発生普及過程を捉えた未知エンティティの発見

赤崎　智† 吉永　直樹†† 豊田　正史††

† 東京大学大学院 情報理工学系研究科 〒 113–8654 東京都文京区本郷 7–3–1
†† 東京大学 生産技術研究所 〒 153–8505 東京都目黒区駒場 4–6–1
E-mail: †{akasaki,ynaga,toyoda}@tkl.iis.u-tokyo.ac.jp

あらまし テキスト中で言及される現実世界の事物（エンティティ）を知識ベース上のエントリと対応付けるエンティティリンキングは，ソーシャルメディア等のテキストストリームからの情報獲得に有用な技術であるが，未知のエンティティを適切に取り扱うことは困難であった．ここで未知エンティティとはリンク先の知識ベースに項目が存在しない未知のエンティティを指し，もともと存在していたが広く知られていないため未登録であるか，真に新しく世に現れたため未登録であるかのどちらかである．特に新しく現れたエンティティは，その新規性からエンティティに関する情報を把握することが重要となるため，早期に検出し知識ベースへの速やかな登録ができることが好ましい．本稿では未知エンティティが初期に出現し普及していく段階では，それと区別できる特徴的な表現が用いられることに着目し，これを利用して未知エンティティを発見する手法を提案する．実験では我々の研究室で収集している Twitter 投稿データを用い，ある時点の Wikipedia の見出し語一覧を知識ベースとし，それ以降に現れたエンティティを未知のエンティティとしその発見を試みる．

キーワード 自然言語処理，エンティティリンキング，マイクロブログ，知識ベース

1. はじめに

モバイル端末などの普及により Web に大量の情報が溢れるようになって久しい．特に， Twitter などのテキストストリームには日々大量のテキストが発信されているため，ユーザーはその中から自らに有益な情報を取捨選択する必要がある．このとき，テキストの中から実世界の事物（以下，エンティティ）に関する情報を認識することができれば，ユーザーの関心のある事物のみの情報を収集することが可能である．このような情報の獲得を支援するための技術として，テキスト中に存在する実世界上の事物に関しての言及（以降，メンション）を既存の知識ベースのエントリ（エンティティ）に結びつけるエンティティリンキングと呼ばれるタスクが研究されている [1] [2].

エンティティリンキングを行う際には結びつけたいメンションに対応するエンティティが知識ベースに存在することが大前提となるが，現実世界では新しい芸術作品や，組織，人等が次々に現れるため，それらのエンティティを常に網羅することは困難を極める．現状，これらのエンティティは人手により知識ベースに登録されているが人間が把握可能な範囲には限界がある．そのため，知識ベースに登録されていないエンティティ（以降，未知エンティティと呼ぶ）をテキスト中のメンションから自動的に発見し，できるだけ早く登録することが望まれる．

本研究ではテキスト中のメンションから未知エンティティを発見するタスクに取り組む．未知エンティティを発見するにあたり，テキスト中からメンションにあたる部分を切り出す必要があるが，これに対し我々は，未知エンティティを指すメンションがエンティティが現れてすぐの段階ではそれと分かるよう強調して"「」"などの括弧記号で囲まれやすいという観察にもとづき，手がかり表現を利用して未知エンティティの候補となるメンションを切り出す．切り出されたメンションが未知エンティティか否かを判定するには 2 値分類器を用いる．我々は未知エンティティが出現し普及していく過程では，その新規性を示唆するような特徴的な表現が同時に出現することに着目し，これらを分類器の特徴量とする．

実験では，ある時点での Wikipedia の見出し語一覧を知識ベースとみなし，それ以降に登録された見出し語を未知エンティティとみなしてそれらをテキスト中から発見できるか確認する．我々の研究室が所有する約 5 年分の Twitter 投稿を利用して，実際に投稿中から未知エンティティを発見しその手がかりの有効性を示す．

2. 関連研究

エンティティリンキングの研究はこれまで盛んに行われているが，未知エンティティを対象とするものは極めて少ない．以下では本研究と同じ Twitter 投稿を利用したエンティティリンキングの研究と，未知エンティティを対象とするようなエンティティリンキングの研究について紹介する．

