福井大学リポジトリ
UNIVERSITY OF FUKUI
REPOSITORY

Soccer sound commentary system using Twitter data

メタデータ	言語: jpn 出版者: 公開日: 2019-03-20 キーワード (Ja): キーワード (En): 作成者: 三輪, 将吾, 小高, 知宏, 黒岩, 丈介, 白井, 治彦, 諏訪, いずみ, Shogo, MIWA, Tomohiro, ODAKA, Jousuke, KUROIWA, Haruhiko, SHIRAI, Izumi, SUWA メールアドレス: 所属:
URL	http://hdl.handle.net/10098/10587

Twitterデータを利用したサッカー音声実況システム

三輪　将吾*　小高 知宏*　黒岩 丈介**　白井 治彦*** 諏訪いずみ**

Soccer sound commentary system using Twitter data

Shogo MIWA*,　Tomohiro ODAKA*, Jousuke KUROIWA**,
Haruhiko SHIRAI*** and Izumi SUWA**

(Received February 1, 2019)

In recent years, the Internet and communication technology have been dramatically spreading. Many people are using services on the Internet regardless of time or place. Among them, social media are used by many people. In this research, we will focus on Twitter among social media. We aim to develop a system that enables sound commentary of football using Twitter data. We acquire only the necessary data from Twitter, process it properly, and generate commentary sentences. Speech output can be enabled from generated commentary sentences using text-to-speech engine. To extract and output the optimum Tweet, a document summarization algorithm is utilized in the system.

Key words : Twitter data,Natural language processing,LexRank,Synthesized voice

1. はじめに

近年、インターネットや通信技術が飛躍的に普及してきている。それに伴って、スマートフォンやタブレットなどの携帯端末が普及し、多くの人々が時間や場所を問わずにインターネット上のサービスを利用するようになっている。

その中でもソーシャルメディアと呼ばれるサービスが世の中の人々に広く受け入れられ、利用されている。ソーシャルメディアとは、インターネット上で展開される情報メディアのあり方で、個人による情報発信や個人間のコミュニケーション、人の結びつきを利用して情報が社会に拡散されるように設計されたメディアのことを指す。ソーシャルメディアでは一般の人々から著名人、企業や国の組織など別け隔てなく利用されており、全ての人々が情報を世界に向けて発信することが出来る。

近年、ソーシャルメディアに注目しているのは一般的なユーザーのみにとどまらない。多くの研究機関や企業がソーシャルメディア上のデータの二次的活用について注目している。ソーシャルメディア自体が持つ特徴や、そこで投稿される情報の特徴を活かして新たな知見を得るための様々な研究が行われている。

例を上げると、Twitter上の投稿を用いて、災害時の救助活動の支援に役立てるような研究がある[1]。また、その他の研究としてTwitter上の投稿からスポーツ速報を行うシステムの研究も行われている[2]。

これらの研究では、テキスト形式の投稿に対し様々な処理を行い、テキスト形式で結果を出力させるのが一般的である。そこで本研究では、Twitter上のデータと合成音声の技術を用いることで、ラジオのような形式で様々なイベントの状況把握が可能なのではないかと考えた。

本研究では、Twitter上でサッカーの試合観戦者による投稿を元にして、試合のリアルタイム音声実況を可能にするシステムの開発を目的とする。リアルタイムで試合観戦者のみのTweetを取得し続け、それらを適切に処理し、テキスト読み上げエンジンを用

* 大学院工学研究科 原子力・エネルギー安全工学専攻
** 大学院工学研究科 知能システム工学専攻
*** 工学部技術部
* Nuclear Power and Energy Safety Engineering Course, Graduate School of Engineering
** Human and Artificial Intelligence Systems Course, Graduate School of Engineering
*** Technical Division

いて音声出力する。そうすることで Twitter 上のデータからリアルタイムに試合の状況把握を可能にする。また本システムでは、得点などの重要な出来事が発生した際、その時間帯に得られた Tweet から重要内容が含まれる Tweet を抽出する手法を導入する。そうすることで、取得したデータ量に関わらず、リアルタイム性を保持したより的確な音声実況を再現する。

本論文の流れは以下のとおりである。2 章では、本システムの全体構成と、データの取得や処理に関する点について具体的に述べる。3 章では、実際のサッカーの試合を対象にデータの取得実験を行った結果と、重要 Tweet の抽出に関する実験結果について述べ、最終的に音声出力される実況 Tweet の例を示す。4、5 章では、本システムに対する考察と、今後の展望について述べる。

