一般社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
IEICE Technical Report
TL2016-8,NLC2016-8(2016-06)

マイクロブログを対象としたスポーツの試合における生成型要約

田川　裕輝†　　嶋田　和孝†

† 九州工業大学大学院 情報工学府 先端情報工学専攻
〒 820–8502 福岡県飯塚市川津 680–4
E-mail: †{y_tagawa,shimada}@pluto.ai.kyutech.ac.jp

あらまし　近年，マイクロブログを対象としたスポーツの試合における要約生成の研究が盛んに行われている．このような研究の要約生成手法の多くは抽出型要約手法によるものである．しかし，抽出型要約手法では要約文中に必要な要素と冗長な要素が混在するといった問題がある．本研究では，冗長な要素を含まない生成型の要約手法を提案する．スポーツの試合といった event 中にゴールなどの sub-event が起きるとその sub-event に関する投稿数が急増するバーストと呼ばれる状態が起こる．バースト毎に要約文を生成することで試合中の重要な場面を網羅した要約生成が可能となる．また sub-event 中には“選手 A のパス”や“選手 B のヒット”といった sub-event を構成する Sub-Event Element (SEE) が存在する．このような SEE を 1 文に繋ぎ合わせることで，冗長な要素を含まず，sub-event を詳細に説明する要約文を生成する．

キーワード　マイクロブログ，スポーツ，要約生成，バースト

Abstractive summarization of sports games from microblogs

Yuuki TAGAWA† and Kazutaka SHIMADA†

† Kyushu Institute of Technology, Graduate School of Computer Science and Systems Engineering　Kawadu 680–4, Iidukasi, Fukuoka, 820–8502 Japan
E-mail: †{y_tagawa,shimada}@pluto.ai.kyutech.ac.jp

Abstract　Recently, summary generation of sports games for microblogs is studied actively. Summary generation methods in the studies are mostly based on extractive summarization. However extractive summarization suffers from the disadvantage of existing both required elements and redundant elements in the summary. In this paper, we propose an abstractive summarization method to reduce the redundant elements. In the method, we focus on burst situations in which many users post tweets when a sub-event in a game occurs. We handle tweets in the burst situations as the inputs for the summarization. Each Sub-Event Element (SEE) contains some phrases that express actions in a game, such as“Player A made a pass to Player B”and“Player B made a shot on goal”. Our method integrates these phrases for the summarization, and then generates an abstractive summary, such as “Player B made a shot on goal from the Player A's pass”.

Key words　Microblogs，Sports，Summary generation，Burst

1. はじめに

近年，Twitter(注1) などに代表されるマイクロブログは多くの人々に利用されている．Twitter では tweet と呼ばれる 140 文字以内の短いテキストを介して様々な情報が投稿されており，その手軽さから，速報性のある情報源として注目されている．特に，ユーザの関心が高いスポーツの試合などの event に関する投稿は膨大な量が存在する．スポーツの試合は時間の経過とともに得点や選手交代など sub-event が変化するため，投稿される内容も時間の経過とともに変化するという特徴がある．このような投稿に対して要約を生成することで，膨大な量の投稿を全て閲覧するといった作業をすることなく，いつ何が起きているか，いつどのような状況かを知ることができる．マイクロブログを対象とした文書要約技術はこのような作業を解決する有効な手段の 1 つとなりうる．

Nichols ら [1] と Kubo ら [2] は，マイクロブログを対象とし

(注1)：https://twitter.com

たスポーツの試合における要約を生成している．彼らは試合中に重要な sub-event が起きると，時間当たりの tweet 数が急増するバースト [3] と呼ばれる状態に着目し，sub-event の時間的変化を捉えている．そして，バースト毎に要約を生成することで，試合中の重要な sub-event を網羅した要約を生成している．また，彼らは要約生成手法として抽出型要約手法を用いている．しかし，抽出型要約手法により生成された要約文には，要約に必要な要素と冗長な要素が混在するといった問題点が挙げられる．抽出型要約手法を用いた，得点という sub-event に対する要約例を以下に示す．

- きたあああ!目を放した隙にスアレスのパスから神の子メッシがゴール!バルセロナ先制!