Liu [3] らは，1 投稿辺りの情報量が少なく，メンションの表記ゆれも多い Twitter を対象とするエンティティリンキング手法を最初に提案した．Liu らは Twitter の投稿量の多さを利用し，特定のエンティティを指す単語表現として類似しているもの集め，通常のメンションに加え表記ゆれなどのメンションについてもエンティティリンキングを試みた．Liu らの手法では未知エンティティそのものについて扱っておらず，またテキスト中でメンションとなる表記（表記ゆれのメンションも含む）が予め与えられていることが本研究と異なる．

Wu ら [4] は，これまでのエンティティリンキングの手法では適切に取り扱われなかった未知エンティティに注目し，テキスト中のメンションが未知エンティティか否かを判定するモデルを提案した．Wu らは，メンションの周辺単語，単語分散表現，トピック，検索エンジンのクエリ，統語的特徴といった様々な特徴量を用いることで未知エンティティのモデリングを行った，Wu らの手法においてもメンションとなる表記が予め与えられている．

Lin ら [5] はテキスト中から未知エンティティを発見し，それらに「人名」,「飲み物」等のタイプ付けを行う手法を提案した．Lin らの手法はまず過去から現在に至るまでの Wikipedia に登録されていないメンションが出現するテキストを集め，それらのメンションの出現頻度を時期ごとに集計する．その後 Lin らは，メンションの集計された頻度から時系列に沿って回帰直線を引いたときの傾き度合いが急なものは，人々の注目が集まったメンションだと仮定し，それらを未知エンティティと判定すると同時に，出現する文脈からエンティティのタイプも推定した．Lin らの手法はメンションの出現頻度をある一定以上の期間蓄積する必要があるため，未知エンティティの早期の発見は難しい．

3. 提 案 手 法

本研究ではエンティティの普及過程における手がかり表現を用いて Twitter の投稿から未知エンティティを発見する．本節ではまず，テキスト中から未知エンティティの候補となるメンションを抽出する手法について説明する．その後，本研究で利用する普及過程の手がかり表現について述べ，実験に用いるデータセットの収集方法と，未知エンティティの発見に用いる分類器について述べる．

3.1 メンションの抽出

テキスト中から未知エンティティを発見する際には，まずテキスト中のどの部分がメンションであるかを認識する必要がある．このとき，言語環境として英語などを対象としたエンティティリンキングでは，テキスト中のどこがメンションであるかを問題としない場合が多い．これは英語がスペース区切りでかつ，メンションとなりやすい固有名詞などの頭文字が大文字となるためである．したがって，そのような単語に対して未知エンティティであるかどうかを判定するだけで良い場合が多い．しかし，日本語や中国語などの分かち書きが必要でかつ大半の固有名詞の頭文字が大文字 (注1) とならないような言語では当然この手法を適用することは出来ない．

テキスト中からメンションを切り出すタスクは一般的には固有表現抽出によって解かれる [6]．しかし，固有表現抽出は切り出すメンションのカテゴリが場所や団体等に制限されておりかつ，マイクロブログ等のくだけた表現に対しての精度は高いとはいい難い [7] [8]．また，未知エンティティは学習するコーパス等にも一般には出現しないため，メンションの切り出しの難しさにより拍車をかけている．

(注1)："orange juice"などの頭文字が大文字で表記されないものも存在する．

そこで我々は日本語において，未知エンティティが発生し周知される過程において"「」" や "『』" などの括弧記号で囲まれてしばしば言及されることを利用して，括弧記号に囲まれた部分をメンションとみなす．この手法の妥当性を検証するため，2014/12/31 と 2016/12/20 の Wikipedia の見出し語の差分に含まれかつ Twitter に 100 回以上の投稿がある未知エンティティ (注2) の 3,191 件について括弧記号に囲われているメンションが（それらを含む Twitter 投稿中に）存在するかどうかを確認し，うち 82%となる 2,643 件にそのような表現が認められることが分かった．このことから，括弧記号に注目することは有効であると考えられる．更に，分かち書きの際の過分割なども起きないため，括弧記号に囲われているものに限り正確なメンション抽出が可能である．このようにして得られた各メンションに対し，未知エンティティか否かの 2 値分類問題を解くことで未知エンティティを発見することが可能となる．