2. Twitter を用いたサッカー音声実況システム

2.1 システム構成

本研究では、Twitter データを用いてプロサッカーの試合の音声実況を可能にするシステムの開発を行う。そのためにはリアルタイムに取得した試合観戦者の Tweet を適切に処理して、音声出力を行う必要性がある。また、取得するデータ量に関わらずリアルタイム性を確保した音声実況を行うために、一定時間に得られた Tweet から重要な文のみを音声出力することが適切であると考えた。

これを実現するための手法について本章で具体的に述べていく。まず、試合観戦者による Tweet の投稿から音声出力までの本システムの流れを図 1 に示す。

本システムはサッカーの試合観戦者が存在し、Twitter 上に実況 Tweet を投稿するということを前提としたシステムである。試合開始と共に本システムを起動させることによってリアルタイムでの音声実況がスタートする。Twitter API を用いて本研究で考案した取得条件で、試合観戦者のみの実況 Tweet の取得を行う。観戦者による実況 Tweet はリアルタイムで取得され、ストリーミング形式でクライアント PC に保存される。具体的な取得条件については 2.2 で述べる。

取得した、実況 Tweet に対してはいくつかの処理を行う。具体的な処理内容については 2.3 で述べる。この処理を行うことで、より適切な実況文を生成して、音声出力を行うことを可能にしている。処理後の実況文を音声データに変換して、出力することで試合の音声実況を実現する。

また、データ量が膨大であった際にも、リアルタイム性を保持するためのアルゴリズムを導入している。

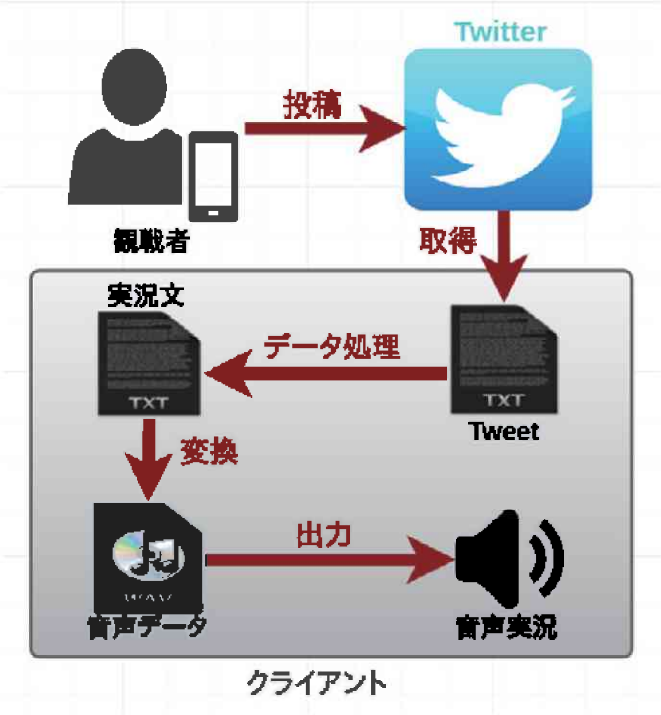

図 1 システム構成図

文書要約アルゴリズムを用いることで、複数の Tweet から適切な実況文を出力可能となる。これについては 2.4、2.5 にて詳細に述べる。

実況 Tweet の取得から音声出力までの流れは、試合時間帯に絶えず行われ、ストリーミング形式でリアルタイム性を保持して音声出力が行われる。

本章の各節において、データの取得条件や具体的なデータの処理手法、音声実況のためのアルゴリズムについて詳細に説明を行う。

2.2 Twitter データの取得

本システムでは Twitter API を用いて、必要なデータの取得を行う。使用する API は Filter realtime Tweets API の Status/filter である。これによって、リアルタイム性を確保してストリーミング形式で Tweet の取得を行う。