この要約文には，“きたあああ!”といった感情的な表現や，“目を離した隙に”といった投稿者の状況，“神の子”といった比喩表現など冗長な要素が存在する．この問題を解決するためには冗長な要素を含まない生成型の要約手法の確立が必要である．

上述した要約例では，“スアレスのパス”，“メッシがゴール”，“バルセロナ先制”といった部分は要約文に必要な要素である．この必要な要素を繋ぎ合わせることで，以下のように必要な要素のみを簡潔にまとめた要約文が生成できる．

- スアレスのパスからメッシがゴール，バルセロナが先制．

本論文では冗長な要素を含まない要約生成を目的とした Twitter からのスポーツの試合における生成型の要約手法を提案する．スポーツの試合の sub-event 中には，sub-event を詳細に説明するために必要な要素である Sub-Event Element (SEE) が存在する．SEE を繋ぎ合わせることで sub-event を詳細に説明し，かつ冗長な要素を含まない要約生成を目指す．

2. 関 連 研 究

マイクロブログを対象とした要約生成の関連研究を以下に説明する．抽出型要約手法を用いた研究として Nichols ら [1] や Kubo ら [2]，Takamura ら [4] の研究がある．

Takamura らは Twitter を対象としたスポーツの試合における要約を生成している．ある文書ともう 1 つの別の文書との被覆度が最大になるようにいくつかの文書を選択する施設配置問題 [5] と呼ばれるモデルにより算出された被覆度の高い tweet を時系列順に提示することで sub-event の時間的変化を捉えた要約手法を提案した．被覆度を測る際には，文書間の単語の一致率に基づいた含意関係を利用している．

Nichols らと Kubo らは Twitter を対象としたスポーツの試合における要約をリアルタイムで生成している．Nichols らは，まず時間当たりの tweet 数の増減からバーストを検出する．次に，バースト中の tweet 集合から単語グラフを作成し，各 tweet に対して単語の出現頻度を用いたスコアリングを行う．そして，スコアが上位の tweet から順に単語の重複がないように，指定された数の tweet を選択し，要約として提示する手法を提案した．しかし，バースト開始直後に出現する tweet には感情的な単語を含む tweet が多く，Nichols らの手法により生成される要約文は単語の頻度に依存するため，感情的な tweet が要約文として選択されやすいといった問題点がある．

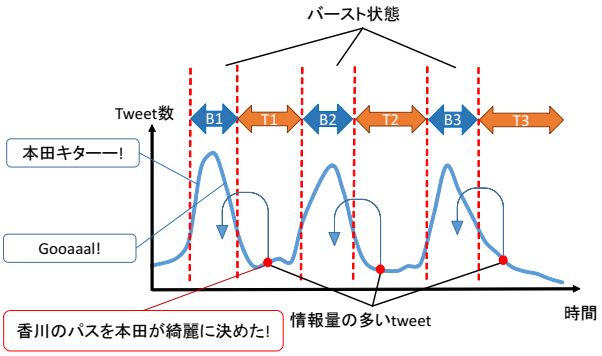

図 1　バースト状態

Kubo らは Nichols らの問題点に対して，各 tweet 毎に選手名と“ゴール”や“パス”といったスポーツ専門用語の共起回数を考慮した tweet スコアを算出し，これを説明性の指標の 1 つとした．また，選手名と専門用語の共起回数が多い説明的な tweet をバースト開始直後であっても，投稿しているユーザのスコアが高くなるようなユーザスコアを考慮することで，説明的な tweet を速報として提示する手法を提案した．

マイクロブログを対象とした文書要約では抽出型要約手法が主流となっている．抽出型要約手法では文や tweet を抽出単位とするため，ある程度の文法性が担保されるといった利点があり，口語表現，誤字脱字といった言語現象が多く見られるマイクロブログにおいて有効である．しかし，抽出した文や tweet の一部が重要な情報を含み，他の部分が冗長な要素である場合が考えられる．この場合，要約文中に必要な要素と冗長な要素が混在するといった問題がある．

