D-035

番組に対する視聴者入力情報からの時系列キーワード抽出の改善に関する検討

On an Improved Method Extracting Time-based Keywords from Viewers' Input Information of TV Program

大黒 泰平† 加藤 友規‡ 土居 清之‡ 亀山 渉†
Taihei DAIKOKU† Tomonori KATO‡ Kiyoyuki DOI‡ Wataru KAMEYAMA†

1. はじめに

我々は、視聴者が番組コンテンツのどの部分でどのような感想を持ったのかについて、チャットの発言における有意なキーワードを定義し、時系列を伴ったデータと共に抽出することを試みてきた[1][2]。本稿では、視聴者の発言の意味的内容に加え、発言そのものの形式的な側面から視聴者の感想を抽出することについて検討を行ったため、報告する。

2. 提案方式

2-1. 発言の分析手法について

Web 上で行われるチャットでの発言内容を用いて視聴者の番組に対する感想情報を測定する際、着目すべき情報として、以下の 3 点が挙げられるものと考える。

- 発言データそのもの。形式的な情報
- 発言内容の意味。質的、意味的な情報
- ①と②の中間に位置する情報

上記を踏まえ、本稿では、視聴者からの発言内容を図 1 に示すように 3 層に階層化して捉える。

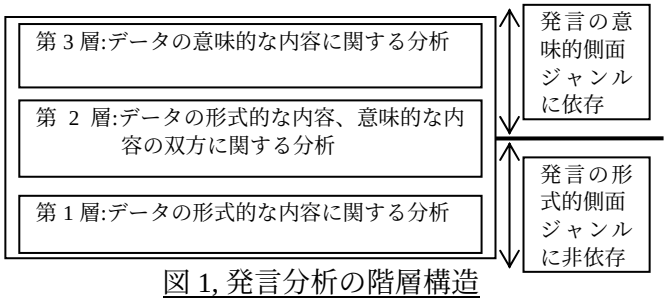

図 1, 発言分析の階層構造

2-2. 発言の分析項目について

具体的分析項目として、以下のようなものを使用することを考えている。

第 1 層:
① 発言の頻度
② 発言に費やされた文字数
③ 文字数あたりの全発言中における割合
④ 各発言間のインターバル
⑤ 発言中の漢字・ひらがな・カタカナの混在率

第 2 層:
⑥ 特定の記号(!、?、顔文字)の発生頻度
⑦ 特定の文字列(面白い etc…)の発生頻度

第 3 層:
⑧ 番組特有の文字列(固有名詞、キーワード)の発生頻度

感想情報の抽出手法に関しては、下層から上層になるに従い、番組コンテンツのジャンルに依存したものとなる可能性があることを報告した[1]。コンテンツのジャンルに依存しない感想情報を測定するためには、下層の情報分析手法が重要なものになってくるものと考える。