試合観戦者による Tweet のみを適切に取得するために今回は以下の様な条件を与え、 Tweet データの取得を行う。

- 日本語による Tweet のみを取得
- #”対戦チーム名” を含む Tweet を取得

本研究では、日本語によって投稿された Tweet のみを対象としているため、日本語 Tweet のみを取得する。

Twitterでは、ユーザーがTweetの投稿を行う際に任意のハッシュタグを追加することができる。このハッシュタグを用いることで、ユーザーは自身の投稿がどのようなジャンルについての内容であるのかカテゴライズすることができる。そこで本システムでは、ハッシュタグに注目した。サッカーの試合観戦者によるTweetの中には、対戦しているチーム名を表記しているハッシュタグを記述しているものが多く存在する。試合時間帯においては、このようなハッシュタグが記述されたTweetは、試合観戦者によるものがほとんどであると考えられる。そのため本システムではハッシュタグによって、必要なTweetの取得を行う。

2.3 Twitterデータの処理

2.2述べた取得条件でで取得したデータは、大多数が試合観戦者によるTweetであるが、試合状況把握か可能であるかという点で、不必要と考えられる内容のTweetも多く見られた。そのため取得した後に、実況文として不適切なTweetを出力対象とさせないためのアルゴリズムを導入する。また、本システムは最終的な実況Tweetの出力手法として、テキスト読み上げエンジンを用いた音声による出力を検討している。音声による出力を行うために、取得したTweetのテキストについて、適切に加工を行う必要がある。本説では取得したデータを適切に音声出力するためのデータの処理に関して述べていく。処理を行った最終的なTweetを実況文の出力対象とする。

具体的には、与えた条件で取得したTweetから以下のようなTweetを判別して実況文の対象Tweetから排除する。

- リツイート
- リプライ
- URLを含むTweet

本システムで使用するTwitter APIであるStatus/filterでは、指定した条件に該当していれば、リツイートやリプライといったTweetデータも全て取得される。リツイートとは他のユーザーによるTweetを自分のフォロワーに向けて引用してTweetをする機能であり、他者のTweetを拡散するために用いられる。つまり、リツイートは元となるTweetと同一内容の情報ということになる。同一内容のTweetを多数出力してしまう可能性を排除するためにリツイートと判断したTweetは出力対象から排除する。

リツイートと同様に、リプライも出力の対象からは除外するようにしている。リプライとは基本的に特定のユーザーへの返信という意味で使用されるため、プライベートな内容となる可能性がある。また、個人に対しての投稿であるため、試合に対してのコメントが乏しいのではないかと考えられる。

それ以外にURL付きのTweetを不適切なTweetと考え、排除するようにしている。サッカーの試合中には、TVやニュースサイトや、その他の公式サイトなどのアカウントによるTweetが投稿されている。そのようなアカウントによるTweetは、外部サイトに誘導させるためにリンク付きのTweet投稿されている場合がある。リンク付きのTweetには、外部サイトのURLが付加されている。このような、URL付きのTweetは試合以外の情報が含まれているため、不要であるとして出力対象から排除する。

不要と考えられるTweetを識別した後に、出力対象となるTweetを本システムで使用するのに適切な形に加工する必要がある。具体的には各Tweetに対して以下のような加工を行う。

- 全てのハッシュタグの削除
- 絵文字や顔文字の削除
- Tweetの末尾にに句点を追加

今回取得するTweetは、全てハッシュタグが付加されたTweetである。本システムでは必要となるTweetを取得するためにはハッシュタグは必要であるが、音声出力による実況を行うには不必要な情報である。そのためハッシュタグは出力対象となるTweetから削除する。ハッシュタグは、様々なソーシャルメディアで用いられる記述方法であるが、TwitterではTweetの末尾に付加するのが一般的となっている。その特徴を利用して#以下をTweetから削除することで、Tweetから全てのハッシュタグを削除するようにしている。

また、出力対象となるTweetからは絵文字や記号の削除も行う。Twitterユーザーのほとんどがスマートフォン上からTweetの投稿を行っている。そのためTweetにはスマートフォン用の絵文字や、顔文字が含まれていることが多々ある。図で示したいくつかのTweetの例でも絵文字を用いていることが分かる。これらは不要な情報であり、音声ファイルへの変換や、後に説明する重要Tweet抽出のアルゴリズムにおいても支障をきたす可能性がある。そのためこれらがTweet中に含まれる場合は削除する。

また、Tweetからハッシュタグや絵文字、記号を削除した後に、Tweetの文末に句点を追加するようにしている。これは、後に説明を行うLexRankによる重要内容Tweetの抽出を適切に行うための処理であ