3.2 普及過程の初期の手がかり表現

知識ベースに未知のエンティティとして新たに登録されるエンティティは，マイクロブログなどでそのエンティティが普及する過程においての初期段階を示唆するような単語が同時に出現することが多い．例えば，以下は「君の名は。」という映画作品に対するメンションが現れる複数の Twitter の投稿であるが「新作」,「決定」,「予告編」などといった単語が同時に出現している．

> ・新海誠の新作『君の名は。』神木隆之介らキャストほか、RADWIMPS の主題歌を収めた予告公開 https://t.co/i6SDgpyMv5
> ・新海誠の最新作「君の名は。」ワールドプレミア開催決定！ 声優を務める神木隆之介からも喜びのコメント到着 | ダ・ヴィンチニュース https://t.co/HOdNtIf3F5 @d_davinci さんから
> ・ヤバす。新海さん久々の新作！ 沖縄でも上映しないかのぉ (*´ω｀*)『君の名は。』(2016 年 8 月 26 日公開予定) 予告編を配信中 【解説】『星を追う... https://t.co/jpctgwn3M4

これらの単語の多くはエンティティのカテゴリ間に差異はあるものの，"作品名 + 決定"，"新作 + 発表"といった共通の組み合わせで出現することが多い．また，普及過程の初期段階に出現するため，これらの単語を Bag of Words や n-gram などの分類器の特徴量として用い未知エンティティの発見を試みる．

3.3 未知エンティティの発見

3.1 節で述べた通り，我々は教師ありの 2 値分類器を用い，テキスト中の括弧記号に囲まれたメンション候補 (注3) に対してそれが未知エンティティか否かの判定を行う．分類器に用いる特徴量を以下に箇条書きで列挙する．

共起 n-gram　3.2 節で説明したマイクロブログ投稿からの手がかり表現の特徴量として単語 1-グラムと 2-グラムを用いる．

(注2)：多義な未知エンティティは除いている．
(注3)：括弧記号に囲まれるものはメンション以外の文字列（台詞など）も存在するため，以降では括弧記号に囲まれたものをメンション候補と統一する．

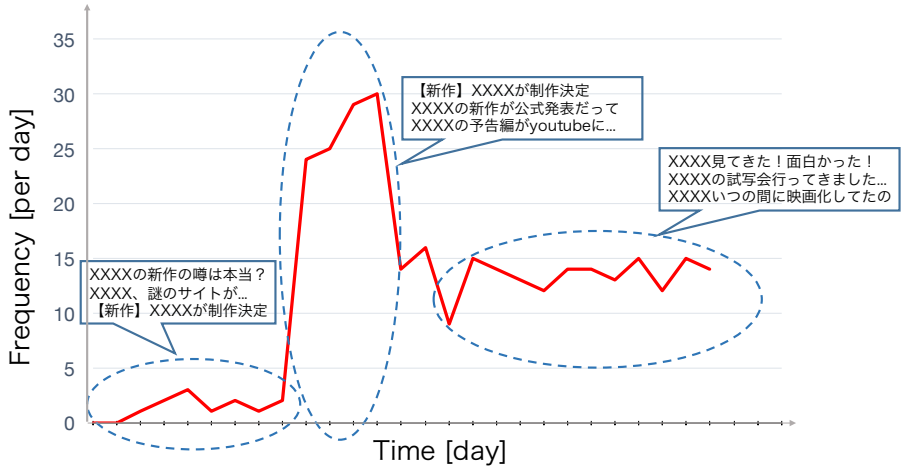

図 1　エンティティ"XXXX"が出現する Twitter 投稿の頻度をプロットした図．点線部分は左が普及段階，右が普及後の投稿

ここでは，メンション候補と共起する単語すべてを 1-グラム，2-グラムとして用い，基本特徴量とする．

前後 *n*-gram　共起 *n*-gram に加え，メンション候補の直前，直後の位置の単語 1-グラム，2-グラムを用いる．ここで，直前と直後の単語については区別するようにする．こうすることで，メンション候補の直前直後に現れやすい単語を捉えることができる．

length　対象としているメンション候補の文字列長を特徴量として用いる．これにより，文字列長が極端に長い/短いメンション候補を区別する効果があると期待できる．

URL　メンション候補が出現する投稿に URL 文字列が出現するか否かを特徴量として用いる．これはユーザーが未知エンティティについて言及するとき，同時にニュースサイト等の URL を参照することが多いことを捉える効果がある．

