3K2-NFC-3-5

TV コンテンツに対する実況コメントの収集と自己組織化手法の提案

Collection and Self-Organization of Live Comments for TV Contents

坂口 琢哉[*1]
Sakaguchi Takuya

[*1] 安田女子短期大学
Yasuda Women's College

In this study, we proposed a self-organization method for live comments about TV programs which users watching them posted simultaneously. Those comments were collected at BBS and dealt them with morphological analysis to calculate a similarity value between each comment, which was formulated with their corresponded words rate and posted time lag. We constructed a model to apply to practical data of a stream of live comments about a baseball game and obtained some essential comments including an appropriate subject for each relative scene, which is the result implying that the model is available as a basic method for summarization systems.

1. はじめに

Web が新しいメディアとして普及して以来、既存メディアとの様々なコラボレーションが実現している。ユーザが TV コンテンツなどを視聴し、その場で感じたことや考えたことをリアルタイムに Web 上でコメントしていく文化もそのひとつであり、実際に BBS や Twitter など様々なサービスで、こうした書き込みを確認することができる。

こうしたユーザによる「実況コメント」を有益なデータと見なし、動画要約や盛り上がり推定などに応用する研究が近年展開されつつあり、具体的には上原らによるもの[上原 2004]や宮森らによるもの[宮森 2005]、小林らによるもの[小林 2011]などが挙げられる。また、筆者らは主にスポーツ中継番組の各場面における代表的なコメントを自動抽出するため、実況コメントの自己組織化モデルについて提案してきた[坂口 2011]。このモデルには 3-Gram 法による類似度の計算手法が用いられていたが、単なる絶叫などあまり有益でないコメントが大量に抽出されるという問題があった。本研究ではこれに対し、各コメントを形態素解析した上で、単語の重複率に基づいて類似度を計算する新しいモデルを提案する。

実況コメントは本来、コンテンツ視聴と同時並行的に書き込みが行われるため、短絡的で意味の無い内容のものが多い。形態素解析を導入することで、そうした不完全なコメントをある程度排除し、各場面をより正確に説明した実況コメントの抽出が期待できる。

2. 提案手法

2.1 実況コメントの収集と形態素解析

本研究では、日本の大手掲示板サイト「2 ちゃんねる掲示板」の「実況板」と呼ばれる BBS 群に書き込まれた実況コメントを対象とし、データを収集した。実況板には，放映中のコンテンツに対応した書き込みの場が「スレッド」と呼ばれる形で用意されている。各スレッドにはユーザ識別 ID やコメント入力時刻などのメタ情報とコメント本文がシリアルに表示されており、1 つのスレッドには不特定多数のユーザによるコメントが最大 1000 個記録される。1 つのコンテンツに対して複数のスレッドが消費されることも多く、本研究では同一番組に対するこれら全てのスレッドからデータを収集した.

連絡先：坂口琢哉，〒731-0153 広島県広島市安佐南区安東 6-13-1，sakataku@sfc.keio.ac.jp

2.2 類似度の定式化

収集した実況コメントの内容や入力時刻に基づき、コメント間の類似度の計算を行った。ここでは、任意のコメント M および N に対し、入力時刻が互いに近く、コメントに含まれる単語が両者の間で重複しているほど類似度が高いものとし、次式により定式化した。

$$s_{MN} = \frac{1}{1+\tau(t_M - t_N)^2} \cdot \frac{c_{MN}}{c_{MM} + c_{NN} - c_{MN}} \quad (1)$$

ただし、s はコメント間の類似度、t はコメントの入力時刻、c は 2 つのコメントに共通する単語の個数を表している。c_{MM} および c_{NN} は、単純にコメント M およびコメント N に含まれる単語数に置き換えられる。また、τ は変数のオーダーを調整するパラメタである。各コメントに含まれる単語の抽出には、形態素解析エンジン「MeCab」[MeCab 2012]を用いた。

2.3 類似度に基づいたコメントの自己組織化

前節で計算したコメント間の類似度を、実況コメントの自己組織化モデルに適用した。本モデルでは、各コメントに対して「活性値」を定義し、この値をコメント間でやりとりしていくことで、次第に特定少数のコメントのみが高い活性値を獲得するようになる[坂口 2011]。具体的には任意のコメント M および N の活性値 a は、自己組織化のプロセスに伴いそれぞれ以下の式に従って変化していく。ただし、d は変化の割合を調整するパラメタである。

$$\begin{cases} \Delta a_M = +d a_N s_{MN} \\ \Delta a_N = -d a_N s_{MN} \end{cases} \quad (\text{ただし、} a_M > a_N) \quad (2)$$

上記の式は、活性値の総量が変わらないこと、活性値が特定のコメントに集約していくこと、およびコメント同士の類似度が高いほど活性値の移動量が大きく、コメントの集約が顕著に進行することを示している。このプロセスを全てのコメント間で繰り返し行うことにより、最初は均等だった活性値が徐々に特定のコメントへ集約されていき、自己組織化が実現する。

2.4 モデルの実装

提案手法を実装したモデルの流れを、以下に示す。

・step1: 対象となるスレッドからメタ情報と実況コメント文を獲得
・step2: MeCab により各コメントを形態素解析
・step3: 任意のコメント M に対し、活性値 aM の値を α で初期化
・step4: 任意のコメント M、N について類似度 sMN を計算
・step5: sMN の大きさに応じて、活性値 aM および aN の値を変化
・step6: step5 を一定回数(k 回)繰り返す
・step7: 各コメントの書込時刻と活性値をプロット

