深層学習を用いたTwitterからの趣味情報の抽出

Extraction of Interests Information from Twitter Using Deep Learning

若宮 悠希[1*]　砂山 渡[2]　畑中 裕司[2]　小郷原 一智[2]
Yuki WAKAMIYA[1]　Wataru SUNAYAMA[2]　Yuji HATANAKA[2]　Kazunori OGOHARA

[1] 滋賀県立大学大学院工学研究科
[1] Graduate School of Engineering, The University of Shiga Prefecture
[2] 滋賀県立大学工学部
[2] School of Engineering, The University of Shiga Prefecture

Abstract: In recent years, research has been actively conducted to extract personal information from Twitter. However, since comments posted on Twitter are relatively short, and comments on various topics are often made, comments on specific interests appear only locally. Therefore, in this paper, to the related vocabulary related to the designated interest be learned widely by deep learning starting from the word set related to the particular interest, whether or not each Twitter user has a designated interest determine by the proposing system.

1　はじめに

近年，インターネットの発展に伴い，Twitter[1]やfacebookといったSNSなどが広く普及することで，オンライン上でのコミュニケーションの場が広く普及してきた．これにより，近しい周囲の人間にとどまらず，距離に縛られない広い範囲の人間と交流を気軽に行うことができるようになった．

これらのサービスを利用するユーザはそれぞれが違う性格や考え方，趣味，嗜好を持つため，交流を行うときは相手の持つ個性が自分のものと合うか，もしくは受け入れられるかどうかが重要となる．そのため，交流相手の個性を十分に理解して受け入れる，もしくは自らが受け入れやすいユーザを選んで交流することが求められる．

交流相手と気が合うかどうかは，同一の話題について興味をもっているかを判断材料とすることができる．Twitterのような不特定多数のユーザが匿名で多様な話題について自由にコメントできるサービスにおいては，検索機能を利用して同じ趣味を持つユーザを選択して交流することもできるが，プロフィールで趣味を明言しているユーザや，サーチワードを含むコメントを投稿したユーザのみしかヒットせず，潜在的に話題に興味を持っているユーザを探すことができないことがあるため，選択の余地が狭まる．

そのため，Twitterに投稿されたコメント集合から，ユーザが特定の趣味を持つか否かを抽出することができれば，交流相手の個性の理解や，新たな交流相手としてユーザを推薦するなど，ユーザ間の交流を手助けが行えることを期待できる．

Twitterから個人情報を抽出する研究は近年盛んに行われるようになってきているが，Twitterにおいては投稿されるコメントが比較的短く，また様々な話題についてのコメントがなされることが多いため，特定の趣味についてのコメントは局所的にしか現れてこない．

そこで本研究では，特定の趣味に関わる単語集合を起点とした深層学習 (Deep Neural Network) により，指定の趣味に関わる関連語彙を潜在的に幅広く学習させた上で，各Twitterユーザの投稿文であるツイートの集合から指定の趣味を持つか否かを自動抽出することを目的とする．

以下本論文では，2章で関連研究について述べる．3で趣味情報抽出システムについて述べる．4章で提案システムの趣味情報抽出制度の評価について述べ，5章で本論文を締めくくる．

2　関連研究

Twitterに投稿されたコメントから個人情報を抽出する研究は近年盛んに行われてきている．

*連絡先：滋賀県立大学大学院工学研究科 電子システム工学専攻 若宮悠希
〒 522-8533 滋賀県彦根市八坂町 2500
E-mail: of23ywakamiya@ec.usp.ac.jp
[1]https://twitter．com

これまでに，各ユーザの家族構成や所有物，趣味嗜好などを抽出することで個人情報を推定する研究 [1][2] が行われている．これらの研究では，投稿されたコメント (ツイート) に含まれる単語により個人の情報を抽出する手法を提案しており,「俺のギター」「私の子ども」など，一人称所有格の後に名詞が現れているツイートから，そのユーザの所有物を抽出し,「ギター」であれば「音楽」など所有物が趣味嗜好と関係のあるものならば，所有物を起点にユーザの趣味を抽出する．