生成型要約手法を用いた研究として Sharifi ら [6] の研究がある．Sharifi らは，Twitter を対象としたスポーツの試合や政治，事件などの event に対する要約を生成している．まず，event に関する tweet 集合からキーとなる単語を中心とした単語グラフを作成する．そして，グラフ中の単語ノードに対して，その単語の出現頻度などによる重み付けをして，そのキーとなる単語を含む最も重みが大きくなるパスを要約として提示している．しかし，Sharifi らは，1 つの event に対して 1 文の要約文を生成しており，sub-event の時間的変化を捉えようとした研究ではなく，問題設定が本研究と異なる．

3. 提 案 手 法

本節では 2 節で挙げた先行研究をふまえ，生成型の要約手法を提案する．バースト毎に要約を生成することで試合中の重要な sub-event を網羅した要約生成が可能となる．また，バースト中の tweet から“選手 A のパス”，“選手 B のヒット”といった sub-event を詳細に説明する Sub-Event Element (SEE) を獲得し，繋ぎ合わせることで冗長な要素を含まない要約生成手法を提案する．以降では，バーストの検出と要約生成手法について説明する．

3.1 バーストの検出

本論文では，試合中に重要な sub-event が起きると tweet 数が急増するバーストに着目する．提案手法では，まず 30 秒あ

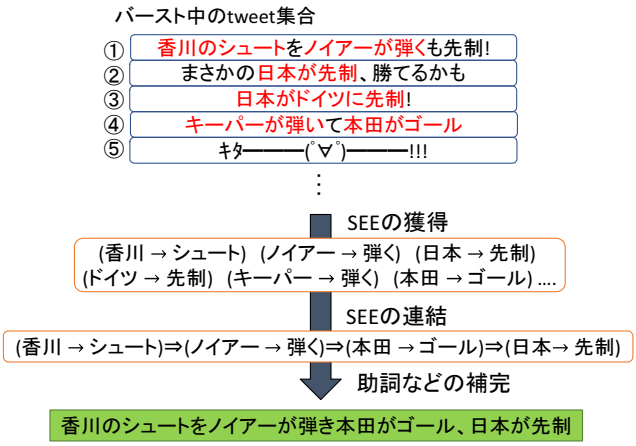

図 2　要約生成の概要図

たりの tweet 数の中央値を算出する．そして，ある 30 秒間の tweet 数が算出した中央値の 3 倍以上になっている場合，その 30 秒間をバーストとする．また，近接するバーストは同一の sub-event に言及していると考えられるため，バーストが連続する場合やバースト間の時間が 30 秒の場合は 1 つのバーストとする．

図 1 に示すように，バースト中の tweet には感情的な tweet が多く，非バースト中にも情報量の多い tweet が多く存在する．そのため，非バースト中に投稿された情報量の多い tweet をバースト中の tweet 集合へ追加することを考える．バースト中の tweet に最も多く出現する選手名と専門用語の組をその sub-event を表す代表語とする．非バースト中の tweet のうち，その代表語を含む tweet は直前の sub-event について言及した tweet であり，その tweet を直前のバーストに追加する．例えば，図 1 の B1 の代表語を含む T1 中の tweet を B1 に追加するといった処理である．

3.2　要 約 生 成

2 節で挙げた抽出型要約手法での問題点を解決するために，本論文では生成型の要約手法を提案する．図 2 は先制という sub-event での要約生成の概要図である (注2)．バースト中の tweet には“香川のシュート”や“日本が先制”といった，動作主とその動作の関係からなる Sub-Event Element (SEE) が存在する．まず，バースト中の tweet から (動作主→動作) という関係にある SEE を獲得する．そして，必要な要素のみを簡潔にまとめた要約文を生成するため，要約文に必要な SEE のみを連結する．最後に，動詞の活用形の変換や連結された SEE の単語間に適切な助詞を補完することで整合性の取れた最終的な要約文を提示する．

3.2.1　Sub-Event Element の獲得

図 2 に示すように，tweet 中の動作主とその動作という関係から (動作主→動作) といった Sub-Event Element (SEE) を獲得する．動作主と動作の間には，動作主から動作への係り受け関係が存在する．そのため，バースト中の tweet に対して係り受け解析し，SEE を獲得する．しかし，マイクロブログを含む

