刻決定について以下に示す．

3.6.1 暗示的なアンカーを抽出する時間の決定

本項では，暗示的なアンカーを抽出するときに用いた時間のしまい値を決定した方法を説明する．6 つのスレッドを読んでストーリー要約に必要だと思われるレスを決定した．

6 つのスレッドでメインのレスを中心に指定した時間内に書き込まれたサブのレスで暗示的なアンカーとなるレスと，自ら作成した回答の暗示的なアンカーのレスとの比較を行った．指定した時間は，5 分前から 300 分前までと，5 分後から 300 分後までを 5 分間隔で行い，それぞれで F 値を求めた．6 つのスレッドそれぞれで上位 50 位以内に入る範囲を求めたところ，10 分前と 15 分後が最も多く上位 50 位以内に入っていたためこの時刻と決定した．

3.7 出力：要約を html 形式で出力

本節では，提案システムの出力について説明する．3.2 節-3.6 で抽出したレスを html 形式に変換して出力する．出力例を図 2 として以下に示す．

183 レス目を見てみると他のレスよりもフォントサイズが大きい．これがこのスレッドでメインのレスとなっている．それに対しフォントサイズが小さいレスはアンカーで取られたサブのレスとなっている．また，メインのレスである 183 レス目には明示的なアンカーを示す「>>」がある．これは抽出条件に含まれるため 195 レス目が抽出されている．189 レス目は逆にメインへの明示的なアンカーを示しているので抽出されている．184・187・191・199 レス目は「追撃」や後のメインのレスにある「電話」という単語が含まれているため暗示的なアンカーとして抽出されている．

図 2: 提案システムの出力例

4. 評価実験

本章では，提案システムのメインを中心としたストーリー要約に必要なレスが抽出されているかの評価実験について説明する．

4.1 実験内容

提案システムが従来の要約システムより，掲示板のメインを中心としたストーリー要約に必要なレスが抽出されているかを確認した．

システムによって出力されたストーリー要約と原文を見比べてもらい必要なレスが取られているかを回答してもらった．

被験者に行ってもらった手順を以下に示す．

1. システムによって出力されたストーリー要約を読んでもらう．
2. 原文には出力されたレスにはチェックが付いており，出力されたレス以外に必要だと思われるレスにはチェックを付けてもらい，出力されたレスで不必要なレスからはチェックを外してもらう．

実験では 20 のスレッド（th1-th20）を使用した．要約前の 1 スレッドには約 70-100 レス含まれており，要約率は 2 割－8 割となった．また，要約を行う比較システムには展望台システム [6] を用いた．比較システムに入力したスレッドは，各レスの最後に「。」を付けることにより文とし，20 レスごとの段落とした．比較システムはストーリーの主題，副題となるキーワードをもとに各文にテキスト全体の順位と，格段落語との順位を与え，その順位の合計で出力を決定する．

被験者は情報科学を専攻する大学生・大学院生の男女 20 名で，被験者 1 人あたり提案システムで出力されたスレッドを 5 つ，比較システムで出力されたスレッドを 5 つ，計 10 スレッドずつ読んでもらった．

4.2 実験結果

本節では評価実験で得られた結果を述べる．1 スレッドあたり合計 10 人にストーリー要約に必要なレスを回答してもらった．10 人の内訳は提案システムを使用した人数が 5 人，比較システムを使用した人数が 5 人となっている．

10 人中 6 人以上がストーリー要約に必要なレスと回答したレスとの適合率と再現率を表 1 として示す．

次に，明示的なアンカーによる出力文の適合率と再現率の平均を表 2 として示す．

暗示的なアンカーによる出力文の適合率と再現率の平均を表 3 として示す．

4.3 考察

本節では，前節の実験結果に基づいて，提案システムに関する考察と，暗示的なアンカーの抽出に関する考察を行う．

4.3.1 提案システムに関する考察

表 1 の平均値から比較システムより提案システムのほうが被験者が必要と判断したレスが効率よく取られていた．すなわち，ストーリーを理解する上で提案システムが出力したストーリー要約は比較システムのものより容易であったことがわかる．提案システムの適合率が高くなった理由として考えられるのは，提案システムはメインとなる人物のレス全てと，メインに関する明示的なアンカーが特定の表現を除くものを全て抽出しているため，ストーリーを読み取る上で最小限必要なものが取られているためだと考えられる．そのため抽出されたほとんどのレスが必要と判断されたからだと考えられる．また，