る。Tweet 抽出の際に、一つの Tweet を一つのセンテンスとして認識させるためにこのような処理を行っている。

ここまでで説明した全ての処理を行った Tweet を音声出力の対象となる実況文として、次の異なる処理を行う。

2.4 音声実況のリアルタイム性保持アルゴリズム

本システムでは、取得した実況 Tweet を音声データに変換した後に、再生することで音声による出力を可能にする。音声出力を実現するために OpenJTalk というシステムを利用した。OpenJTalk とは名古屋工業大学で開発された、日本語テキストの読み上げを可能にするための合成音声システムである。BSD ライセンスを取得しておりフリーソフトウェアとして利用することが出来る。これを用いることで、テキスト形式のファイルを WAV ファイルに変換して音声出力することが可能となる。

音声出力をリアルタイム性を保持して適切に行うためには、考慮すべき点が存在する。音声による出力では、一定時間に出力するデータ量を適切に設定しなければならない可能性がある。実況文の対象 Tweet をそのまま音声出力することは、取得したデータ量によっては問題となる場合がある。それは、音声によるテキストの読み上げを行うことで一定の時間を浪費してしまうためである。これが原因となり、一定時間に対する出力量に関する制約を設けなければ、リアルタイム性が阻害されてしまう恐れがある。

そこで、リアルタイム性を保持した音声出力を常に行えるようにするために、一定時間に対して出力する Tweet を制御するためのアルゴリズムをシステムに取り入れる。Tweet データの取得、処理、加工を行ってから音声出力までのアルゴリズムの流れを図 2 に示す。Tweet の取得を開始してから 10 秒間に得られた Tweet から一つを選択し、それを音声出力の対象とする。これを 10 秒毎に繰り返し行うことで、リアルタイム性を保持した音声出力を行う。

選択する実況 Tweet は無作為に選択するのではなく、文書要約に用いられるアルゴリズムを利用することによって、いくつかの Tweet から出来るだけ重要な内容を含む Tweet を選択する。そのためのアルゴリズムとして LexRank という文書要約アルゴリズムを導入する。

2.5 重要 tweet の抽出

複数の Tweet から 1 つの Tweet を抽出するために、本システムに文書要約に用いられるアルゴリズムを

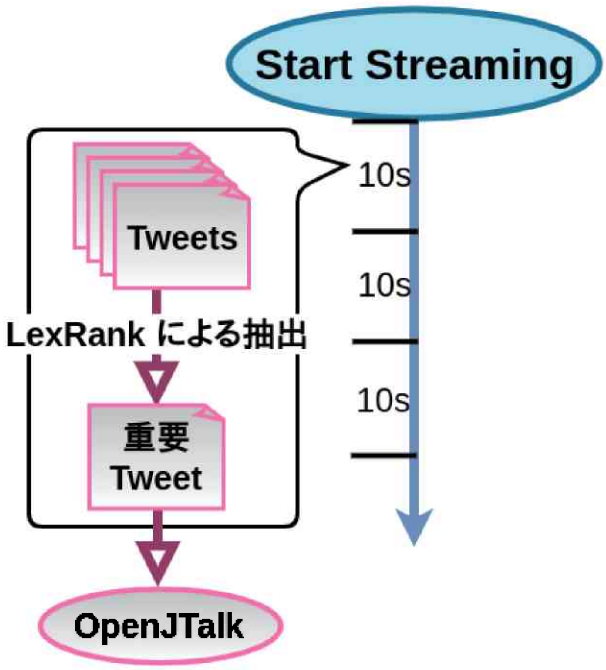

図 2 リアルタイム性保持のためのアルゴリズム

導入する。本節では、そのために適したアルゴリズムのについて検討し、システムに実装する手法について述べる。

文書要約の手法については様々な研究が行われており、現在一般的な文書要約のアルゴリズムは大きく分類すると以下の 2 種類が存在する。

- 抽出型 (Extractive)
- 抽象型 (Abstractive)

抽出型は、要約対象の文章の中から重要と考えられる文を抽出することによって、その文章の中で特徴的な文を要約文とする手法である。抽象型は、人間が文章の要約を行うように、文章の意味を読み取った上で新たな文を生成することによって要約を行う手法である。今回は、いくつかの Tweet からできるだけ重要な内容の Tweet を抽出したいため、抽出型の文書要約アルゴリズムを適用させるのが適切であると考えた。