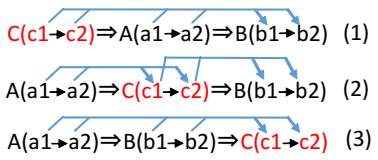

図 3　sub-event element(SEE) の連結位置

web 上の言語資源は口語表現，誤字脱字を含む文書が多く，係り受け解析精度に大きく影響する [7]．そのため，バースト中の tweet に対し，ルール (注3) や Google N-gram [8] を用いて，助詞，読点を補完することで係り受け解析の精度を上げる．助詞，読点を補完した後に，バースト中の tweet に対して係り受け解析し，サ変名詞以外の名詞，またはスポーツ専門用語以外の名詞からサ変名詞や動詞，スポーツ専門用語への係り受け関係を SEE として獲得する．この際に動詞は原形に変換する．また，a から b への係り受け関係と，b から c への係り受け関係が存在する場合，a から c に係る関係が存在するとみなす．係り受け解析器には Cabocha (注4) を用いる．

3.2.2　Sub-Event Element の連結

必要な要素のみを簡潔にまとめた要約生成のために，Sub-Event Element (SEE) を連結する．SEE の連結とは SEE を時系列順に並べることである．時系列順に並べることで，sub-event の正確な状況を説明する要約文を生成する．また，バースト中の tweet から獲得した SEE は非常に多く存在し，要約文に必要のない SEE も存在するため，必要な SEE か否かを判別し，必要なもののみを連結する必要がある．

出現頻度の最も高い SEE に対して，その他の SEE を出現頻度降順に，必要な SEE のみを時系列順に並ぶ位置に連結する．まず，連結する SEE 中の単語と既に連結された SEE 中の単語が重複する場合，その SEE は連結しない．重複しない場合，SEE が時系列順に並ぶ連結位置を決定する．次に，連結位置が決まった SEE に対して，その SEE が要約文に必要な SEE か否かを判別するためのスコアを計算する．このスコアが一定の閾値を超えた場合に時系列順に並ぶ連結位置に連結する．このような処理を繰り返すことで要約文に必要な SEE のみを時系列順で連結する．以降，SEE の時系列順での連結と，要約文に必要な SEE か否かを判別するためのスコアについて説明する．

SEE の時系列順での連結位置を決定するには SEE 同士の時間的順序関係を獲得する必要がある．図 2 に示す ① の tweet では，“シュート”といった単語の後に“弾く”という単語が出現している．また，④ の tweet では，“弾く”から“ゴール”に係り受け関係が存在する．このような単語の出現順序と係り受け関係から，“シュート”の後に“弾く”，“弾く”の後に“ゴール”という時間的順序が獲得できる．このように tweet 中での単語の出現順序と係り受け関係から SEE 同士の時間的順序関係を獲得することができる．

実際に，A(a1 → a2) ⇒ B(b1 → b2) と連結された SEE に C(c1 → c2) という SEE の時系列順での連結位置を考える．図

(注2)：→は sub-event が→の前後の単語から構成されることを表し，⇒は前後の sub-event が連結していることを表す．

(注3)：例えば，選手名と専門用語が連続する場合，“ の ”を補完する．

(注4)：http://chasen.org/～taku/software/cabocha/

3 に示すように C の連結位置は (1)，(2)，(3) の 3 通りある．時系列順に並ぶ位置では，図 3 に示す矢印のような単語の出現順序や係り受け関係が成立する組み合わせが増える．そのため，考えられる全ての連結位置に対して，ある SEE 中の単語の後に別の SEE 中の単語が出現する頻度と，ある SEE 中の単語から別の SEE 中の単語への係り受け関係の頻度の合計を計算し，その合計が最も高い位置が連結する SEE が時系列順に並ぶ連結位置とする．すなわち，求めた頻度の合計が高くなるほどその位置が SEE が連結される位置として適切であることを表している．

選手名，チーム名，スポーツ専門用語を含む SEE や出現頻度の高い SEE は要約文に必要な SEE である．一方，複数のバーストに出現する SEE は sub-event を特徴付ける SEE ではないため，要約文には必要のない SEE である．連結位置が決まった SEE が要約文に必要な SEE か否かを判別する基準として (1) 式を用いてスコア S を計算する．

$$S = \frac{maxFreq \times seeScore \times seeFreq}{wordNum \times tweetNum \times seeBurstnum} \quad (1)$$

