E-071

ライブチャットデータを用いたコメント対象推定手法の提案

Inference of user's preferences by using live-chat data

有安　香子†　　金次 保明†
Kyoko Ariyasu　　Yasuaki Kanatsugu

1．背景・目的

デジタル放送の進化やインターネットの普及に伴い、ニッチなニーズに対するきめの細かい放送サービスの提供への期待が高まりつつある。視聴行動を一元的に捉え数値化する視聴率に変わる基準として、視聴者の満足度があげられる。現在、放送局では視聴者の感想を様々なメディアを用いて収集し、満足度の向上へと努めている。寄せられる感想は多岐に渡り、番組内容以外の外的要因を包含しているため、その分析には多くの時間と人手を要する。我々は、番組に関する感想や意見を含んだテキストデータに焦点を当て、視聴者の反応を自動抽出する手法の検討を行っている。

関連技術として、視聴者のニーズに合った番組を推薦するために、番組の視聴履歴やネット検索履歴などを元に視聴者の嗜好を推測する研究が多数行われている[1][2]。視聴者の行動履歴から視聴者の次の視聴行動を推測することは有効なアプローチであるが、これを用いて満足度などの視聴者の情動を推測することは難しい。

一方、インターネットの発達に伴い自然発生的に出現した、ライブチャットという新しい番組視聴形態が広がりを見せている。番組放送時に視聴者同士がリアルタイムでチャットを行いながら意見を交換し合うものである。これらのライブチャットデータを用いて、盛り上がりを検出し、ダイジェストを生成する研究などが行われている[3]。

本稿では、実際のライブチャットデータの特徴を分析し、その特徴に応じて視聴者の反応を抽出するために必要な情報を補完し、各コメントがどのシーンの誰に対するものかを推定した。また、提案手法の評価を行うため実データをもとに実験を行った。

2．ライブチャットデータの特徴

ライブチャットデータは、コメント生成時間、ユーザ ID、コメント内容から構成された時系列に並んだコメントの集合である。他の番組に関する感想などを含むサイトのテキストとは異なり、ライブチャットデータは、そのリアルタイム性に起因した、多くのデータ特徴を有している。

2.1 ライブチャットデータの特徴分析

予備実験として、大河ドラマ「新選組！」に対する実際のライブチャットデータ 3 サイト 4000 コメントを用いて、表現方法や特徴などの調査を行った。参加者同士のコミュニケーションに関するコメントが全体の 15%あり、番組内容に関するコメントは全体の 85%であった。その他、

- コメント対象が画面に映っている場合には、主語を省略して入力する傾向が強い。主語が明示されたコメントは 22%以下
- コメントの文体は口語体
- 登場人物や感情に対する特定の固有表現が自然発生的に決まり、これを共有し、繰り返し多用する。
- 関心のないものに言及しない
- 時間を遡った言及はしない

等の特徴が予備実験により得られた。これらの特徴は、今伝えたいことに的を絞り、最小限の入力でコメントするために生まれた特徴であると考えられる。特定の固有表現は、少ない入力で端的に気持を表現するために生まれたライブチャット独特の感情表現法である。ユーザは、これらの表現を予め辞書登録し、使用する。

例）∴(つД`)∴（悲しみの表現「泣」で辞書登録）

2.2 データ処理上の問題点

2.1 で述べた特徴を踏まえ、ライブチャットデータから、番組のどの部分に対して、どの様な意見が述べられているかを抽出するために必要なデータ処理手法の検討を行った。

コメント生成時間は、ユーザの入力速度に依存したタイムラグを含んでおり、このタイムラグが未知であるため、タイムスタンプの時刻だけでは番組のどの部分に対するコメントかを特定できない。この問題と大きくかかわるのが主語の省略の問題である。タイムラグからコメント対象が曖昧になり、省略された主語の推定が難しくなる。

また、入力速度を重要視するあまり、表記ゆれ（例：山並 or やまなみ or 山南）が通常のネットから得られるテキストのデータに比べ圧倒的に多く、口語体である・固有表現が多用されているなどの特徴と相俟って、言語処理などの一般的な手法で解析することを難しくしている。

3．コメント対象推定手法

2.2 に記した問題点を解決し、各コメントがどのシーンの誰に対するものかを推定するための手法を提案する。

コメント対象の推定に必要なデータとして、デジタル放送に付加され放送される字幕データを使い、字幕を表示するタイミング、話者、字幕テキスト本体の情報を使用する。また、番組公式サイトや EPG データから番組コンテンツの開始・終了時刻・登場人物などを事前情報として使用する。これらのデータを用いて、時系列に並ぶというチャットデータの特徴を利用した、コメント対象を推定する処理手順を下記に記す。

†NHK 放送技術研究所 システム
Broadcasting Systems, Science &Technical Research Laboratories, NHK