また，この他に抽出対象ユーザのプロフィール文やツイートではなく，相互に交流関係のあるユーザのプロフィール文を元に属性を抽出する研究 [3] が行われている．この研究では，交流関係のある複数のユーザのプロフィール文に頻繁に出現するものを本人に深く関わりのある単語と仮定して取得することで，これらの単語を元にして対象のユーザに関わりがあると考えられる属性を推定する．

本研究では，抽出したい趣味を利用者があらかじめ決定して関連単語を与えることにより，趣味と関連する単語として幅広い語彙を網羅した学習を行うことで，推定対象のユーザのツイート集合のみから，指定の趣味を持つか否かの判断を行う．

3 深層学習を用いたTwitterユーザからの趣味情報抽出システム

3.1 深層学習を用いた趣味情報抽出システムの構成

本研究で提案する深層学習を用いた趣味情報抽出システムの構成を図 1 に示す．

まず，システム利用者が抽出対象として決定した趣味ひとつひとつに対応する趣味情報抽出ネットワークを構築するため，各趣味の学習用データとして Twitter の投稿文であるツイートを複数収集する．収集したデータを深層学習の一種である Deep Neural Network に入力できる形に整形し，これらを用いて趣味情報抽出ネットワークを学習により構築する．次に，複数の抽出対象ユーザが投稿した各ツイート集合をシステムに入力し，構築した趣味情報抽出ネットワークにより，各趣味に興味を持っているかどうかを 1 ツイートずつ 2 値で判定する．この判定結果を可視化インタフェースを利用してユーザごとに数値化し，提示することで，システム利用者がどのユーザが指定の趣味に興味を持っているのかを解釈する支援とする．

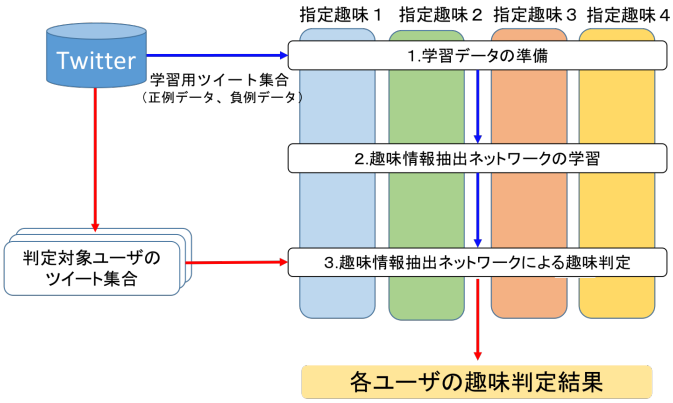

図 1: 趣味情報抽出システムの構成図

3.2 抽出対象の趣味の選定

提案システムは，深層学習により入力データのツイート集合に出現する単語の出現頻度を学習することで趣味情報を抽出するため，抽出対象の趣味として「不変性が高い趣味特有の単語集合を用意できるもの」と定義し，このうち学習用ツイート集合を十分に収集できる趣味を対象として趣味情報抽出システムを構築する．

例えば,「野球」や「将棋」などのように「ルール」「手法」「用語」などユーザによって呟かれる単語が変化しないもので構成されるものと,「jpop」「バラエティ番組」などのように「タレント」「アーティスト」など流行りや趣向によって大きく変化してしまう単語で構成されるものが存在する．そのため，後者の趣味を深層学習により抽出しようと試みると，一部の単語集合のみが学習されることにより，正しく抽出できるユーザに偏りが発生してしまうことから，本システムでの趣味情報抽出には不向きとなる．

3.3 深層学習に用いる学習用データの準備

3.3.1 深層学習に利用するデータ集合

本研究では，趣味情報抽出の対象を Twitter ユーザに定め，各ユーザの投稿したツイート集合を深層学習によって分析し，指定した趣味が文章中に現れているかを抽出する．Twitter ユーザの趣味情報抽出を行う際は，対象ユーザの投稿した複数のツイートを 1 件ずつ推定していくことで，全体のうち何%に趣味が現れているかを元に最終的な判断をする．