実装の際、類似度が明らかに低い値となるコメントの組合せについては擬似的に s=0 とし、これらについては step4 および step5 を省略した。具体的には、2 つのコメントの入力時間差が 60 秒を超える場合、処理の対象外とした。

なお、各種パラメタの値はそれぞれ τ =0.01、d=0.1、活性値の初期値 α =1、繰り返し回数 k=100 とした。

3. 実験結果と考察

モデルの検証を行うため、実データによる実験を行った。

ここでは、2012 年 4 月 17 日のプロ野球公式戦(広島×横浜 DeNA)について、地上波による TV 中継があった 18:40 から 20:54 までの時間帯において、特に広島を応援するスレッドに投稿された実況コメントを対象とし、これらを手動で収集した後、モデルによる自己組織化を行った。該当するスレッド数は 6 個、総コメント数は 4801 個であった。

図 1 は、横軸に経過時刻(秒)を、縦軸に各コメントの活性値をとったグラフである。自己組織化により、活性値が特定のコメントに集約していることが分かる。また、表 1 は活性値の高い上位 10 項目のコメント内容を示したものである。従来のモデルによる結果[坂口 2011]と比べ、人名などを含んだコメントが高い活性値を獲得しており、各場面の主体が理解できる出力結果と言える。これは、MeCab による形態素解析が人名を比較的正しく出力できたことに対し、「ｷﾀ――(ﾟ∀ﾟ)――!!!!!」に代表されるような短絡的なコメントは不完全に分解され、類似度が低く計算されたことが原因として考えられる。

表 1: 活性値の高いコメントと試合経過

t	(試合経過) コメント内容	a
410	(白濱がバント失敗) 大竹＞＞＞＞＞＞＞＞＞＞∞＞白濱	16.2
1073	(白濱が盗塁刺殺) 白濱の肩いけるやん！！！	11.6
4360	(梵が HR で得点) 梵ズきたあああああああああああw	12.4
4377	大切なのは引っ張りだよ	14.9
5665	(白濱が再度バント失敗) いちおつ 白濱は何ができるんや？	〃 17.6
5708	知ってたよ	12.8
5744	(代打前田登場) まぁ呉のファンにも前田見せておいたほうがいいんだろう	19.1
6839	(9 回表、抑え投手サファテ登場) サファテはバチーンの AA ないのね	11.9
7125	(サファテが奪三振) ｷﾀ――ヽ(≧∀≦)ﾉｽﾊﾟｯと三振毎度ありｯ！(≧∀≦)ﾉ――!!!!!	19.5
7483	(試合終了、広島が 3-0 で勝利) 勝ったあああああああああああああああああああああ	13.1

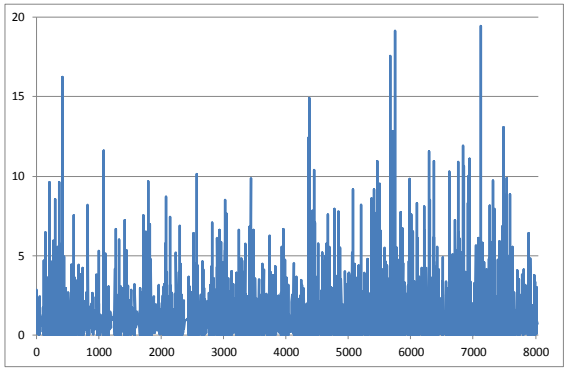

図 1: 各時刻におけるコメントの活性値

4. おわりに

本研究では、TV コンテンツに対する実況コメントとして、2 ちゃんねる掲示板の書き込みに言及し、これを収集した上で形態素解析し、コメント間の類似度を計算する手法を提案した。また、この類似度の値に基づいてコメント同士が自己組織化を行い、最終的に各場面を説明する代表的なコメントを出力するモデルを構築した。プロ野球の公式戦中継に対する実況コメントにモデルを適用した結果、各場面の主体などを含む有益なコメントが活性値を上昇しやすい傾向にあることが示され、提案手法の有効性が示された。

今後はモデルを定量的、多角的に評価し、その有効性を更に実証していくことが必要と思われる。スポーツ中継番組だけでなく、ドラマやニュースなど様々なジャンルにも適用し、モデルの汎用性についても検証が必要である。一方で、本モデルに基づいた実用的な動画要約システムを構築するとともに、その先の可能性として、マルチモーダルなシソーラスの自動構築などへの応用も検討したい。

参考文献

[上原 2004] 上原宏, 吉田健一: インターネット上の対話文に基づくドラマ番組の構造化－注目状態グラフによる視聴者による視聴者コミュニティの嗜好パターン認識, 信学技報パターン認識・メディア理解, Vol.104, No.369, PRMU2004-87, pp.25-30, 2004.

[宮森 2005] 宮森恒, 中村聡史, 田中克己: 番組実況チャットを利用したテレビ番組のメタデータ自動抽出方式, 情報処理学会論文誌, データベース, Vol.46, No.SIG_18(TOD 28), pp.59-71, 2005.

[小林 2011] 小林尊志, 野田雅文, 出口大輔, 高橋友和, 井手一郎, 村瀬洋: Twitter の実況書き込みを利用したスポーツ映像の要約, 電子情報通信学会技術研究報告, Vol.110, No.457, pp.165-169, 2011.

[坂口 2011] 坂口琢哉: 電子掲示板における TV 番組実況コメントの自己組織化と動画要約への応用, 情報処理学会研究報告「数理モデル化と問題解決」, Vol.2011-MPS-86, No.24, pp.1-2, 2011.

[MeCab 2012] MeCab: Yet Another Part-of-Speech and Morphological Analyzer, http://mecab.sourceforge.net/, 2012.