3. 発言内容の分析結果について

以上のことを確かめるため、男性被験者 4 名による視聴者の発言内容の分析予備実験を行った。分析結果を図 2～図 6 に示す。番組コンテンツは野球番組を用いた。

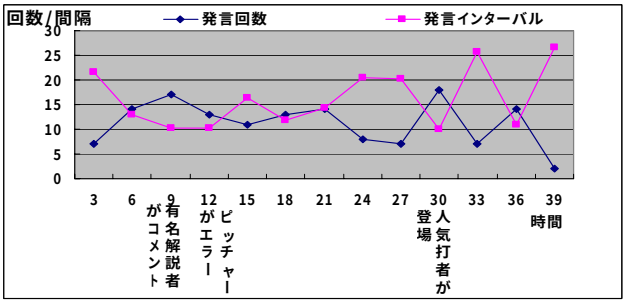

図 2, 発言頻度と発言インターバルの関係

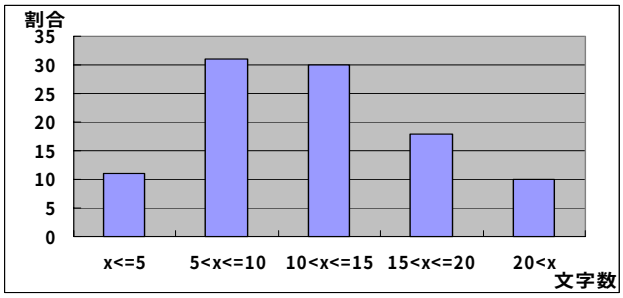

図 3, 発言文字数の割合

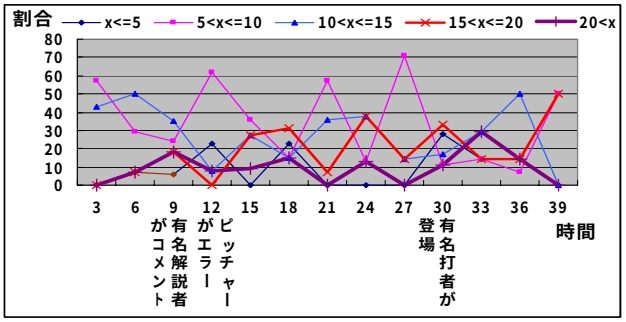

図 4, 発言文字数の推移

† 早稲田大学大学院国際情報通信研究科, GITS Waseda University

‡ 日本テレビ放送網(株), Nippon Television Network Co., Ltd

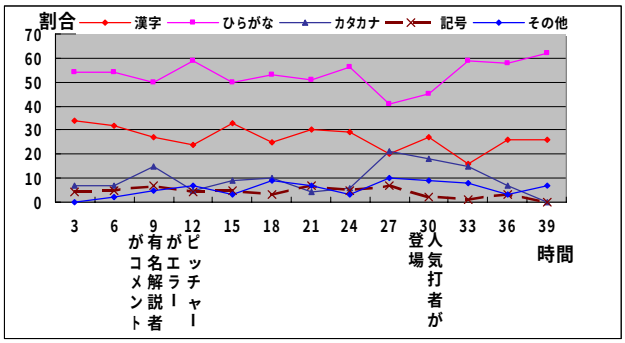

図 5, 各文字種の混在率

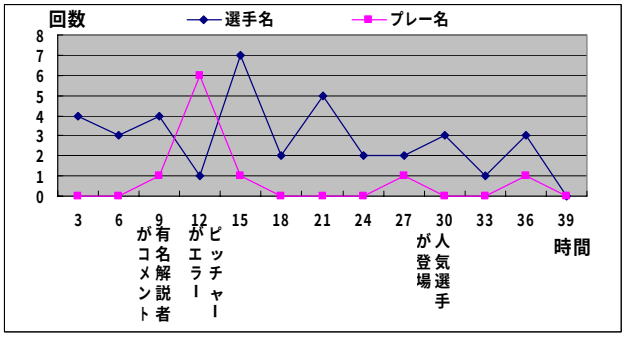

図 6, 番組固有の文字列の発生傾向

4. 結果の考察

発言内容の分析の結果、以下のような考察を行うことができる。

(ア) 発言頻度、発言インターバルについて

図 2 より、時間によってばらつきが見られることが特徴として挙げられる。当然ながら、両者には負の相関が見られる。また、視聴者が番組に対して興味を持っていると考えられる場面と発言頻度が高く、発言インターバルが短い部分がほぼ一致している。

このことから、発言頻度が高く、かつ発言インターバルが短い場面においては、視聴者が番組コンテンツに対して興味を持っている可能性が高いということができる。

(イ) 発言文字数について

図 3、4 より、発言は、約 80%が 5～20 文字のもので占められている。また、5 文字以下の発言は、記号を用いて感情を表していない限り正常な感想を表現しているとは考え難い。

このことから、5 文字以下の発言で記号以外の文字列が多く含まれている場合、その発言は意味がほとんど無いものである可能性が高い。また、5 文字以下の発言であっても記号等を用いて感情を表現しようとしている場合、その発言には有意な感想が包含されている可能性がある、などと考えることができる。

(ウ) 各文字種の混在率について

図 5 より、カタカナの発生頻度は他の文字種に比べて少ない。しかし、カタカナはその単語のみで 1 つの意味ある単語を表現しているケースが多いと考えられる。よって、カタカナは発言中の混在率は低いが、重要な意味を包含し得るワードであると考えられる。

(エ) 特定の文字列、記号の発生頻度について

チャットのログを参照したところ、視聴者の感想を表すような文字列はほとんど現れていなかった。また、図 5 より特定の記号は数は少ないながらも番組中に持続的に発言されていた。

このことから、特定の文字列に関しては、第 2 層の中でも番組コンテンツのジャンルに依存するもの、特定の記号に関しては依存しないものに属するということが考えられる。

(オ) 番組に特有の文字列の発生頻度について

番組全体を通して、選手名やプレー名と言った、番組特有の語句が多く発言されていた。

このことから、スポーツ番組においては、第 3 層にあたる情報は選手名やプレー名であるなどと考えることができる。

以上のことから、視聴者の発言内容を形式的な側面と意味的な側面に分けて考えつつ、視聴者の番組コンテンツに対する感想を抽出するための手がかりを得ることができたと考える。また、スポーツ番組において、ジャンルに依存する情報・特有な情報どのようなものであるのかについて見通しを得ることができた。

これまで我々が試みてきた、発言の意味的側面に則った分析手法と本分析手法とを組み合わせることにより、より確度の高い視聴者の番組コンテンツに対する感想情報を測定できるようになるものと考える。

5. まとめと今後の課題

本稿では、番組コンテンツに対する視聴者からの入力情報を時系列に沿って抽出する手法において、発言内容の形式的な側面・意味的な側面を分けて感想情報を抽出することを試みた。

今後の課題として、以下の 4 点が挙げられる。

- ✓ 発言の形式的側面に着目した、視聴者の番組に対する感想を反映した情報測定手法の更なる検討
- ✓ 様々なジャンルにおける同様の調査の実施
- ✓ 発言の形式的側面、意味的側面の双方の分析結果の統合手法の検討
- ✓ 発言から抽出された視聴者の感想情報を再利用可能なメタデータの形に構築していくための手法の検討

参考文献

[1]大黒泰平, 加藤友規, 土居清之, 亀山渉, “ 番組コンテンツにおけるユーザ入力情報からの時系列キーワード抽出に関する一考察 ”, 情報処理学会第 65 回全国大会 4U-6 (2004 年).

[2] 大黒泰平, 加藤友規, 土居清之, 亀山渉, “ 番組に対する視聴者入力情報からの時系列キーワード抽出の効率化に関する検討 ”, 映像情報メディア学会 年次大会 22-2 (2004 年).