3.3.2 深層学習に利用するデータの選定

深層学習に用いる入力データは，推定対象と同じく，Twitter から収集したツイートを利用する．抽出対象

とする趣味が文章に現れているツイートを正例，対象とする趣味が文章に現れていないツイートを負例として正解ラベルをつける．これらのツイート集合を入力データとして，深層学習により趣味情報抽出ネットワークを構築する．

ここで，ツイートに指定の趣味が現れているか否かは，指定の趣味に関連のある単語を設定し，いずれかの単語が文章中に出現しているか否かを判断する基準とする．例えば,「野球」に関していえば「阪神タイガース(プロ野球チーム名)」や「ホームラン (野球用語)」などが考えられる．この単語集合を本研究では「特有単語」と定義する．

Twitter REST API[2]を用いてツイートを収集し，指定の趣味ごとに設定した特有単語を含むツイート，含まないツイートを選別して正例，負例を選択する．

3.3.3 特有単語の設定

特有単語は対象の趣味と関連が深いと考えられる単語となる．正しく趣味情報抽出を行うためには正例データに含まれるツイートは，対象の趣味の話題が現れているツイートのみを限定し，関係のないツイートを収集しないようにする必要がある．そのため，特有単語として設定することに望ましい単語として，その趣味特有の単語であることが求められる．

この際考えられるものの 1 つとして「趣味名」が挙げられるが，抽出対象趣味によっては一般的に広く知られているものとなり，収集されるツイートが趣味の話題が現れているもののみに限られない可能性がある．そのため，本システムでは趣味名を特有単語として設定することはしない．

特有単語は，抽出対象の趣味を決定する際にシステム利用者が設定する必要がある．よって，利用者による特有単語設定を補助するために TF-IDF により特有単語決定手法を提案する．

TF-IDF とは，単語の出現頻度（Term Frequency．以下 TF と表記）と，逆文書頻度（Inverse Documnet Frequency．以下 IDF と表記）の積で表される指標となる．これは複数異なる条件の文書集合に出現する単語に対し，各文書に対する重要度を評価する手法となる．TF は対象の単語がある 1 つの文書中にどれだけの頻度で出現したかを表し，IDF は全文書中に対象の単語が出現した文書の割合を逆数で表したものとなる．これらの積を求めることで，特定の文書にのみ高い頻度で出現する単語を抽出することができる．

TF-IDF は以下の式 (1) により求められる．ここで $tf(t,d)$ はある文書 d における単語 t の出現頻度，$idf(t)$ は全文書中に単語 t を含む文書数，N は全文書数となる．

$$\underline{tf-idf(t,f)} = tf(t,d)(log\frac{N}{df(t)}+1) \qquad (1)$$

TF-IDF を用いることにより，他の趣味の話題に出現しにくい対象の趣味の特有単語を抽出し，特有単語設定の助けとする手法について説明する．まず抽出対象に設定したい趣味と，それ以外に考えられる複数の趣味について，趣味名を含むツイートを複数収集する．1 つの趣味について収集したツイートを 1 つの文書とみなし，複数趣味の文書集合に出現する全ての単語を対象に TF-IDF を算出する．そして，システム利用者の手により抽出対象の趣味に対する TF-IDF 上位の単語のうち，その趣味特有の単語だと考えられるもの 1 語を特有単語に設定する．野球を例とすると，TF-IDF 上位となった単語として「投手」「選手」「満塁」「匿名」が出現した場合，このうち「投手」「満塁」が野球特有の単語であると利用者が判断できたもののうちから，より TF-IDF 値の高い単語を特有単語として設定することが考えられる．特有単語を複数設定すると各単語と同時に出現する単語を幅広く学習することが可能となるが，一般的な単語について広く学習する可能性が高まるため，1 語に絞る．