そのために Erkan らが PageRank の概念を元に提案した、抽出型文書要約のアルゴリズムである LexRank を用いる [3]。LexRank は、入力文書からグラフ構造を生成して、重要な文のランキングを作ることで重要と考えられる文を発見し、それを要約文として出力する。LexRank では以下の 2 つの概念を重要視して文書要約を行う。

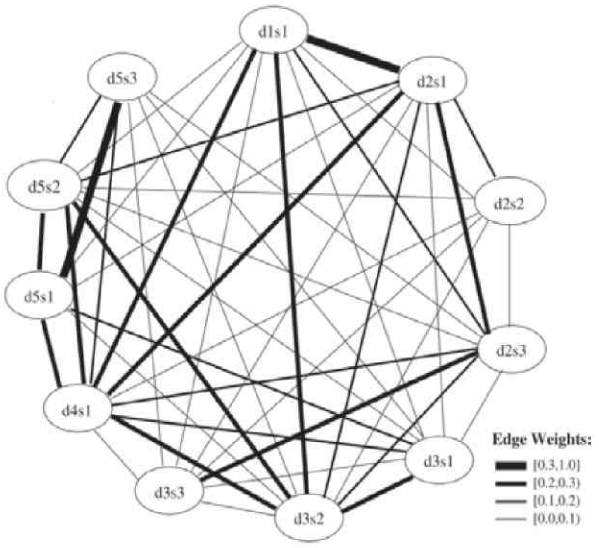

図 3 LexRank による類似度グラフの例

- 多くの文に類似する文は重要である
- 重要な文に類似する文は重要である

この概念を元に、ノードを文、エッジを 2 文間の類似度として入力文章の無方向グラフを生成する。提案論文では、TF-IDF によって、文同士のコサイン類似度を算出している。TF-IDF は情報探査においてよく用いられる手法であり、文章中に出現した単語の重要度を評価する手法の 1 つである。

上記の理論をグラフにして可視化したものを図 3 に示す [3]。各ノードには dXsY と示されている。これは、文書 X の Y 番目の文であるということを示している。エッジ数の多いノードは、多くの文と類似していることを示している。また、エッジの太さは、ノード間の類似度の高さを示している。図 3 の場合は、エッジ数が多く、かつエッジが太い d5s1 や d4s1 が重要度が高いと判断し、要約文の候補となる。本システムでは、重要度が最も高いものを 1 つ抽出するようにしている。

3. 動作実験

3.1 Twitter データの取得実験

前章までで検討を行ったデータの取得方法とデータ処理を行い実況文の対象となる Tweet の抽出までのシステムの動作実験を行った。そこで得られたデータに試合の状況を把握できるような Tweet が含まれているのか、またリアルタイム性を保持した実況を可能にするためにどのような処理が求められるのかについて検討し、リアルタイム性を保持するためのアルゴリズムの実装について考察を行う。そこで今回は 2018 年に開催された FIFA ワールドカップのいくつかの試合を対象に試合観戦者の Tweet の取得実験を行った。例として、ある試合に関してのデータの取得実験の詳細を以下の表 3 に詳細を示す。

表 1 動作実験の詳細

実験概要	詳細
試合日時	2018 年 6 月 24 日
対戦チーム名	イングランド vs パナマ
試合結果	イングランド　6 - 1　パナマ
得点時間 (分)	前半 : 8、22、36、40、46 後半 : 62、78
取得 Tweet の使用言語	ja (日本語)
Tweet の取得条件	# eng (イングランド) or # pan (パナマ) を含む Tweet
取得する Tweet データ	タイムスタンプ、ユーザー名、Tweet
データ処理後の Tweet 数	前半 : 約 800 件 後半 : 約 400 件

この試合は、他に取得対象とした試合と比較して得点などのサブイベントの発生が多数見られた試合であるため例として選択した。サッカーの試合は、前半戦と後半戦の 2 つのパートで構成されている。試合時間はそれぞれ 45 分間である (それぞれ、アディショナルタイムが追加される可能性がある)。得点が多数見られた前半戦の時間帯において約 800 件の出力対象 Tweet が得られ、得点の少なかった後半戦と比較して倍の Tweet が得られた結果となっている。今回取得する Tweet データとして投稿時間を示すタイムスタンプと、投稿者を示すユーザー名も取得するようにしている。取得し、データ処理行った後の Tweet データを取得順にテキストファイルに記述するようにして実験を行っている。