表 1: システム出力文の適合率と再現率

	提案		比較	
スレッド	適合率	再現率	適合率	再現率
th1	0.854	0.897	0.636	0.718
th2	0.677	1.000	0.500	0.870
th3	0.650	1.000	0.206	0.539
th4	0.793	0.958	0.375	0.750
th5	0.811	0.956	0.789	0.911
th6	0.875	1.000	0.817	0.875
th7	0.943	0.943	0.571	0.800
th8	0.942	0.980	0.750	0.780
th9	0.781	0.980	0.629	0.765
th10	0.811	0.977	0.692	0.818
th11	0.816	0.727	0.907	0.891
th12	0.780	1.000	0.557	0.739
th13	0.720	0.900	0.345	0.500
th14	0.727	0.941	0.516	0.941
th15	0.731	0.905	0.357	0.714
th16	0.955	0.913	0.500	0.609
th17	0.744	0.914	0.558	0.829
th18	0.810	0.971	0.674	0.829
th19	0.787	0.980	0.692	0.918
th20	0.817	1.000	0.588	0.816
平均	0.801	0.947	0.583	0.781

表 2: 明示的なアンカーによる出力文の適合率と再現率

	提案		比較	
スレッド	適合率	再現率	適合率	再現率
th1	0.800	1.000	1.000	0.750
th2	0.615	1.000	1.000	0.625
th3	0.429	1.000	0.000	0.000
th4	0.750	1.000	0.923	0.667
th5	0.808	1.000	1.000	0.905
th6	0.871	1.000	0.962	0.926
th7	0.900	1.000	0.889	0.889
th8	0.938	1.000	1.000	0.867
th9	0.744	1.000	0.759	0.688
th10	0.724	1.000	0.790	0.714
th11	0.769	1.000	1.000	0.700
th12	0.815	1.000	0.875	0.636
th13	0.714	1.000	0.750	0.600
th14	0.500	1.000	1.000	1.000
th15	0.692	1.000	1.000	0.778
th16	0.900	1.000	0.800	0.444
th17	0.462	1.000	0.714	0.833
th18	0.667	1.000	1.000	0.500
th19	0.630	1.000	0.882	0.882
th20	0.750	1.000	0.500	0.333
平均	0.724	1.000	0.842	0.687

表 3: 暗示的なアンカーによる出力文の適合率と再現率

	提案		比較	
スレッド	適合率	再現率	適合率	再現率
th1	0.800	0.667	0.385	0.833
th2	0.333	1.000	0.130	1.000
th3	0.625	1.000	0.156	1.000
th4	1.000	0.750	0.121	1.000
th5	0.444	0.667	0.353	1.000
th6	0.429	1.000	0.231	1.000
th7	0.917	0.846	0.333	0.769
th8	0.833	0.909	0.435	0.909
th9	0.500	0.750	0.158	0.750
th10	0.875	0.933	0.520	0.867
th11	0.778	0.583	0.875	0.972
th12	0.467	1.000	0.194	0.857
th13	0.643	0.818	0.217	0.455
th14	0.546	0.857	0.318	1.000
th15	0.625	0.714	0.129	0.571
th16	1.000	0.750	0.381	1.000
th17	0.636	0.700	0.323	1.000
th18	0.563	0.900	0.300	0.600
th19	0.833	0.938	0.471	1.000
th20	0.125	1.000	0.000	0.000
平均	0.649	0.839	0.301	0.829

メインに関する暗示的なアンカーもある程度抽出されたため，明示的なアンカーが用いられていない返信などもある程度抽出できているからだと考えられる．再現率が高くなった理由として考えられるのは，適合率が高くなった理由と同じでストーリーに必要だと思われるレスは抽出したため，被験者が新たに加えるレスがほとんど無かったためだと考えられる．また，今回の実験の被験者の多くは今回用いている 2 ちゃんねるを読んだ経験がほとんど無かった．そのため，2 ちゃんねる特有の表現を除いている提案システムは内容を理解するのに十分であったと考えられる．