3.4　学習・評価用データセットの構築

本項では分類器の学習及び評価に用いるデータセットの構築方法について述べる．本研究ではデータセット構築の都合上，曖昧性のあるエンティティ（同じ表層単語で複数のエンティティが存在する）については扱わない．まず正例については以下の手順で収集を行う．

（1）　異なるタイムスタンプの Wikipedia の見出し語を比較し，その差分から未知エンティティを取得する．具体的にはまず時間 t においての Wikipedia 見出し語と時間 $t^+(t < t^+)$ においての Wikipedia 見出し語との差分を取得する．この操作で残った見出し語 $(E_{t\sim t^+})$ は，時間 t から時間 t^+ の間に Wikipedia に登録されたエンティティとみなせるため，時間 t においては未知のエンティティということになる．(注4)

（2）　次に，得られた未知エンティティそれぞれにつき，そのエンティティが出現する最古のものから順に Twitter 投稿を抽出する．これにより未知エンティティの投稿についての発生の投稿を得ることができ，それらの投稿は普及過程の初期の手がかり表現を含むはずである．そこで，各未知エンティティ投稿の時系列において図 1 のように 1 日に 5 件以上の投稿が出現する最初の区間（図 1 の左点線部）までの投稿を抽出する．こうして集まった用例は初期の手がかり表現と未知エンティティを投稿中に含むため，そのまま正例として用いることが可能である．

負例用例については以下の手順で収集を行う．

（1）　正例を集めた期間 $t\sim t^+$ から，"「」"及び"『』"に囲まれ，かつ $E_{t\sim t^+}$ に含まれないメンション候補の頻度を集計し，その上位 10,000 件 (注5) を抽出する．

（2）　次に，抽出したメンション候補それぞれにつき，$t\sim t^+$ の期間でそのメンション候補が出現する最新のものから順に 100 件ずつ Twitter 投稿を抽出する．このような操作を行うのは，負例として普及後のエンティティに関する投稿のみを可能な限り集めるためである．

以上の手順で集まった用例から 3.1 節で述べたような括弧記号で囲まれたメンション候補が出現する投稿のみを抽出し，学習データとして用いる．データセットの具体的な内容については 4.1 節のデータセットで述べる．

4.　評 価 実 験

本節では，評価に用いるデータセットと分類器について詳細に述べ，それらを用いて実際に未知エンティティの発見を試みた結果について述べる．

4.1　データセット

実験に用いるデータセット構築のため，日本語 Wikipedia のダンプデータ (注6) と，Twitter の投稿データ (注7) を使用した．

我々は，データセットの未知エンティティと正例を収集する

(注4)：この際，同じ表層単語で複数のエンティティが存在するものは，抽出の際に既に普及しているエンティティでの用例が集まってしまう．そのため，予め時間 t 以前の期間で 5 回以上の頻度で現れているものは曖昧性のあるエンティティとし，対象から除く．

(注5)：取り出したメンション候補 10,000 件のうち頻度最大のものは 136,502 件で，最小のものは 598 件であった．

(注6)：https://dumps.wikimedia.org/jawiki/

(注7)：喜連川・豊田・吉永研究室は 2011 年から 2017 年 1 月にかけての期間に収集した約 380 億のツイートを所有している

ため，時間 t として 2014/12/31 ，時間 t^+ として 2016/12/20 をとり，それぞれの時刻の Wikipedia の見出し語一覧である all-titles-in-ns0 データを使用し，3.4 節の手順に沿って見出し語の差分を抽出した．all-titles-in-ns0 はリダイレクトページの見出し語も含むため，それらについては予め除去した．得られた Wikipedia の見出し語を用いて，時間 t から時間 t^+ 間の Twitter 投稿で実験に用いるデータを収集した．その結果，正例として $t \sim t^+$ 間で計 2,240 件の未知エンティティと，手がかり表現を含む 55,784 用例を得た．負例については，複数の括弧記号が出現する正例から派生した 19,455 件と，3.3 節で集めた負例から，正例の数と均等になるようにエンティティ単位でランダムサンプリングした 36,224 件を合わせ，計 55,679 件の負例を得た．