3.3.4 学習用データの整形

深層学習の入力データや推定対象の文章は，それぞれ文章から BoW(Bag of Words) に変換して利用する．BoW とは全文章中に出現する単語を並べ，各文章での単語の出現頻度をベクトルで表現したものである．また，抽出対象の趣味特有の表現を学習したいため，使用する単語は 15 ツイート以上に出現した名詞のみとし，その中でも判定に関わらないと考えられる単語はあらかじめ除去する．除去する単語としては,「リツイート」「リプライ」など，Twitter で趣味関係なく広く使われる言葉や，特有単語の有無に関わらずツイートに出現する単語となる．特有単語を含む正例データと，特有単語を含まない約 220 万のツイート集合との間で，出現した単語全てについてカイ 2 乗検定を行う．ここで有意水準 5 %を下回る単語以外を全て除去する．そのほか,「日」「月」などの 1 文字の単語,「12」「www」などの英数字についても除去している．

3.4 深層学習による趣味情報抽出ネットワークの学習

本研究では，推定対象ユーザが指定の趣味に興味を持っているかを判断するために，深層学習による趣味情報抽出ネットワークを構築，これを元にツイート集合を分類する．

[2]https://developer.twitter.com/en/docs

表 1: 学習時の DNN のパラメータ

中間層ノード数	100
中間層数	3
ミニバッチサイズ	256
エポック数	50
L1 正則化	係数：0.01
活性化関数	中間層：ReLU 出力層：SoftMax
最適化アルゴリズム	Adam
学習率	0.1

深層学習とは，機械学習の一種で入力と出力の関係をパターンとして学習したネットワークを構築する手法である．構築されたネットワークにより，人間に与えられない細かな分類規則によって新たなデータを分類することができる．

本研究では，深層学習ライブラリの 1 つである Deep Learning for Java(DL4J)[3]を利用して， Deep Neural Network を扱う．

3.4.1 趣味情報抽出ネットワークの構築

指定した各趣味に対し，正例，負例の入力データを元に深層学習を行い趣味情報抽出ネットワークを構築する．入力として BoW を用いることから，入力層ノード数は学習用データ中に出現する名詞の総数となるため，抽出対象となる趣味ごとに入力層ノード数が異なる．その他の深層学習のパラメータは，全て以下のもので統一する．

今回，深層学習を行う上で，趣味情報抽出の結果が後述の評価において最も高い精度となったパラメータを採用した．

3.5 趣味情報抽出ネットワークを用いた Twitter ユーザの趣味判定

構築した趣味推定ネットワークを利用して対象ユーザのツイートから趣味情報抽出を行う．趣味情報抽出を行う際には，推定対象の Twitter ユーザが投稿した複数のツイートの集合を対象とし，1 ツイートを 1 データとして，文章中に趣味が現れているかどうかを 2 値で判定する．次に，趣味が強く現れていると正判定されたツイートが，全体のツイート集合に対しどの程度の割合で現れたかを算出してグラフ出力する．抽出結果を提示するための可視化インターフェースとして，統合開発環境である TETDM[4] を利用する．

[3]https://deeplearning4j.org

この時，最終的にユーザが指定の趣味に興味を持っているかどうかは，あらかじめ定めた閾値を超えているかどうかで判断する．指定する趣味の範囲や，どれだけ一般的かにより，文章からの抽出されやすさが異なるため，閾値は一般的な Twitter ユーザの判定結果よりも正判定ツイートの割合が少し高くなるように設定する．ランダムに収集した 15435 ユーザに対し上記の趣味判定を行い，各ユーザの判定結果から正判定ツイート割合の偏差値を算出し，偏差値 60 となる割合を閾値とする．

3.6 趣味情報抽出システムの使用例

提案システムを用いたユーザ推薦の例を挙げる．

システム利用者として，他のユーザとの交流の輪を広げたいユーザを設定する．この利用者は自分の趣味である「将棋」と「筋トレ」のどちらにも興味のあるユーザを捜索することを目的に，この 2 つの趣味を抽出対象として提案システムに設定，無作為に収集した候補ユーザ 19 名に対し趣味情報抽出を行う場面を考える．その結果を図 2 に示す．

