スポーツ解説生成のための選手の意外情報抽出手法

11471005 池内 雄大（灘本研究室）

あらまし：スポーツの試合の放送での解説は，その選手と競技に関する情報がほとんどである．そのスポーツのコアなファンではない視聴者は親しみにくい解説であると感じている場合がある．そこで本研究では，サッカーを対象とし，選手自身の Twitter を用いて，選手に関するサッカー以外の意外性のある情報を抽出する手法の提案を行う．

1．はじめに

現在，サッカーワールドカップや，3 年後の東京オリンピックなど，様々なスポーツの国際大会が開催されている．それに伴い，様々なメディアが，これらのスポーツ大会を取り上げている．これにより，人々の関心がスポーツに大きく寄せられ，普段観戦しないようなスポーツの放送も視聴しようとしている．このように，そのスポーツについてコアなファンでない視聴者層の増加が見込まれる．

一方，現在のスポーツ放送において視聴者が取得することが可能な情報として，年齢やポジション，所属クラブ，プレーの特徴など，そのスポーツに関係のある情報の解説が行われることが一般的である．このような解説に対して，そのスポーツについてコアなファンでない視聴者は，親しみにくい解説であると感じている場合がある．それに対し現在，細かすぎる解説をするスポーツジャーナリストの増田明美さんの解説が注目されており，選手自身に着目した個性的な解説を行っている．例えば，選手の趣味や家庭での生活状況など様々である．

また，近年 Twitter などの SNS が普及し，誰でもインターネット上に気軽に投稿を行っている．また，プロスポーツ選手も同様に様々な情報の投稿を行っており，選手の私生活の情報を取得することができる．

本研究では，このような個性的な解説を自動で生成し，視聴者に提示するために，選手の意外性のある情報を取得することを目的とする．これにより，スポーツ放送において，より一層親しみやすさを感じさせ，オリンピックのような大きな国際大会のみならず，隔週で行われているような規模の小さい試合まで興味が惹かれるような，視聴意欲の向上に繋げる解説を生成することが可能となる．本研究では，様々なスポーツの中でも，サッカーにスポットをあて，選手に着目した意外性のある個性的な面白い情報を推薦する．

2．関連研究

佃[1]らは，人物名，地域名，製品名，施設名，および組織名の 5 つのカテゴリの主題語を用い，Wikipedia を用いて，それぞれの同位語間の関係を考慮し，意外な情報を発見する手法を提案している．それに対して，本研究では選手自身のツイートに着目し，その選手に関係のある話題語を抽出する点が異なる．

伊藤[2]らは，ウェブサイト All About のデータをコーパスとし，ユーザの発話から名詞，動詞，形容詞，未知語を特徴語として抽出することで，各記事との関連度，意外性を算出し、ユーザに意外性のある記事を提示する手法を提案している．それに対して，本研究では，特徴語を抽出する対象は，ユーザに提示する情報の主題者であり，抽出する特徴語は名詞のみを用いる点が異なる．

3．全体の流れ

以下に提案手法の処理の流れを示す．

① 情報を取得したい選手名を入れる．
② その選手自身の Twitter から 200 ツイートを取得する．
③ 取得したツイートから選手に関する話題語を抽出する．
④ 抽出した話題語からサッカーに関する語と，私生活語らしからぬ一般語を取り除く．
⑤ 選手名と私生活キーワードをユーザに提示し，興味のある話題を選択させ，情報を提示する．

4．意外情報抽出手法

4.1 話題語の抽出

取得した各選手自身の 200 ツイートに対してクラスタリングを行う．この際クラスタリング手法には短文のクラスタリングに優れているといわれる Repeated Bisection 法[3]を用いる．クラスタリング対象は名詞のみとする．また各クラスタには話題を示す名詞であるトピックが複数存在するが，本研究では各クラスタの中心ベクトルの語を話題語とする．

4.2.1 サッカー以外の語の抽出

抽出した話題語からサッカー用語を取り除く

ために，辞書を用いて比較した差分を取得する．使用する辞書は，手作業で選定したサッカー用語から，Word2Vec を用いて語を拡張する．それに加え，国名リストも追加した 4502 件を使用する．ここで用いた Woed2Vec は 2016 年 11 月 3 日から 2017 年 6 月 7 日までに収集した 295,759,657 ツイートを用いた．