$where$

$$seeScore = \begin{cases} 4 & if \quad \text{選手名またはチーム名と，スポーツ専門用語を含む SEE} \\ 3 & elsif \quad \text{選手名を含む SEE} \\ 2 & elsif \quad \text{専門用語を含む SEE} \\ 1 & otherwise \end{cases}$$

$maxFreq$ は時系列順に並ぶ連結位置を決定する際に算出した最も高い単語の出現順序と係り受け関係の頻度の合計である．$seeScore$ は連結する SEE に対する重みであり，$seeScore$ の値は情報量の多い SEE に対して高い値となるように経験的に決定した．$seeFreq$ はバースト中での SEE の出現頻度であり，$wordNum$ は連結後の連結された SEE の単語数である．また，$tweetNum$ はバースト中の tweet 数である．$seeBurstnum$ は直前までのバーストのうち，連結される SEE が出現するバーストの数である．スコア S は情報量が多く，時系列順に並ぶ適切な位置に連結される SEE に対して高い値となる．スコア S が一定の閾値を越えた場合，時系列順に並ぶ位置に連結する．

3.2.3 Sub-Event Element への単語の補充

Sub-Event Element (SEE) は，動作主とその動作を表す 2 単語からなる．しかし，この 2 単語だけでは，sub-event を詳細に説明する要素として情報が不足している場合がある．

例えば，以下のような SEE が獲得されたとする．

- (選手 A →決める) : 7
- (選手 A → PK) : 6

出現頻度降順に単語の重複がないように SEE を連結していくため，(選手 A →決める) という SEE が連結された場合，(選手 A → PK) という SEE は，連結されず，“ PK を決める ” という情報が要約文に含まれない問題が存在する．

そのため，(選手 A → PK →決める) のように (選手 A →決める) という SEE に“ PK ”という単語を補充することで，“ 選手 A が PK を決めた ” といった情報を要約文に含むことができる．このように，連結された SEE の先頭の SEE から順に単語の補充が必要な SEE にのみ適切な単語を補充する．

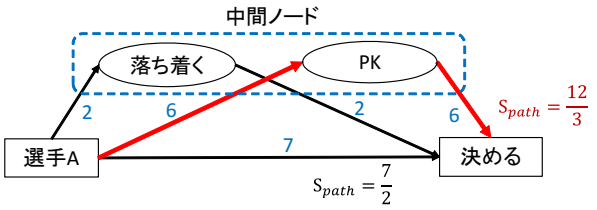

図 4 “ 選手 A ”から“ 決める ”への係り受けグラフ

例として，(選手 A →決める) といった SEE への単語の補充を考える．はじめに，“ 選手 A ”という単語から“ 決める ”という単語への係り受け関係からなるグラフを作成する．係り受けグラフを図 4 に示す．このグラフは係り元の単語から係り先の単語へ係り受け関係の頻度で重み付けされたエッジで繋いだグラフである．係り受けグラフ中の全てのパスに対して以下の (2) 式を用いてスコア S_{path} を計算し，スコア S_{path} が最も高いパスの単語を SEE に補充する．しかし，SEE は 1 つの動作主とその動作によって構成されるため，新たに動作主を表す選手名とチーム名は補充しない．

$$S_{path} = \begin{cases} 0 \ if \ \text{中間ノードに選手名またはチーム名を含む} \\ \dfrac{\text{係り受け頻度の合計}}{\text{単語数}} \quad otherwise \end{cases} \quad (2)$$

3.2.4 最終的な要約文の生成

連結された Sub-Event Element (SEE) の単語間に助詞や読点を補完することで，整合性の取れた最終的な要約文を生成する．例えば，(香川→パス) ⇒ (本田→決める) ⇒ (日本→先制) といった連結された SEE が存在する場合，“ 香川のパスから本田が決め日本が先制 ” といった読みやすい要約文を生成する．そのためには動詞の活用形の変換，単語間への適切な助詞と読点の補完の 2 つの処理が必要である．