4.3.2 明示的なアンカーの抽出による考察

表 2 の平均値から，適合率は比較システム，再現率は提案システムが高いことがわかる．提案システムの適合率が比較システムより下がった理由として，出力レス数の差がある．提案システムの明示的なアンカーの出力数の平均は 16.9 レスだったのに対し，比較システムでは 10.6 レスであった．また，同じような内容のレスが連続して存在した場合，被験者はそのうち 1 つか 2 つのみを必要と判断していたため，明示的なアンカー部分の回答が減ってしまっていると考えられる．そのため出力レス数の差と回答レス数の差より適合率が下がったと考えられる．提案システムの再現率が高くなった理由としては，明示的なアンカーを全て抽出しているからと考えられる．

4.3.3 暗示的なアンカーの抽出に関する考察

表 3 の再現率から，システムの種類に関わらず必要だと思われた暗示的なアンカーも多く取られている．しかし，適合率は提案システムのほうが高い．すなわち，暗示的なアンカーの抽出は提案システムのほうが優れているということになる．適合率が高くなった理由を考える．提案システムの抽出条件はメインのレスに含まれる名詞と判断された単語とサブのレスに含まれる名詞と判断された単語が 1 つでも一致した場合のみとなっているため，メインのレスに関係があると思われるレスが多く取られたからと考えられる．再現率が高くなった理由を考える．提案システムはメインのレスの名詞が 1 つでも含まれていれば抽出するものであった．しかし，2 ちゃんねるのレスは名詞が含まれないことがある．したがって，システムの条件によって名詞が含まれる文章として読みやすいレスのみが抽出されたため，被験者たちは新たにレスを追加しなくてもストーリーとして成り立つと判断したと考えられる．

5. 結論

本論文では電子掲示板の中で，主要人物が存在するスレッドを主要人物に関するストーリー要約するシステムを提案した．本システムは，入力されたスレッドから，除外対象のレスを除去し，メインの特定とそのレスの抽出，メインのレスに関する明示的なアンカーの抽出，暗示的なアンカーを抽出し，抽出されたレスたちをまとめることによりストーリー要約の html ファイルを作成する．評価実験を行い，従来の要約システムより，電子掲示板のスレッドをメインの人物に関して要約され，スレッドの内容を理解しやすいかを確認した．

今後の課題を示す．メインへの明示的なアンカーだが，多く取りすぎてしまっているため適合率が下がっていた．したがって，明示的なアンカーに対しても時間制限などを用いて絞り込む必要がある．また，暗示的なアンカーでうまく抽出できなかったものを比較システムでは多く抽出していることがわかったので，暗示的なアンカーを抽出するための時間の変更や，今回用いた展望台システムの評価式を提案システムに導入するなどハイブリッドなシステムにする必要がある．また，現在の提案システムでは要約率を指定できないため，レス数にバラつきが生じてしまっている．したがって，要約率を指定する方法や，出力の際の優先順位などの決定をする必要がある．

参考文献

[1] 松尾豊，大澤幸生，石塚満：電子掲示板における会話からのトピック発見と要約，人工知能学会全国大会第 16 回論文集 pp206-209, 2002

[2] 竹内啓祐，浦島智，畑田稔，安宅彰隆：電子掲示板からの評判情報抽出における P/N 判断，電子情報通信学会技術報告，KBSE，知能ソフトウェア工学，vol.106，No.473，pp55-60，2007

[3] 知久大興，井上久祥，岡本敏雄：意見交換を支援する対話情報視覚化ツール”議事録”の開発〜”話題の推移”部分の抽出と構造化〜，電子情報通信学会ソサイエティ大会公演論文集，Vol2. pp37，2007

[4] 岩崎礼次郎，荒木健治：コールセンターの対話データを対象として営業日報自動生成のための重要文抽出法，人工知能学会全国大会第 19 回論文集，pp65-pp67，2005

[5] 松本裕治，山下達雄，平野義隆，松田寛，高岡一馬，浅原正幸：形態素解析システム『茶筌』，Ver.2.4.0，使用説明書，2007

[6] 相良直樹，砂山渡，谷内田正彦：サブトピックを考慮した重要文抽出による報知的要約生成，電子情報通信学会論文誌．D，情報・システム，Vol.J90, No.2, pp.427-440, 2007