4.2.2 私生活キーワードの決定

話題語と辞書との差分を私生活キーワードとして決定する．このとき取得した語には，「今週」，「緊張」のように，その語だけでは何の話題か理解することが難しい語が含まれる．これらの語に対して，形態素解析 Juman[4]を用いて，表 1 のルールに当てはまる語を削除する．また表 2 に抽出した私生活キーワードの一例を示す．この私生活キーワードを用いて，Web 検索をした結果を意外情報と呼ぶ．

表 1．私生活キーワードからの除外条件

No	条件
1	人名である語
2	数詞
3	時相名詞
4	ドメインを持たない抽象物

表 2．乾選手の私生活キーワード例

スタバ	リンカーン	ワンピース
ワイン	水族館	ピアニカ
寿司	パパ	家族

4.3 提示方法

抽出した私生活キーワード及び Web ページを提示するユーザインタフェースを作成する．ユーザには，はじめに意外情報を取得したい選手を提示する．ユーザは情報を取得したい選手を選択すると，システムはその選手の私生活キーワードを抽出し提示する．そこから，選択された選手名と私生活キーワードを検索クエリとして Web 検索をしたページを提示する．図 2 に作成した Web ページを示す．

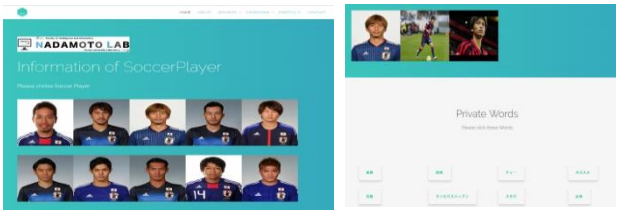

図 2 ：提示例

6．実験

提案手法の有用性を示すために実験を行った．

6,1 実験条件

実験にはサッカー選手 5 名のツイート 200 件を用い，提案手法により私生活キーワードを抽出し，抽出した語が正解であるかを手作業で判定する．ここで正解とするのは，選手名と私生活キーワードを検索クエリとして Web 検索をした際，上位 5 件にサッカー以外の情報の記述がされているページが存在することとする．

表 3.実験結果

選手名	適合率	再現率	キーワード数
長友選手	0.7586	0.9167	29
乾選手	0.4490	0.7857	49
槙野選手	0.5741	0.7333	54
香川選手	0.5333	0.7273	15
岡崎選手	1.0	1.0	2

6.2 結果と考察

サッカー選手 5 名における私生活キーワードの実験結果を表 3 に示す．全選手に共通して適合率より再現率が高い結果となったが，これは不要な語の取り除きが不十分であるためと考えられる．また不要な語の中にサッカー用語はほぼ含まれていなかったため，サッカー用語を取り除くべく作成した辞書に問題は無く，ルールベースの部分でルールの改定，追加を行う事で精度が向上するのではないかと考えられる．

7．まとめと今後の課題

本研究では，選手自身のツイートに着目し，私生活キーワードを用いて，選手に関するサッカー以外の意外性のある情報を抽出する手法の提案を行った．今後の課題として，ユーザの趣向にも着目して，意外情報とのパーソナライゼーションを行い，意外情報を解説文として抽出する研究を進めることである．

参考文献

[1]佃洸摂,大島裕明,山本光穂,岩崎弘利,田中克己,”語の認知度と同位語間の関係に基づく意外な情報の発見”,日本データベース学会論文誌 Vol.11,No.3pp.21-26,2013.

[2]伊藤直之,西川侑吾,大野和久,松本征二,中川修,”語の意外度に基づき話題展開する非タスク指向型対話システム”,JSAI2015,2L5-OS-07b-2,2015.

[3] 花井俊介,灘本明代,”酷似レシピ抽出のためのクラスタリング手法の提案”,DEIM2014,F8-6 ,2014.

[4] JUMAN: http://nlp.ist.i.kyoto-u.ac.jp/