まず，連結された SEE 中の末尾以外の動詞を連用形に変換する．例えば，“ 決める ”は“ 決め ”，“ 弾く ”は“ 弾き ”に変換する．また，“ 同点に追いつかれる ” などといった文脈からは (同点→追いつく) といった SEE が獲得される．そのため，“ 追いつく ”を“ 追いつかれる ”という元の形に変形する必要がある．よって，連結された SEE 中の動詞がバースト中の tweet で出現する回数のうち，半数が接尾動詞が語尾についた形で出現している場合，接尾動詞が語尾についた形に変形する．このような処理を連結された SEE 中の全ての動詞に対して行う．

次に，連結された SEE の単語間 (→と⇒の部分) へバースト中の tweet から頻出する助詞や読点を補完する．補完語を決定する際にはルールを用意する．ルールの例を表 1 に示す．ルールにマッチしない場合，以下の Step1〜3 を処理する．補完語が存在しない場合，次の Step に移り，補完語が決定した時点で処理を終了する．

Step1

バースト中の tweet に出現する補完部の前の語から後ろの語への係り受け関係のうち，前の語に接続する助詞で最も出現頻度の高い助詞を補完する．

表 1　助詞，読点の補完に用いたルールの例

ルール	補完語
補完部の前の語が選手名またはチーム名かつ後ろの語がスポーツ専門用語	の
末尾の SEE で 前の語が選手名またはチーム名かつ後ろの語がスポーツ専門用語	が
補完部の前の語が“クロス”，“パス”，“アシスト”，“キック”	から

Step2
バースト中の tweet に「補完部の前の語+助詞+補完部の後ろの語」という単語の並びで出現する助詞のうち，最も出現頻度の高い助詞を補完する．
Step3
バースト中の tweet に「補完部の前の語+助詞」という単語の並びで出現する助詞のうち，最も出現頻度の高い助詞を補完する．補完する助詞が存在しない場合，読点を補完する．

4. 実　　験

4.1 データセット

サッカーの 6 試合に関する tweet を用意した．6 試合に関する tweet はそれぞれの試合に関連するハッシュタグを含むものを収集した．収集したデータの統計量を表 2 に示す．

選手名とサッカー専門用語は wikipedia(注5)，サッカーニュースサイト「ゲキサカ」(注6)，日本サッカー協会の web サイト(注7)から取得した．

4.2 実 験 設 定

ある 1 試合のスコア S の閾値は，他の 5 試合を用いて決めた．

本研究と比較する手法として Kubo ら [2] の手法を用いる．Kubo らは抽出型要約手法を用いて，要約をリアルタイムで生成している．リアルタイムでの要約生成という点で本研究と手法は異なるが，提案手法が抽出型要約手法の問題点を解決できているかを比較するため，Kubo らの手法を採用した．バーストの検出は提案手法を用いて，提案手法の要約文と Kubo らの手法の要約文を評価した．

4.3 可読性の評価

抽出型要約手法では文を抽出単位としているため，ある程度の文法性は担保される．しかし，提案手法では適切な助詞などの補完に失敗している場合，読みにくい要約文が生成される．要約において，読みやすい文であることは重要であり，提案手法の要約文の読みやすさを可読性という観点から評価する．3 人の被験者で要約文の可読性を 1 文ごとに 1〜4 の 4 段階(注8)で評価した．提案手法の要約文の可読性の評価結果を表 3 に示す．

表 3 から，可読性の評価では全ての試合において 3 以上の評価を得ており，平均値では 3.59 と高く評価されていることが確認できた．この結果から，提案手法により，適切な助詞などの補完ができたといえる．

(注5)：https://ja.wikipedia.org/wiki/サッカー用語一覧
(注6)：http://web.gekisaka.jp
(注7)：http://www.jfa.jp
(注8)：1:読みにくい〜4:読みやすい

表 2　データの統計量

ハッシュタグ	対戦チーム	試合日程	tweet 数
#CL など	バルセロナ対ユヴェントス	2015/6/6	9,945
#CWC など	バルセロナ対リーベルプレート	2015/12/20	5,258
#nadeshiko, #W 杯など	日本対オランダ	2015/6/24	4,734
	日本対オーストラリア	2015/6/28	6,831
	日本対イングランド	2015/7/2	7,862
	日本対アメリカ	2015/7/6	16,507