例に示した、試合において前半戦で、 Tweet 数がどのように推移したのかを時系列グラフにしたものを図 4 に示す。

グラフを見ると、ある時間帯に Tweet 数が上昇していることが分かる。表 1 と 図 4 を対応して見ると、Tweet の上昇が見られる時間帯は、試合中に得点発生した時間帯とリンクしていることが分かる。それぞれ

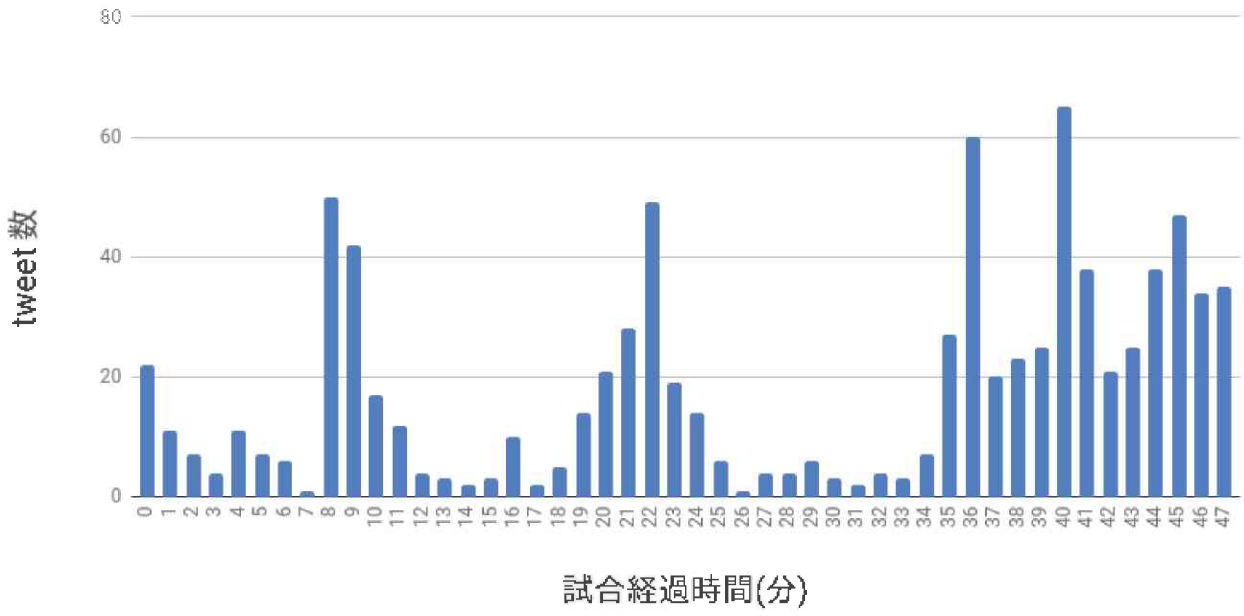

図 4 実況文対象 Tweet の時間ごとの変位 (前半戦)

の得点時において、得点発生から約 1 〜 2 分間に極端に Tweet 数が上昇している。このことから、Tweet 数の上昇が見られた時間帯においては、得点に関する情報が含まれた Tweet を多数取得しているということが考えられる。

Tweet の上昇が見られた際の実況 Tweet の一部を表 2、3 に示す。ここに示したのは、前半の 8 分と後半 17 分において取得できた Tweet の一部である。表 2、3 を見ると、得点に関する情報が含まれていることが分かる。どちらのチームの得点であるのか、どの選手による得点であるのかといった大まかな情報から、得点に至るまでの経緯を解説するような内容まで様々な内容の Tweet が存在していた。その他には、得点に対してユーザーの感情のみを表すような内容の Tweet も多数見受けられることが分かった。表 2 では、「イングランドによる得点」「ジョン・ストーンズ選手による得点」「先制点」といった、サブイベントに付随する重要な内容を含む Tweet が多数存在しており、適切に処理、出力することで得点に関する殆どの情報を把握できると考えられる。

今回示した例以外の得点発生時においても同様な内容の Tweet が多数取得されており、得点というサッカーの試合における主なサブイベントの発生を把握することが可能である内容の Tweet を取得、抽出できているという結果といえる。また、選手交代や決定機などの際にも、 Tweet の上昇が見られる点があり、同様にそれらのサブイベントに関して言及した Tweet が見られるということが分かった。

