E-019

Twitter と感性情報を用いた野球中継速報文の生成
Summarization of Sports Broadcast Using Twitter and Sensibility Knowledge

才田 涼† 森田 和宏† 泓田 正雄†
Ryo Saita† Kazuhiro Morita† Masao Fuketa†

1. はじめに

多くのマスメディアではスポーツの試合中に起こったイベントを速報として配信しているが，これらは人手で作成されている．これらを自動で生成することができればコストの削減や，より早い配信が期待できる．野球の場合，1 イニング毎や試合ごとにマスメディアより発表される試合速報からダイジェストを生成する研究が村上ら[1]や岩永ら[2]によっておこなわれているが，試合の流れや結果が正確に把握できる一方で，情報ソースとなる試合速報がイニング毎や試合終了時といったある一定の区切りごとに発表されるために速報性を損なっている．そこで本研究では，情報ソースとして速報性に長けたマイクロブログの中でも投稿数が多い Twitter を使用し，リアルタイムにイベントの発生を検知することでこの問題を改善する．

また，Twitter の特徴として，ハッシュタグと呼ばれるデータタグを用いることで同一の話題についてのツイートを一括して閲覧することが可能という点が挙げられる．ある特定のハッシュタグに関連するイベントが発生したとき，そのハッシュタグ内に限りツイート数が増加する現象が発生することがある．通常“バースト”とは文章中における単語の出現頻度が急増する現象のことを指すが，本研究ではある特定のハッシュタグを持つツイートの急増と定義する．本研究では，試合中のプロ野球チームのハッシュタグを持つツイート群にバーストが発生したとき，対応するチームに関するイベントが発生したと仮定し，バースト中のツイートを要約することで速報性と可読性を損なわずに速報文を生成することを目的とする．

2. 関連研究

前述の村上らの研究をはじめ，スポーツの試合ダイジェストを生成・抽出する研究はいくつかおこなわれている．久保ら[3]は Twitter でサッカーの試合中に発生したイベントを迅速かつ詳細に投稿するユーザを“良い実況者”と定義し，そのユーザの投稿をそのまま抽出することでスポーツの試合速報を生成する抽出型要約手法を提案した．しかし，久保らの手法では投稿をそのまま速報として抽出しているため，口語表現や感情的な表現などの冗長な要素を含み可読性が損なわれるおそれがあり，“良い実況者”が実況に参加していないタイミングに発生したイベントを捕捉することができない問題点がある．本研究では，村上ら[1]や岩永ら[2]と同じくテンプレートを用いた生成型要約手法を用いることで可読性を確保しながら，リアルタイムにバーストを検知することで速報性を損なわない手法を提案する．

†徳島大学大学院 先端技術科学教育部，Graduate School of Advanced Technology and Science, Tokushima University

3. 提案手法

3.1 ツイートの取得

取得したいプロ野球チームの球団別ハッシュタグを本文中に含むツイートの取得を 15 秒ごとに試みる．収集の際にリツイートは除外し，ツイート中の URL とハッシュタグを削除する．

3.2 ツイート中のバースト抽出

取得したツイートの総数が 150 件を超えるか 45 秒間ツイートが投稿されなかったとき，取得したツイート群に対して Kleinberg[4]のバースト検知アルゴリズムを適用し，バーストを検出する．Kleinberg の手法では，時系列データの各要素ごとにコスト関数を用いて計算されたバーストレベルを定めている．

3.3 主体とイベントの抽出

検出したバーストから，そのバーストを発生させた原因である主体とイベントを抽出する．総ツイート数Nのバースト D に対して形態素解析をおこない，式(2),(3)を用いて選手名の tf-idf 値を求め，最も値の高いものをそのバーストの主体と定める．次に，バースト D 中に出現する名詞 i を対象に式(1)を用いて Score を付与し，最も高い値を持つ固有名詞をイベントとする．ここで$n_{i,D}$はバースト D 全体における名詞 i の出現回数を表し，df_iは名詞 i を含むツイート数を表す．また，A は主体と名詞 i の共起回数を表す．この際に主体となった選手名は名詞 i に含まず，式(1)中の A の値やイベント候補となる固有名詞の出現回数が閾値以下だった場合は主体とイベントの関連性が低いと判断し，次節以降の処理をおこなわない．

$$Score_i = tf_{i,D} * idf_i + \log(\mathrm{A}+1) \quad (1)$$

$$tf_{i,D} = \frac{n_{i,D}}{\sum_k n_{k,D}} \quad (2)$$

$$idf_i = \log\frac{|N|}{df_i} \quad (3)$$

形態素解析に用いる辞書には NPB 公式サイト(注 1)より取得した各チームの選手名を人名，Wikipedia(注 2)や野球情報サイト「BASEBALL MONSTER(注 3)」から取得した野球用語を固有名詞として登録し，「ホームラン」に対する「本塁打」，「HR」など同義の用語は同じ語と見なす．

3.4 速報文の生成

前節で抽出したイベントに対応したテンプレートを用いて速報文を生成する．テンプレートの例を表 1 に示す．対応するイベントが同じテンプレートに関しては，Yoshinari

(注 1): http://npb.jp/
(注 2): https://ja.wikipedia.org/wiki/野球用語一覧
(注 3): http://baseballmonster.nobody.jp/

