Twitter 及び電子掲示板における実況の自動検出と分析

Automatic detection and Analysis of JIKKYOU in Twitter and BBS

石原　浩[*1]，谷口　祐治[*2]
Hiroshi ISHIHARA[*1]，Yuji TANIGUCHI[*2]
*1 国立大学法人琉球大学工学部情報工学科
*1 Department of Information Engineering，University of the Ryukyus
*2 国立大学法人琉球大学総合情報処理センター
*2 Computing and Networking Center，University of the Ryukyus
*1 Email: iaia@osn.u-ryukyu.ac.jp, *2 Email: taniguchi@cc.u-ryukyu.ac.jp

あらまし：実況とは番組に対する反応をインターネット上のコミュニティに書き込むことである．既に実況を用いた研究は行われているが，「リアルタイム以外での実況」はデータとして用いられていない．リアルタイム以外での実況を検出するための手法を検討し実験を行い，またそれらの実況の有用性について評価する．
キーワード：Twitter，２ちゃんねる，実況，テキスト解析

1. はじめに

「実況」と呼ばれる行為が存在する．テレビやラジオのスポーツ中継などでアナウンサーが行う状況説明を実況と呼ぶが，そこから転じて電子掲示板などのインターネット上のコミュニティに，アナウンサーなどが行うようにある状況についての説明を書き込むことも「実況」と呼ばれる．

実況が盛んに行われるコミュニティとして「2 ちゃんねる」(以下 2ch)や，最近では「Twitter」が挙げられる．利用者はコミュニティに集まり実況をすることによって，ネット上の実況者同士で感情の共有や一体感を感じることが出来る．

実況は，番組に対する視聴者の反応であり，これを活かそうという研究がいくつか存在する．宮森垣らは番組の配信側から提供されるハイライトを集めたダイジェストなどでは製作者の意図を反映しているのみであり，視聴者視点が取り入れることが出来ていないことに着眼し，2ch の実況を用いた視聴者視点を取り入れた動画要約を生成している．(1)

坂口琢哉はあまり有益でない実況書き込みを排除する目的で単語の重複率に基づいて類似度を計算するモデルを提案した．(2)

リアルタイムだけでなく録画などで視聴する場合にも実況行為を行うこともある．近年，テレビ番組の放送内容を放送局独自のサイト上や「ニコニコ動画」などの動画共有サイトなどでも公開することが増えている．これは時間的，地理的理由からリアルタイムでの視聴が出来なかった人にも見てもらうため，また動画共有サイト独自の機能を含めて楽しみながら視聴したい人などが増えているからである．このようなリアルタイム以外での視聴においても実況が行われる．

本研究では，そのような「リアルタイム以外での実況」も有益なデータとみなし，この実況をデータとして取り入れるための自動検出システムを作成する．そのために 2ch 及び Twitter での実況を用いてテキスト解析し，リアルタイム以外での実況を検出出来るような書き込みを選出する分析システムを作成する．
また「リアルタイム以外での実況」のデータの有用性について検討する．

2. 実況について

番組中，それぞれの場面に関して，登場人物のセリフや感想，疑問や解説，感情や実況特有の書き込みなど様々な書き込みが行われる．

実況には勢いなどと呼ばれる書き込み数が急増する盛り上がりが存在する．これは番組内容の重要なシーンや印象的なシーンに見られる現象である．そのような書き込み数が急増するシーンを番組におけるハイライトとする．ハイライトでは多くのユーザが似たような書き込みを行うので，リアルタイム以外で動画視聴を行う場合でも同様の実況が行われると予想される．ハイライトを象徴するような書き込みを選出し，リアルタイム以外での実況の検出に用いる．以下にそれぞれのメディアにおける実況について解説する．

2.1. 2ch における実況

2ch は日本最大の電子掲示板サイトである．図 2 のように階層構造になっている．本実験で用いる実況は，実況 ch というカテゴリの，各放送局ごとに分けられた板に存在する，番組についてのスレッドに書き込まれるレス(レスポンスの略)を用いる．