取得実験の結果、得点発生時などには過度な Tweet の集中が見られ、そのまま音声出力を行うことは、出力内容のリアルタイム性を損なう恐れがあることが分かった。この試合において最も Tweet を取得した時間帯 (後半 33 分) には約 70 件の Tweet が取得されている結果となっている。この約一分間の Tweet を取得順に音声出力した結果、約 80 秒の時間が必要であった。その他のイベント発生時の時間帯においても、一分間に 60 秒を超える出力量になることが考えられる。そのため取得したデータ量が膨大だった際、一定時間に出力する Tweet を制御しなければ、リアルタイム性を保持した音声出力が困難であるため、図 2 で示したアルゴリズムを導入し、重要 Tweet のみを抽出して音声出力を行う。

3.2 重要 Tweet の抽出実験

得点時に見られた Tweet を対象に LexRank による重要 Tweet の抽出が適切に行われるのかについて実験を行う。表 2、3 に示しているのは、得点時の約 10 秒間に取得した後に不要 Tweet を排除し、加工を行った Tweet である。今回は表 2、3 で示した Tweet に対して重要 Tweet の抽出を行った結果を例として示す。

今回はそれぞれの Tweet から、重要度の高いと判断された上位 2 つの Tweet を例として示す。表 4 で抽出された Tweet では、「セットプレー」から「イングランド」の「先制」という、表 2 中の Tweet に含

表 2 イングランドの得点時における Tweet の内容

Tweet の取得時間	Tweet の内容
21:08:53	ジョン・ストーンズゴーーーーーーーーール イングランド先制。
21:08:54	ストーーーンズイングランド先制! どフリーや。
21:08:54	イングランド先制すげえな。
21:08:54	キタセットプレー理想的。
21:08:55	わーおイングランド先制ドンピシャヘッド。
21:08:55	イングランド先制。
21:08:57	ペナ内でドフリー、完璧な形からイングランド先制。
21:08:57	セットプレーからイングランド先制。
21:08:58	早速イングランド先制きたー。
21:08:58	イングランドまたセットプレーやんなんやこのチーム。
21:08:59	しゃあーーイングランド先制ジョン・ストーンズナイスヘディング。
21:08:59	ストーンズ。
21:08:59	もうシャワー行っていいかな。

まれる重要なキーワードを含んだ Tweet が要約の上位に位置づけられている結果となっている。表 5 で抽出された Tweet では、「イングランド」による得点であり、「ハリー・ケイン」という選手が「ハットトリック」を記録したといった、この場面での得点について詳細に把握することの出来る内容の Tweet が Tweet が要約の上位に位置づけられている結果となっている。得点以外に言及した内容の Tweet は抽出されず、得点に関する重要な内容を含んだ Tweet が出力できるような結果となった。

10 秒間に得られた Tweet に対して、この手法で抽出を行い、上位 2 つの Tweet を音声出力の対象として、OpenJTalk によって音声ファイルへと変換し出力する。これを繰り返し行うことによっていくつかの Tweet から重要内容を含んだ Tweet を、リアルタイム性を保持して出力することが可能となる。

表 3 イングランドの得点時における Tweet の内容

Tweet の取得時間	Tweet の内容
22:21:00	イングランド 6 点目無慈悲なスコア（決まり方そのものが無慈悲）ハリー・ケインのハットトリック(暫定得点王浮上)。
22:21:01	9 番ケインハットトリック達成。
22:21:01	イングランド 6-0 パナマケイン、今大会 2 人目のハットトリック。
22:21:01	わーい、ヴァーディ来る〜。
22:21:02	ケインハットトリックおめでとう。
22:21:02	ケインのハットトリックでイングランド 6 点目。
22:21:02	サッカーロシア W 杯前半イングランド 6—0 パナマイングランド、ケインのゴール、ハットトリック。
22:21:06	イングランド 6 点目あれケインのゴールになるのか笑意外な形でハットトリック達成。
22:21:08	ゴール確定ハットトリックか。
22:21:10	ともかく、ケインおつヴァーディ。

表 4 重要 Tweet の抽出 (表 2)

ランク	抽出された Tweet
1	セットプレーからイングランド先制。
2	イングランドまたセットプレーやんなんやこのチーム。

表 5 重要 Tweet の抽出 (表 3)