システムにはユーザ@A〜S という 19 名の候補ユーザそれぞれのツイートが 100 件ずつ入力されており，各ユーザの判定結果が出力されている．ここで，右のパネルを確認すると将棋に興味があると判定されているユーザは 3 人，筋トレに興味があると判定されているユーザは 3 人存在している．その中で，ユーザ@R というユーザは将棋，筋トレのいずれにも興味があると判定されており，条件に当てはまるユーザとして水色のハイライトが表示されている．よって R さんを選択し中央と左のパネルにより詳細を確認する．中央のパネルにより詳細を確認すると，将棋，筋トレについて正判定されたツイートが複数存在しており，実際に将棋，筋トレについてのツイートを投稿していることがわかる．左のパネルにより詳細を確認すると，将棋，筋トレともに正判定ツイート割合が偏差値 60 を上回っている．さらに将棋は 132.57 と，Twitter ユーザの中でも特に強く興味を持っていることがわかる．よって，Twitter 上から収集した無作為な候補ユーザ群より，利用者が設定した複数の条件に当てはまるユーザをシステムにより発見することができた．

4 趣味情報抽出システムの有効性の評価

3 章で述べた提案手法の有効性を評価するため，収集したテストユーザに対して複数の趣味に対し構築した趣味情報抽出ネットワークを構築，実際に趣味情報

表 2: 評価用の各趣味における特有単語とテストユーザ数

趣味名	特有単語	正ラベルユーザ	負ラベルユーザ
ポケモン	メタモン	25	175
コーヒー	タリーズ	25	175
将棋	棒銀	25	175
ゴルフ	キャロウェイ	25	175
乃木坂 46	まいやん	25	175
ラグビー	ノーサイド	25	175
筋トレ	腕立て	25	175
プログラミング	CSS	25	175

図 2: 趣味情報抽出システムの使用例

表 3: 評価用の各趣味における学習用データ数と入力層ノード数

趣味名	学習用データ数	入力層ノード数
ポケモン	18528(正例 9264)	695
コーヒー	11528(正例 5764)	441
将棋	470(正例 235)	18
ゴルフ	284(正例 142)	13
乃木坂 46	17616(正例 8808)	542
ラグビー	9430(正例 4715)	313
筋トレ	10932(正例 5466)	369
プログラミング	11098(正例 5549)	493

抽出を行い，各ネットワークの趣味情報抽出精度を評価した．また，本システムを Twitter 検索機能と比較するために，Twitter 検索で探したユーザに対する趣味情報抽出の評価を行った．

4.1 趣味情報抽出精度の評価

提案システムを評価するため，抽出対象の趣味を複数設定し，各趣味に対応して構築された趣味抽出ネットワークの抽出精度をテストユーザを用いて精度指標を示す．今回の評価では，3 章で述べた趣味の定義に従い,「ポケモン」「コーヒー」「将棋」「ゴルフ」「乃木坂 46」「ラグビー」「筋トレ」「プログラミング」と 8 つの趣味を設定した．

設定した 8 つの趣味それぞれに対し，TF-IDF を活用して特有単語を筆者自身が設定，これを元に学習用データを収集して趣味抽出ネットワークの構築を行った．趣味ごとに設定した特有単語と収集したテストユーザの数を表 2 に示す．各趣味に対しプロフィールやツイートの情報から趣味を持っていると思われるユーザを収集し，これを正ラベルユーザとして設定する．一方で趣味について興味を持っていないユーザとして負ラベルユーザを用意することにより，正負を正しく判定できるかを評価する．今回は負ラベルテストユーザとして,「他 7 つの趣味の正ラベルテストユーザ」を用いる．これらのユーザは対象の趣味に興味を持っていないことを人手で確認している．各テストユーザについて最大 100 件の最新ツイートを収集し，これをシステムに入力するツイート集合とする．学習時に使用した全単語数，すなわち入力層ノード数と使用した学習用データ数は表 3 に示す．この他の学習時のパラメータは全ての趣味で 3 章で述べた表 1 の値を用いている．

各テストユーザを提案システムに入力して趣味抽出を行い，正判定に対する Precision，Recall，F 値により各趣味に対応する趣味抽出ネットワークの抽出精度を示す．ここで，本章において正判定とは「対象の趣味についての対象ユーザの正判定ツイート割合が閾値