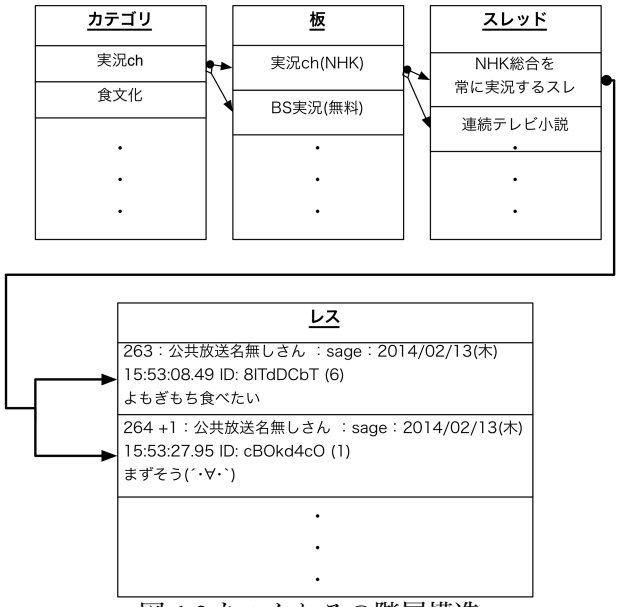

図 1:2 ちゃんねるの階層構造

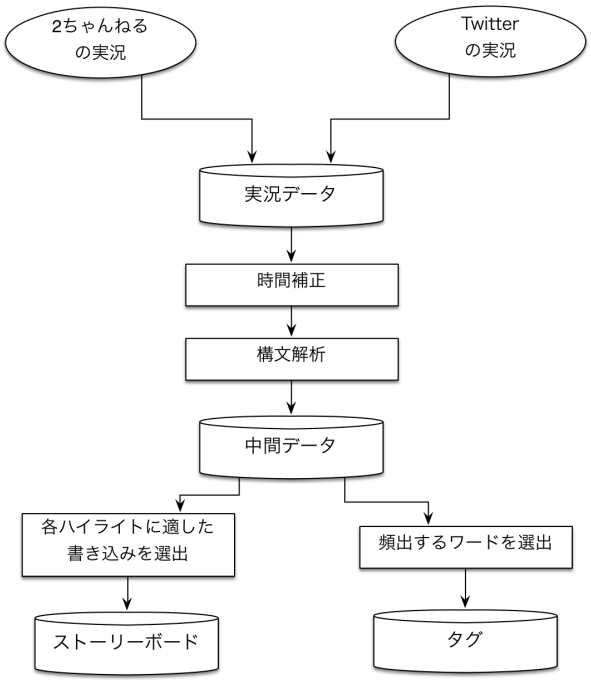

図 2:分析システム

2ch での実況は，まず番組が始まる前にその番組用のスレッドがユーザによって作られる．番組が始まると「ｷﾀｰ(ﾟ∀ﾟ)ｰ!」や「◯◯はじ また」など番組始まりを意味する文章 や顔文字が書き込まれる．

2.2. Twitter における実況

Twitter は 140 文字以内の「ツイート」と呼ばれる短文を投稿できるサービスである．

特定のトピックに関する投稿を検索しやすいように，トピックのキーワードの前に#をツイートにつけて投稿するハッシュタグが存在する．実況に用いられるハッシュタグには放送局や番組名，番組の略称や愛称など様々なものが存在する．

ツイートを手動でまとめることでひとつの記事のようにすることができる Togetter というサービスがある．Twitter での実況も Togetter でまとめられることもあるが，手動であるために非常に手間がかかり，またまとめた者の意図が含まれる．

また Twitter では録画などリアルタイム以外での実況が比較的容易である．

3. 技術概要

3.1. 分析システム

分析について大まかな流れを図 2 に示す．

自動検出は独自に作成したものを用いる．これは 2ch と Twitter 両方から実況を検出するものである．実況の書き込みから，本文及び投稿時間をデータとして検出し分析に用いる．

文章が長くなればなるほど投稿時間は遅くなる．出来る限りその場面の時に書き込み時間を近づけたい．そこで本文の文字数によって投稿時間に巻き戻しの補正をかける．

時間ごとの書き込み数はハイライトの検出として用いる．ハイライトは全書き込み数の平均以上であればハイライトとした．次に不要な文や言葉を削除する．主にハッシュタグやリンクである．ここまでの書き込みデータを中間データとする．

ハイライト時の書き込みをそれぞれ形態素解析にかける．名詞や動詞が主である単語表を作り出す．

単語ごとに重複した数だけ得点をつける．次にそれぞれの書き込みごとにその書き込みに含まれる単語の得点を全て足し，適応度とする．適応度が一番高い書き込みをその場面を最も説明している書き込みとして選出する．選出された書き込みはストーリーボードに時間とともに書き足されていく．

実況には独特の書き込みが多く見られる．AA(アスキーアート) や顔文字，「www」や「ワロタ」などだが，これらは，実況として有益なデータであるがその後の自動検出において非常に扱いづらいので極力排除する．

3.2. 自動検出

自動検出におけるおおまかな流れを以下に示す．

Step1. 番組が放送開始時から 2ch と Twitter の実況を時間と本文をデータとして検出する．

a) 2ch は対象の番組スレッド，もしくは放送局スレッドから検出する．

b) Twitter では，予め放送局や番組のハッシュタグを選び出し，そのハッシュタグを検索にかけ検出する．

Step2. 番組が放送終了時に検出を終了する．

Step3. 検出データを 2ch と Twitter で分けて保存する．

Twitter においては RT(ReTweet．他のユーザの投稿を自分のアカウントで再投稿すること) は実況に値するか判断が難しく，また全く同じ実況書き込み増えてしまうことになるため除外する．

上記した自動検出及び分析システムを用いて作成されたストーリーボードを用いて，リアルタイム以外での実況を検出する．

3.3. リアルタイム以外での実況の自動検出

Twitter のユーザを一人選びそのツイートを全て保存する．各ツイートとストーリーボードに記述された各書き込みを形態素解析にかけ，両者の類似度を計算する．類似度が高いものがある場合に，その書き込みに対してその番組の経過時間をつける．その経過時間から視聴開始時間と視聴終了時間を逆算する．その間の書き込みを実況とみなし，データとして検出する．

4. おわりに

実況についての研究で，リアルタイム以外での実況は今までデータとして用いられ無かった．本稿で述べたようにリアルタイム以外での実況の有用性を本実験で示せたので実況についての研究に一部寄与出来た．

5. 参考文献

参考文献

(1) 宮森垣，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検出方 式」，『情報処理学会論文誌』，vol.46，pp.59-71，(2005-12-15)，一般社団法人情報処理学会．

(2) 坂口 琢哉，「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」， The 26th Annual Conference of the Japanese Society for Artificial Intelligence，(2012).