ランク	抽出された Tweet
1	イングランド 6 点目無慈悲なスコア（決まり方そのものが無慈悲）ハリー・ケインのハットトリック(暫定得点王浮上)。
2	サッカーロシア W 杯前半イングランド 6—0 パナマイングランド、ケインのゴール、ハットトリック。

4. 考察

4.1 Tweet の取得実験に関する考察

今回は、ハッシュタグに着目してサッカーの試合観戦者の Tweet の取得を行った。与えた条件で取得した Tweet から、リツイートやリプライ、URL 付きの Tweet を排除した。その結果ほとんどの Tweet が試合に対して言及した Tweet であったため、今回の取得条件と取得後後の処理手法で試合観戦者による Tweet を適切に取得することが可能であると言える。

サッカーの試合では、得点や決定機、選手交代といった事象が試合における重要なサブイベントであるが、それらが発生した時間帯では Tweet の上昇が見られた。またそのポイントで取得した Tweet はそれらの事象に言及したものがほとんどであり、試合状況を把握可能である内容となっていた。

4.2 音声出力とリアルタイム性に関する考察

本システムでは OpenJTalk を利用することで、音声による実況文の出力を可能にした。取得した Tweet を音声ファイルに変換し音声による出力を可能にすることで、本システムをラジオを聞くような感覚で利用することができる。テキスト形式による実況と音声による実況をうまく使い分けることで利用者は様々なシュチュエーションで本システムを利用できるのではないかと考えられる。

取得実験の結果、得点時において過度な Tweet の集中が見られることがわかった。そのため一定時間に出力するデータ量を制限しなければリアルタイム性を確保した音声出力が困難となる場合があることが分かった。そこで 10 秒間に取得した Tweet のうち 2 つのみを音声出力の対象とすることによってリアルタイム性を保持した音声実況を可能にした。

4.3 重要 Tweet の抽出に関する考察

複数文書要約の手法である LexRank を用いた、重要 Tweet の抽出実験を行った。10 秒間に得られた Tweet からこの手法を用いて出来るだけ重要な内容を含む Tweet を出力することで、より正確な音声実況を行うことが可能となると考えられる。得点時においては、観戦者の感情のみを示した Tweet と、どのようなシュチュエーションでどのチームの誰による得点かについて説明した Tweet が混在していた。複数の文から類似する文を判断し、それに類似する文を重要文とする LexRank を用いることで、後者のような得点についての情報を多く含んだ Tweet を抽出出来ている結果となっていた。

5. まとめと今後の課題

サッカーの試合観戦者が Twitter 上で試合に関して言及した Tweet をリアルタイムで取得、処理、出力することで、試合状況を把握することが出来るシステムを提案した。音声による実況を行うことで、本システムをラジオ実況を聴く感覚で利用することが出来る。そのため利用者は、手の離せない状況でも音声によって試合状況を把握することが可能である。

今後は、本システムを実際に使用してもらい、試合状況が適切に把握出来たかどうかについて検証する必要がある。

本稿ではサッカーの試合実況に重点を置いて論述したが、本システムは災害時における現地の状況把握を計るためのシステムとしての発展を目指している。災害時においてはサッカーの試合実況と異なり、注目しなければならない事象が単一ではないため、異なる取得条件や処理手法を検討する必要がある。また、個人が欲している情報なのか、消防署や政府が欲している情報なのかといった情報ニーズを意識する必要がある。本研究では、被災者の救助要請 Tweet に着目し、消防署や自衛隊の救助活動を支援するようなシステムを開発できないかと考えている。

本システムを災害状況把握に利用するために、今後は災害時の Tweet について調査し、適したデータの処理手法を検討していきたい。

参考文献

[1] Muhammad Imra, Fernando Diaz, Carlos Castillo, and Sarah Vieweg. Processing social media messages in mass emergency:survey summary. WWW 2018, April 23-27, 2018, Lyon, France, pp. 507511.

[2] 久保 光証, 笹野 遼平, 奥村 学, 高村 大也. “ 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 第 19 回年次大会 発表論文集 (2013 年 3 月). pp.138 - 141

[3] G. Erkan and D. R. Radev. Lexrank: Graph-based lexical centrality as salience in text summarization. Journal of Artificial Intelligence Research, pp. 457 479, 2004.