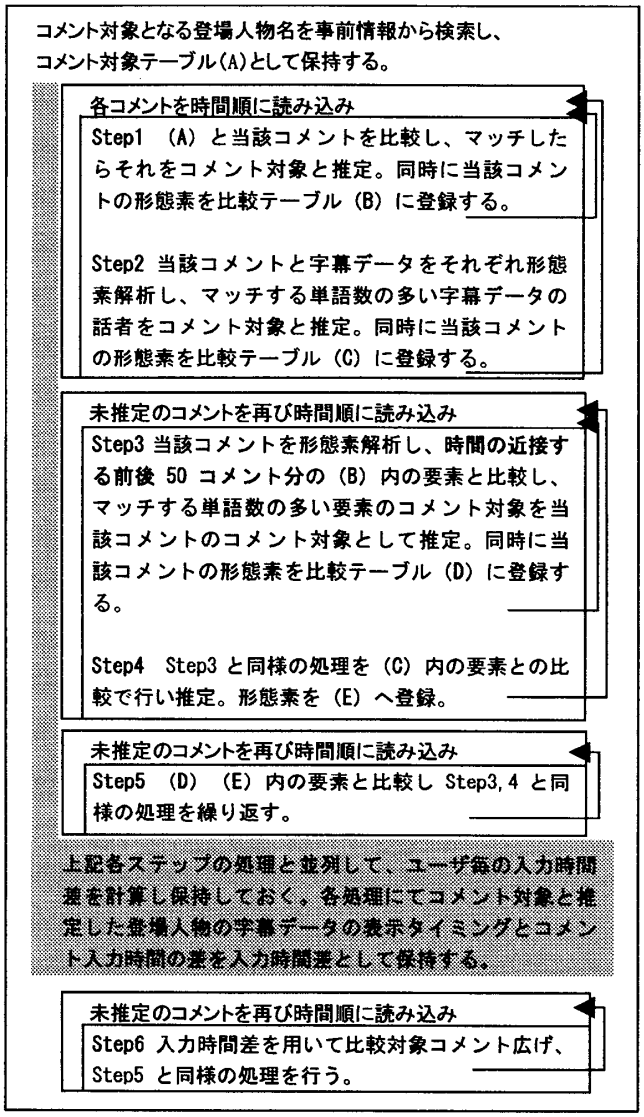

図 1: コメント対象推定処理手順

各コメントの中で、コメント対象が推定し易い、主語が明示さているコメント(Step1)、及び、台詞内の単語が明示されているコメント(Step2)に関する処理を先に行い、これらのコメントに含まれる表現と、同じ表現をもつものを、時間の近いコメントの中から探す処理を Step3 及び Step4 にて行う。その際に処理されたコメントとの類似性を用いて Step5 の処理を行う。極端に入力時間がかかるユーザのコメントを処理するため、ユーザ毎の入力時間差を計算し、Step6 において処理を行う。この様にしてコメント対象の推定を行うコメント数を増やしていく。

4. ライブチャットデータによる実験

提案手法の評価を行うため大河ドラマ「新撰組！」に関するライブチャットデータ 8000 コメントを用いて実験を行った。筆者らが[4]において述べたライブチャットデータ整形手法を用いてデータを整形して番組内容に関するコメントを抽出し、提案手法を用いて推定を行った。実際の放送映像と各コメントを照らし合わせ、推定結果の正誤判定を行った結果を表 1 に示す。尚、8000 コメントのうち番組内容に関するものは 87%、6930 コメントであった。（推定コメント数の総計に同じ）

表 1:コメント対象推定結果正解率

処理方法	推定コメント数	正解数	正解率	累計(%)
Step1.	1223	1177	0.962	17%
Step2.	1547	1366	0.883	36%
Step3.	985	546	0.55	45%
Step4.	903	381	0.42	50%
Step5.	610	281	0.46	60%
Step6.	1662	703	0.42	64%

Step1 及び Step2 の推定処理は高い正解率を示した。しかし、これら番組側の情報からの推定処理だけでは全コメントの 36%しか正しくコメント対象を推定できない。そこで、コメントが時系列に並んでいるという特徴を用いて、他コメント内の単語との一致をキーに推定処理(Step3-Step6)を行うことにより、64%まで推定率を上げることができた。

推定結果が一致しなかったデータの実際のコメント対象は大きく 3 つに分けられる。

1) 物語のキーとなる、「虎徹（日本刀）」「神輿」などの物質名に対するコメント
2) 視聴率や他の大河ドラマなどテレビ番組全体に対するコメント
3) 役者の動作や表情など演技部分に関するコメント

1)及び 2) については、図 1 のテーブル (A) の代わりに、各コメント形態素の名詞で構成することで誤推定を軽減できると考えられる。

また、上記「虎徹」を持っているのは近藤勇だが、土方が刀について台詞内で言及したため、コメント対象として土方がタグ付けされ、虎徹に関するコメントは全て不正解となっていた。この様な結果は、本稿目的であるコメント対象推定としては不正解だが、同じ場面においてコメント内容が類似したユーザ同士に、同じタグ付け処理が行われたと捉えると、コメントの類似性を基にした、ユーザクラスタリングなどに本推定手法を用いることが可能である。

6. まとめ

視聴者の番組に関する直接的な感想や意見を表しているライブチャットデータを使って、視聴者の番組に対する感想を推定する手法を提案した。実際のライブチャットデータの特徴を分析し、タイムラグや主語省略などの課題を解決し、明示されていないコメント対象を推定する手法を提案した。実際のライブチャットデータを用いて実験を行い、手法の有効性を検証した。今後はこれらの結果を用いた視聴者クラスタリング手法の提案を行う予定である。

参考文献

[1] Realization of Personalized Presentation for Digital Contents Based on Browsing History, S. Fukumura, PACRIM2003, pp. 605-608, Aug. 2003.

[2] TV Content Recommender System, Srinivas Gutta, AAAi2000, pp. 1121 - 1122, Oct. 2000

[3] Personal TV Viewing by Using Live Chat as Metadata, H. miyamori, WWW2005, pp. 948-949, May 2005

[4] ファンサイトのチャットデータを用いた番組メタデータ自動生成, 有安, FIT2006