上記の操作により得られた 111,463 件の用例中の"@"から始まるユーザー名を表す文字列を"USER"に，URL を表す文字列を"URL"に置き換えた後，辞書に ipadic を用いた MeCab(注8) により分かち書きを行った．その後，セット間でエンティティが被らないように 8:1:1 に分け，それぞれ訓練データ，開発データ，評価データとした．

4.2 未知エンティティの発見

本項では，未知エンティティの発見に用いる分類器について詳細に述べる．ベースラインとして，全ての用例を正例と判定するモデル (ベースライン 1) と，投稿中に URL 文字列が存在すれば正例と判定するモデル (ベースライン 2) の 2 つを用いた．

2 値分類器としては，Passive Agressive アルゴリズム I (PA-I) [9] [10] と Kim らの Convolutional Neural Network (CNN) を用いた分類を行うモデル [11] の 2 つを用いた．PA-I の分類器の特徴量は 3.4 節で述べたものを用い，線形カーネルを適用し評価を行う．CNN については 3.4 節の特徴量は用いない．パラメータについては窓幅{1,2,3}のフィルタをそれぞれ 100 個ずつ使用し，3 層からなる全結合層に確率 0.5 でランダムにニューロンを停止する Dropout を適用する．パラメータの学習のため Adam($\alpha = 0.001, \beta 1 = 0.9, \beta 2 = 0.999, \epsilon = 10^{-8}$) を用いて確率的勾配降下法を行う [12]．PA-I，CNN の実装としてはそれぞれ opal [10](注9)， Chainer [13](注10) を用い，いずれも開発データを用いてパラメータチューニングを行った．その後，訓練データを用いて学習を行い，評価データを用いてテストを行った．

4.3 実験結果と考察

表 1 に，各手法を用いて用例のメンションが未知エンティティかどうかの判定を行ったときの分類精度と，正例についての適合率，再現率，F 値を示した．表 1 より，提案手法である PA-I，CNN ともに概ね 0.72 から 0.74 という 分類精度で未知エンティティの発見ができていることがわかる．最も F 値が高いのは PA-I で，0.741 という値を示している．ただし，これらの値はエンティティ単位ではなく用例単位で計算を行っていることに注意されたい．つまり，あるエンティティについて複数の用例があったとき，それらを全て区別して値を算出している．

表 1　評価データでの分類結果

手法	分類精度	適合率	再現率	F 値
ベースライン 1	0.497	0.497	1.000	0.664
ベースライン 2	0.595	0.560	0.871	0.681
CNN	0.723	0.731	0.746	0.739
PA-I	0.746	0.752	0.731	0.741

表 2　PA-I を用いた abulation test の結果

特徴量	分類精度	適合率	再現率	F 値
全ての特徴量	0.746	0.752	0.731	0.741
–共起 n-gram	**0.793**	**0.801**	**0.776**	**0.788**
–前後 n-gram	0.742	0.746	0.727	0.737
–length	0.767	0.779	0.740	0.759
–URL	0.746	0.751	0.732	0.741