である偏差値 60 を超えているという判定」，負判定は反対に「正判定ツイート割合が閾値を下回ったという判定」と定義する．また，正判定に対する Precision はシステムが正判定したユーザのうちの正ラベルユーザの割合，Recall は正ラベルユーザのうちの実際にシステムが正判定できたユーザの割合，F 値は Precision と Recall の調和平均となる．正判定に対するこれらの精度指標は式 2，式 3，式 4 により表すことができる．この時，TP は正ラベルユーザを正判定できた数とされる真陽性ユーザの数，FP は負ラベルユーザを誤って正判定してしまった数とされる偽陽性ユーザの数，FN は正ラベルユーザを誤って負判定してしまった数とされる偽陰性ユーザの数となる．

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

$$F\ 値 = \frac{2Precision * Recall}{Precision + Recall} \tag{4}$$

表 4: 評価用の各趣味における趣味抽出ネットワークの評価結果

趣味	Precision	Recall	F 値
ポケモン	0.706	0.960	0.814
コーヒー	0.821	0.920	0.868
将棋	0.690	0.800	0.741
ゴルフ	0.821	0.920	0.868
乃木坂 46	0.864	0.760	0.809
ラグビー	0.686	0.960	0.800
筋トレ	0.545	1.000	0.704
プログラミング	0.455	0.800	0.580

評価用の各趣味に対する趣味抽出ネットワークの評価結果を表 4 に示す．

各趣味について用意したテストユーザの正判定に対する Precision，Recall，F 値を確認すると，全ての趣味で Recall が 0.7 から 0.9 と高い水準を見せていることに対し，Precision は 0.7 から 0.8 と高い数値を見せているものから，0.3 から 0.6 とあまり高いとは言えない結果となった趣味が存在している．高い Recall を見せていることからもわかる通り，各趣味は正ラベルユーザの多くを正しく正判定することができている．その一方でほぼ同数以上の負ラベルユーザを正判定してしまい，偽陽性が多く出現している趣味があることがわかる．

提案システムでは，DNN により趣味ごとに判定に関わる重要な単語や，単語の組み合わせを学習し，未知ツイートが入力された際は学習された単語やその組み合わせが出現しているか否かにより正判定，負判定を行う．そのため，学習時にどの単語が判定に重要であると学習されたのかを理解することで，結果考察の助けとする．安藤らの研究 [5] では，BoW を用いたテキストベースの DNN において分類パターン上での重要度の高い単語，すなわち重要単語の解釈を行える枠組みを開発している．これを用いて各趣味における趣味抽出ネットワークの正判定に対する重要単語を出力し，各趣味の上位 20 単語を表 5 に示す．

表 5 には評価に使用した 8 つの趣味の重要単語がそれぞれ示されており，各趣味に強く関連する単語が上位 20 単語として出現している．

これらの中にも，ポケモンの「個体」，コーヒーの「当方」，将棋の「相手」など，表中でアンダーラインを引いている単語は一般的で広く使われると考えられ，現在の提案システムでも学習データから一般語を除去しきることはできていないことがわかる．これら一般語が学習時に残っていることが偽陽性が発生する要因の 1 つと考えられる．

ラグビーの重要単語として,「紅白歌合戦」や「きのこ」「たけのこ」など，一見して関係がなさそうな単語が含まれている．これらの単語は，ラグビーを題材としたドラマの主題歌が紅白歌合戦で歌われたことや，Twitter にて「#きのこたけのこ国民総選挙ノーサイド」というハッシュタグが流行していた時期に正例ラベルデータを収集したことが関係していると考えられる．また，プログラミングの Precision が他趣味に比べ低い結果となっている．誤って正判定としてしまった負ラベルユーザを確認したところ，プログラミングでは「初心者」「ルール」「レッスン」「筋トレ」などの多趣味に関連する単語が重要単語として学習されていた．そのため，これらを含む将棋，ゴルフ，筋トレの正ラベルユーザを誤って正判定してしまっていた．提案システム単体では趣味抽出の学習の解釈はできず，システム利用者はどのような基準で正判定が行われたかを理解することが難しい．そのため，安藤らのシステムと連携して判定の際にツイートに含まれていた表 5 の重要単語を抽出し，可視化インタフェースで出力することにより解釈の助けとすることが考えられる．