表 3　可読性の評価結果

試合名	可読性
バルセロナ対ユヴェントス	3.57
バルセロナ対リーベルプレート	3.76
日本対オランダ	3.14
日本対オーストラリア	3.85
日本対イングランド	3.33
日本対アメリカ	3.88
平均値	3.59

表 4　被験者による比較評価と圧縮性の評価結果

試合名	被験者評価	文圧縮率 [%]
バルセロナ対ユヴェントス	3.00	58.08
バルセロナ対リーベルプレート	3.42	51.35
日本対オランダ	3.14	23.26
日本対オーストラリア	2.23	23.55
日本対イングランド	2.78	31.19
日本対アメリカ	2.38	23.51
平均値	2.83	35.16

4.4 比 較 評 価

6 人の被験者で同じバーストから生成された提案手法の要約文と Kubo らの手法の要約文を比べ，どちらが適切な要約文か評価した．評価の際には，被験者に各バースト中の tweet 集合を与えた上で，1〜4 の 4 段階(注9)で評価した．すなわち，中間値である 2.5 以上で提案手法の要約文が優れていることを意味する．

また，圧縮性という観点から文圧縮率は要約生成手法に依存せず評価することが可能であり，要約評価において重要である．以下の (3) 式により，Kubo らの手法の要約文長に対する提案手法の要約文長の割合を算出した．

$$\text{文圧縮率 [\%]} = \frac{\text{提案手法の要約文長}}{\textit{Kubo}\text{ らの手法の要約文長}} \times 100 \qquad (3)$$

提案手法と Kubo らの手法の要約文の被験者による比較評価の結果と文圧縮率を表 4 に示す．

表 4 から，比較評価の平均値は中間値の 2.5 を超えており，文圧縮率も考慮すると，必要な要素のみを簡潔にまとめた要約生成ができたといえる．6 試合の被験者による評価結果から提案手法の有効性が確認できた(注10)．

(注9)：1:Kubo らの手法の要約文が適切〜4:提案手法の要約文が適切
(注10)：要約評価で広く利用される ROUGE による評価も行った．ROUGE-1

表 5　生成された要約文の例

sub-event 名	手法	要約文
セーブ	提案手法	スアレスのシュートをブッフォンがセーブ
	Kubo らの手法	グレイトセーブ！ブッフォン！
ゴール	提案手法	マルキージオのヒールパスからテベスのシュートをテアシュテーゲンが弾きモラタがゴール
	Kubo らの手法	きたぁぁーー !! あのヒールパス素晴らしすぎるタイミングだ←モラタ美味しいなー笑
キックオフ	提案手法	前半戦のキックオフ
	Kubo らの手法	いよいよなでしこの決勝トーナメントオランダ戦キックオフ．切り裂け川澄，通せ宮間，気分はバンクーバーの弾幕
選手交代	提案手法	大野の交代，岩渕が投入
	Kubo らの手法	日本 1 枚目．大野アウトで岩渕イン．岩渕，ドリブルで切り裂いたれ !!

5. 考　　察

生成された要約文の例を表 5 に示す．表 5 の sub-event 名は人手で最もらしいものを名づけた．セーブという sub-event に対する提案手法の要約文では“スアレスのシュート”といった Kubo らの手法の要約文には含まれていない情報が含まれている．これは，“スアレスのシュート”という Sub-Event Element (SEE) と“ブッフォンがセーブ”という SEE が複数の tweet に分かれて記述されているため，抽出型要約手法を用いた Kubo らの手法では“スアレスのシュート”といった情報を要約文に含むことができていない．同様に，ゴールという sub-event に対する提案手法の要約文は“マルキージオのヒールパス”，“テベスのシュート”，“テアシュテーゲンが弾く”といった SEE を獲得し，要約文に必要な要素として連結していることが確認できる．このように要約文に必要な情報が複数の tweet に分かれて記述されている場合でも，提案手法では 1 つの要約文に必要な情報を含むことができている．