表 1:テンプレートの例

番号	テンプレート	対応するイベント例
1	<主体>が<イベント>で出塁しました	ヒット，2 塁打，バントなど
2	<主体>に<イベント>で出塁されました	
3	<主体>が<イベント>で出場します	代打，リリーフなど
4	<主体>が<イベント>で得点しました	ホームラン，スクイズ犠牲フライなど
5	<主体>が<イベント>で得点されました	
6	<主体>が<イベント>を達成しました	サイクルヒットなど

表 2:ROUGE スコア例

	久保らの手法	提案手法
試合 A	0.169	0.281
試合 B	0.181	0.333
試合 C	0.125	0.533
試合 D	0.175	0.290
試合 E	0.134	0.433
18 試合平均	0.141	0.380

表 3:生成された速報文の例

	久保らの手法	提案手法
例 1	筒香のツーランホームランキター(ﾟ∀ﾟ)!特大ホームラン！ これは日本の 4 番！！おかえりなさーーーーい！！！	筒香が 2 ランホームランで得点しました
例 2	５点ビハインド 1 アウト満塁 4 番中谷犠牲フライ！ めーーちゃええ当たり。	中谷が犠牲フライで得点しました

ら[5] の感性表現辞書をもとにバーストを構成するツイートの極性を判断する．極性 positive を持つツイートが多ければそのチームにとって好ましい内容のイベントが発生していると仮定し，表 1 の番号 1 のように当該チームにとって好意的な内容を持つテンプレートを使用する．同様に，極性 negative を持つツイートが多ければ，表 1 の番号 2 のようにそのチームにとって好意的でない意味を持つテンプレートを使用する．

4. 評価実験

4.1 実験設定

提案手法の有効性を確認するため，評価実験をおこなった．データセットとして，2017 年 5 月 23 日から同年 5 月 28 日にかけて NPB セントラル・リーグ公式戦でおこなわれた 18 試合の試合時間中に投稿され，球団を表すハッシュタグを持つツイートを使用した．本研究との比較手法として，久保ら[3]の手法を用いる．評価方法として久保らの手法で用いられている修正版 ROUGE-1 を利用した．これは決められた時間差内の正解要約に該当する単語が出現した場合のみ考慮するものであり，久保らの手法ではその差を 3 分と定めている．本研究ではその差を 1 分とし，正解データは Yahoo!スポーツナビ(注 4)のテキスト速報を用いた．

4.2 実験結果

実験結果の一部を表 2 に，出力例を表 3 に示す．18 試合全ての ROUGE-1 スコアを平均した値は 0.38 となり，いずれの試合においても久保らの手法より高いスコアを示した．また，速報性という観点についても，正解とみなす要約文を 1 分以内に出力されたものと限定した上でリアルタイムにバーストの検知を行っている点から，一定の有用性があると確認できた．一方でマイクロブログの性質上，ネットスラングや選手に対する別称が多く含まれる文に対して，主体が適切に抽出できない事例が確認できた．他にも，一つのバーストに複数の話題が混在しているケースがある．これらの問題点に対しては複数の別称を形態素解析辞書に登録し，同一の人名と見なす処理や，共起関係にある主体とイベントに対して重み付けをおこない，同一バースト中の複数の速報文を時系列順に出力することで改善を図る．

(注 4): https://baseball.yahoo.co.jp/npb/

5. おわりに

本稿では，イベントに呼応して発生するバースト現象に着目し，速報性と可読性を両立した速報文の生成手法を提案した．評価実験の結果，話題が複数存在するバーストに対するイベント抽出方法やテンプレート選択方法について改善をおこなう必要があることが確認できた．

今後の課題としては，問題点の改善やテンプレートの拡充によって ROUGE スコアのさらなる向上を図るとともに，速報文生成精度の向上やボールカウントや打球の行方，出塁状況といったより詳細な情報を迅速に出力する手法の開発が考えられる．

参考文献

[1] 村上 聡一朗,笹野 遼平,高村 大也,奥村 学, “打者成績からのイニング速報の自動生成”,言語処理学会論文誌,Vol.22,pp.338-341(2016).
[2] 岩永 朋樹,西川 仁,徳永 健伸,“テキスト速報を用いた野球ダイジェストの自動生成”,言語処理学会論文誌,Vol.22,pp.238-241(2016).
[3] 久保 光証,笹野 遼平,高村 大也,奥村 学, “良い実況者”に着目した Twitter からのスポーツ速報生成”,言語処理学会論文誌,Vol.19,pp-138-141(2013).
[4] J. Kleinberg, “Bursty and hierarchical structure in streams”, Proc.8th ACM SIGKDD, pp.91-101 (2002).
[5] Tomoko Yoshinari,El-Sayed Atlam,Kazuhiro Morita,Jun-ichi Aoe, “Automatic acquis-ition for sensibility knowledge using co-occurrence relation”, Interna-tional Journal of Computer Applications Technology and Resarch, Vol.33, pp.218-225 (2008).
[6] Chin-Yew Lin.Rouge, “A package for automatic evaluation of summaries”, In Proceedings of ACL Workshop Text Summarization Branchesa Out, pp.74-81 (2004).