4.2 Twitter 検索により探索したユーザに対する趣味情報抽出の評価

本システムによる趣味情報抽出を用いることで指定趣味に興味を持つユーザ推薦を有効に行えるかを示すため，Twitter 検索によるユーザ探索を比較とした評価を行った．

用意したテストユーザを用いた評価では，一般語や他

表 5: 評価用の各趣味における趣味抽出ネットワークの重要単語上位 20 語

ポケモン	コーヒー	将棋	ゴルフ
海外	コーヒー	相手	ゴルフ
色違い	ドトール	将棋	アイアン
ポケモン剣盾	ラテ	矢倉	メンズ
ポケモン	ロイヤルミルクティー	斜め	ベスト
オシャボ	付近	嬉野流	マーベリック
個体値	チョコ	振り	クラブ
レイド	ミルクティー	戦法	ドライバー
個体	当方	四間	セット
国産	缶コーヒー	腰掛け	モデル
ザシアン	タピオカ	将棋ウォーズ	ボール
特性	ジャーポテドッグ	三間	シャフト
マスボ	募集中	動画	矢野
図鑑	珈琲	石玉	ゴルフシューズ
性格	ドリンク	定跡	
外国産	スノーマンラテ	後手	
臆病	ベア	原始	
ココガラ	抹茶	棋譜	
火炎	ハニーミルクラテ	先生	
王冠	真斗		
ミュウ	ホットドッグ		

乃木坂 46	ラグビー	筋トレ	プログラミング
レーン	ユーミン	腹筋	コース
乃木坂	紅白	筋トレ	サイト
シンクロニシティ	松任谷由実	プランク	コーディング
飛鳥	ラグビー	背筋	エンジニア
紅白	きのこ	セット	積み上げ
バスラ	紅白歌合戦	筋肉	初級
白石麻衣	ゲーム	回数	中級
真夏	たけのこ	トレ	レベル
乃木	きのこの山	体幹	フレームワーク
与田	麗美	体重	案件
みなみ	たけのこの里	積み上げ	基礎
ななみん	党首	逆立ち	アニメーション
白石	キャンペーン	キロ	要素
桃子	戦い	ブリッジ	ドットインストール
まちゅ	ラグビー日本代表	クランチ	初心者
ガルル	米津玄師	アブローラー	プロパティ
生駒	日本代表	パカ	レスポンシブ
募集中	名曲	大胸筋	上級
迷惑メール	大晦日	ランニング	ウェブカツ
秋元真夏	総選挙	プッシュアップバー	言語

趣味関連語を除去できていないことにより，Precision が低い結果となってしまった趣味がある．そのうちの 1 つであるコーヒーと，高い Precision を見せたポケモンについて，実際にシステムを活用する際に想定する方法で収集した候補ユーザに対して趣味情報抽出を正しく行えるのかを確かめた．

趣味情報抽出を行う際，提案システムでは抽出対象として対象の趣味名を含むツイートを投稿したユーザを収集し，このうちから趣味に興味を持つユーザを抽出することを想定している．よって「ポケモン」「コーヒー」をそれぞれ含むツイートを投稿したユーザを収集し，正ラベルユーザ，負ラベルユーザに分類したものをテストユーザとした．趣味名で収集したユーザはポケモンはうち 7 割，コーヒーはうち 3 割が正ラベルとなった．ポケモンなど娯楽となりやすい趣味よりも，コーヒーなどのより生活習慣に近い趣味は多くのユーザの話題として出現しやすいためであると考えられる．