次に，キックオフと選手交代という sub-event で生成された要約文について考察する．Kubo らの手法の要約文には“切り裂け川澄”や“岩渕ドリブルで切り裂いたれ !!”といった個人の意見と思われる冗長な要素が含まれている．一方，提案手法の要約文は sub-event を説明するのに必要な情報のみを簡潔にまとめた文であり，冗長な要素を含んでいないことが確認できる．

また，文圧縮率では，バルセロナ戦 2 試合と日本戦 4 試合では約 30 ポイントの大きな差が見られた．日本戦 4 試合は W 杯という注目度の高い試合であり，サッカーに精通していないユーザも含む多くのユーザが試合に関する tweet を投稿すると考えられる．そのため，日本戦 4 試合のバースト中の tweet には，sub-event を詳細に説明する tweet はあまり出現せず，(選手名→専門用語) のような情報量の多い SEE がバースト中に多くは出現しない．このような現象から，日本戦 4 試合における要約生成では 1 つや 2 つの SEE しか連結されず，比較的文長の短い要約文が生成された．一方，バルセロナ戦に関する tweet を投稿するユーザはサッカーに精通しているユーザが多いと考えられ，sub-event を詳細に説明する tweet が多く投稿される．よって，バースト中に (選手名→専門用語) のような情報量の高い SEE が多く出現し，バルセロナ戦 2 試合では，日本戦 4 試合で生成された要約文と比べて文長の長い要約文が生成された．このように，試合を実況しているユーザの質により，生成される要約文長が大きく変化したため，文圧縮率も同様に変化したと考えられる．

6. おわりに

本論文では，Twitter を対象としたスポーツの試合における要約生成を行った．sub-event を詳細に説明する Sub-Event Element (SEE) に着目し，要約文に必要な SEE のみを時系列順に連結することで冗長な要素を含まない生成型の要約手法を提案した．

提案手法の要約文と Kubo らの手法の要約文を分析した結果，提案手法の要約文は，要約文に必要な要素のみをまとめた文であることが確認でき，抽出型要約手法の問題点を解決できた．

今後の課題として，サッカー以外のスポーツへの適用やリアルタイムでの要約生成が挙げられる．

謝辞　この研究の一部は科研費 26730176 の助成を受けたものです．

文　　献

[1] Jeffrey Nichols, Jalal Mahmud, and Clemens Drews. Summarizing sporting events using twitter. In Proc. of IUI '12, pp. 189-198, 2012.

[2] Mitsumasa Kubo, Ryohei Sasano, Hiroya Takamura, and Manabu Okumura. Generating live sports updates from twitter by finding good reporters. In Proc. of WI '13, pp. 527-534, 2013.

[3] J. Kleinberg, Bursty and hierarchical structure in streams. In Proc. of KDD '02, pp. 91-101,2002.

[4] Hiroya Takamura, Hikaru Yokono, and Manabu Okumura. Summarizing a document stream. In Proc. of ECIR '11, pp. 177-188, 2011.

[5] 高村大也，奥村学. 施設配置問題による文書要約のモデル化. 人工知能学会論文誌, Vol.25, No.1, pp. 174-182, 2010.

[6] Beaux Sharifi, Mark-Anthony Hutton, and Jugal Kalita. Summarizing microblogs automatically. In Proc. of HLT '10, pp. 685-688, 2010.

[7] 池田和史, 柳原正, 服部元, 松本一則, 小野智弘. 口語文書の解析精度向上のための助詞落ち推定および補完手法の提案. 情報処理学会研究報告, Vol.2010-DBS-151 No.39, 2010.

[8] 工藤拓, 賀沢秀人. Web 日本語 N グラム第 1 版. 言語資源協会発行, 2007.

[9] Lin Chin Yew, Eduard Hovy. Automatic Evaluation of Summaries Using N-gram Co-occurrence Statistics. In Proc. of NAACL and HLT '03, pp. 150-157, 2003.

で 0.07（提案）と 0.14（Kubo）と大きな差があったが，(1)Kubo らの手法とは要約率が異なること，(2)ROUGE では単なる表層的な一致度しかみないことなどの理由から，本評価では被験者評価を優先している．