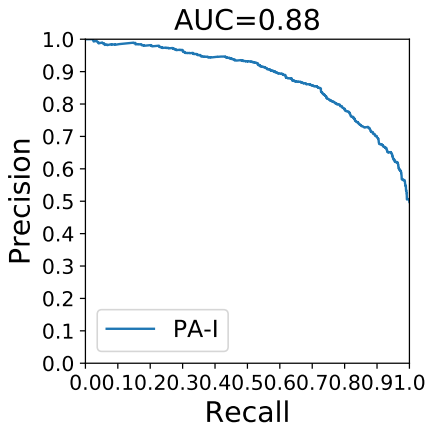

図 2　PA-I(–共起 n-gram) での適合率-再現率曲線

ここで，分類器に用いた特徴量でどれが最も性能に寄与しているかを確認するため，PA-I を用いて Ablation test を行った．表 2 は 3.3 節の特徴量をそれぞれ取り除いて評価を行ったものである．表 2 より，最も性能に寄与しているのは前後 n-gram だということが確認できる．これは，「... 新作映画『君の名は。』来夏公開...」のように，メンション候補の前後に手がかりとなる表現が出現しやすいため，これらを捉えることで判定精度に寄与していることを示唆している．逆に共起 n-gram については取り除くことで精度が向上している．これは，投稿中のメンション候補から遠い単語は判定のノイズとなっていることを示唆している．これは括弧記号が複数出現（＝メンション候補が複数存在）する用例においては特に顕著で，そのような場合はメンション候補の直前直後の単語のみを用いて判定を行う方がうまく判定できる場合が多い．length と URL については特に精度向上に寄与していないため，投稿外の情報や文の構造的な特徴量を分類器に組み込む必要がある．

図 2 は，PA-I で特徴量から共起 n-gram を除き，適合率-再現率曲線を描いたものである．これより，正例に関して適合率を 0.9 程度に保ちつつ，再現率を 0.6 まであげられることが確

(注8)：http://taku910.github.io/mecab/
(注9)：http://www.tkl.iis.u-tokyo.ac.jp/ ynaga/opal/
(注10)：http://chainer.org

表 3　PA-I (–共起 n-gram) で特徴量の重みが上位のもの．メンション候補の直前/直後の出現を単語_b/単語_a で表す．$len \geqq n$ はメンション候補の文字列長が n 以上であることを表す．

素性 top15（重み正）	素性 top15（重み負）
に_a 改名_a	主題歌_a
朝ドラ_b は_b	発表_a 「_a
に_a 変更_a	の_a 主題歌_a
棋戦_b は_b	有料_b サービス_b
名称_b は_b	$len \geqq 50$
アルバム_b は_b	ディズニー・オン・アイス_b
新作_b	ない_b ?_b
新曲_b	OP_a
タイトル_b は_b	な_a のに_a
で_a 日本人_a	フォン_b RPG_b
を_a 作ろ_a	代行_b サービス_b
と_a 命名_a	痛い_b RPG_b
タイトル_b が_b	は_b …_b
に_a 決定_a	姉ちゃん_b の_b
姉妹_b 」_b	ミュージカル_b

認できる．残りの 0.4 についてどのようなものがあるか確認したところ，投稿中に手がかり表現が存在しない例や，手がかり表現がメンション候補から遠い位置にある例，メンション候補が複数存在する例などが多く確認できた．詳細については誤り分析のところで述べる．このような例は文字列特徴量などの手がかりだけでは判定が難しいため，投稿外の情報や知識ベースの情報を利用する必要がある．

表 3 に PA-I で共起 n-gram を除いたものの特徴量の重み上位 15 件を示した．正の重みに「新曲」,「新作」,「タイトル」等の未知エンティティの普及過程の初期段階であることを示す単語があることが確認できる．負の重みには，文字列長が 50 文字以上であることを示す特徴量などが確認でき，これは括弧記号に囲まれやすいキャラクターの台詞などにペナルティを与えていると考えられる．

以下は未知エンティティを発見できた例である．太字が判定対象のメンション候補である．未知エンティティには 1 つ目のような芸術作品が多く，ニュースサイトなどの URL と同時に出現しやすい．また，2 つ目のような芸術作品以外の電化製品などの未知エンティティも発見できている．3 つ目は漫画作品の未知エンティティであるが，この例は漫画作品の作者の Twitter 投稿から発見した例であり，このような未知エンティティはマイクロブログを参照することでいち早く発見できる．

・やばい ≫ 山田孝之主演 「 勇者ヨシヒコ 」続編決定！タイトルは「**勇者ヨシヒコと導かれし七人**」 URL
・軍用規格対応「**TORQUE G02**」発表 。4.7 型 IGZO 液晶、1300 万画素カメラ搭載 URL ...
・【 宣伝 】本日 5 月 11 日発売の週刊少年サンデーより熊之股の新連載 「**魔王城でおやすみ**」が始まります! ...