趣味情報抽出の結果として正判定に対する Precision は，ポケモンが 0.886，コーヒーが 0.667 となった．ポケモンは高い Precision をみせたが，コーヒーは一般語や他趣味関連語による誤判定が多く見られ，正判定のうち 3 割程度は誤判定となってしまった．一方で，趣味情報抽出の結果の Precision は，趣味名を元に検索し収集したユーザのうちの正ラベルユーザの割合を上回っているため，Twitter ワードサーチによる単純なユーザ探索と比較して，そのうちから提案システムによりユーザを絞り込むことで趣味に興味を持つユーザをより高い割合で見つけることができると考えられる．

表 6: 趣味名を含むツイートを元に収集したテストユーザ数

趣味名	テストユーザ総数	正ラベルユーザ	負ラベルユーザ
ポケモン	145	105	40
コーヒー	146	51	95

表 7: 趣味名を含むツイートを投稿したテストユーザの趣味情報抽出評価結果

ポケモン			
		正解ラベル	
		正	負
予測結果	正	70	9
	負	35	31
Precision		0.886	
Recall		0.667	
F 値		0.760	

コーヒー			
		正解ラベル	
		正	負
予測結果	正	10	5
	負	41	90
Precision		0.667	
Recall		0.200	
F 値		0.300	

以上の結果より，提案システムでは多様な単語表現を学習することによって高い Recall を示せたことにより対象の趣味に興味のあるユーザを網羅して抽出することができているが，一方で一般的な単語や他趣味に関連する単語を学習してしまっているため，偽陽性ユーザが発生してしまうことがわかった．趣味名を含むツイートを行うユーザを収集し，これを元に趣味情報抽出を行った結果，Twitter ワードサーチと比較して交流ユーザ探索により有用であることを示せたが，入力する候補ユーザ群によっては多すぎる偽陽性が出現してしまうと使用例として想定するユーザ推薦にも支障をきたしてしまうことが予想される．そのため，偽陽性の原因となる一般的な単語や他の趣味に関連する語などを TF-IDF 等を利用して抽出，除去し，抽出精度を上げていくことが今後の課題となる．

5 おわりに

本研究では，Twitter ユーザを対象として趣味抽出を行い，システム利用者が設定した任意の趣味についての抽出結果を提示することを目的として行った．任意の趣味についての学習用データを Twitter より収集し，テキストベースの深層学習を用いて抽出対象ユーザの投稿文の集合より対象の趣味に興味を持っているかどうかを判定し，結果を提示するシステムを開発した．

提案システムの趣味抽出精度の評価として，設定した 8 つの趣味について用意したテストユーザに対する Precision，Recall を求めた．その結果，現在のシステムでは高い Recall を示せたことにより対象の趣味に興味のあるユーザを網羅して抽出することができ，Twitter ワードサーチを用いた単純なユーザ探索に比べより有用であることを示せたが，一方で一般的な単語や他趣味に関連する単語を学習していることにより偽陽性が発生し Precision が低下することがわかった．

今後の課題として，これら判定に悪影響を及ぼす単語を除去し，抽出精度を上げていくことを目標としていきたい．

参考文献

[1] 馬繰美穂, 徳久良子, 寺嶌 立太: ユーザの嗜好と所有物の関係性を用いた属性分析研究報告情報基礎とアクセス技術 2014-IFAT-114,pp.1-6 (2014)

[2] 那須川哲也, 西山莉紗, 金山博, 吉田一星, 大野正樹: 一人称所有格を用いたプロフィール推定, 言語処理学会第 19 回年次大会発表論文集,pp.952-955 (2013)

[3] 上里 和也, 浅井 洋樹, 山名 早人:Personalized PageRank を利用した網羅的 Twitter ユーザ属性推定, DEIM 2016 第 8 回データ工学と情報マネジメントに関するフォーラム D2-2 (2016)

[4] 砂山渡, 高間康史, 徳永秀和, 串間宗夫, 西村和則, 松下光範, 北村侑也: 統合環境 TETDM を 用いた社会実践, 人工知能学会論文誌 32 巻 1 号, pp.NFC-A_1-12(2017)

[5] 安藤 雅行, 河原 吉伸, 砂山 渡, 畑中 裕司:テキストベースの深層学習における分類パターンの解釈支援, 知能と情報 Vol.31, No.4, pp.779-787 (2019)