最後に誤り分析として分類器が判定を誤った例についていくつか取り上げ考察する．以下は偽陰性の用例で，太字が判定対象のメンション候補である．

・「はらたく細胞」じゃなくて「**はたらく細胞**」です!笑 血液中の細胞擬人化漫画となっております URL
・!!重大発表!! 本日のライブで改名を発表「F:ma」改め...『**Fiima**』となりました!! 読み方は変わらず …

1 つ目の例は手がかり表現が全く出現していない例であり，このような例はメンション候補ごとにツイートを収集し，バッチ処理で判定することが望ましい．2 つ目の例は手がかり表現がメンション候補から遠い位置に出現する例で，このような例は離れた位置の単語の係り関係などを考慮する必要がある．

以下は偽陽性の用例である．

・任天堂、ついに「**スーパーマリオ**」完全新作を iPhone・iPad 向けにリリースへ 片手で遊んで対戦可能な『SUPER MARIO RUN』URL
・【速報】TV アニメ『ラブライブ! サンシャイン』OP 主題歌 シングル「青空 Jumping Heart」1 話挿入歌「**決めたよ Hand in Hand**」！ジャケット&CM 動画解禁! URL
・麻倉もも・雨宮天・夏川椎菜による新世代トライアングルガールズユニット「**TrySail(トライセイル)**」1st アルバム「Sail Canvas」が 5 月 25 日 (水) に発売決定! URL #BARKS 23
・コブクロ新曲「**STAGE**」が刑事ドラマの主題歌に抜擢 アルバム『TIMELESS WORLD』6 月 15 日発売決定...
・おぉ、なんか夢のようですねえ > 現実世界に 3D ホログラムを重ねて表示する「**Windows Holographic**」とヘッドマウントディスプレイ「Microsoft HoloLens」を Microsoft が発表 - GIGAZINE URL

1 つ目の例のメンション候補「スーパーマリオ」は明らかに既知のエンティティであるが，未知エンティティだと判定してしまっている．この例では同時に「SUPER MARIO RUN」という未知エンティティが出現している．この例のように複数のメンション候補が出現し，どちらか一方が未知エンティティである場合は判定が難しくなる．2 つ目の例のメンション候補「決めたよ Hand in Hand」は，時間 t 以降の Wikipedia の見出し語に存在しないが,「決めたよ Hand in Hand/ダイスキだったらダイジョウブ!」という見出し語は存在しており，該当する Wikipedia 記事も例の内容を指している．このような表記ゆれの問題は Twitter 等のマイクロブログでは頻繁に起きるため，Wikipedia の見出し語をそのまま教師データとして使うことのデメリットとなる．3 つ目の例のメンション候補「TrySail(トライセイル)」も表記ゆれの問題であり,「TrySail」という見出し語は Wikipedia に存在するが,「TrySail(トライセイル)」という見出し語は存在しないため誤って正例と判定してしまっている．4 つ目の例のメンション候補「STAGE」はここでは負例のエンティティであるが，これと類似した CD 等の芸術作品で Wikipedia に見出し語が新しく登録されるもの（つまり正例）もあるため，Wikipedia の曖昧な見出し語の登録基準によって

起きる誤りである．5 つ目の例のメンション候補，「Windows Holographic」は日本語 Wikipedia では見出し語が存在しないが，英語 Wikipedia では見出し語が存在するため，必ずしも負例とは言い切れない例である．これもまた Wikipedia の見出し語を教師データとして使うことのデメリットに挙げられる．

ここで説明したように，Wikipedia を用いてデータを構築しているがため起きる判定誤りやラベル誤りが多いため，手法の正確な評価には人手によるアノテーションが必須である．

5. まとめと今後の課題

本研究では，文中から未知エンティティを発見するタスクに取り組んだ．提案手法では，未知エンティティの初期の出現と同時にその普及過程を示す手がかり表現が現れることを利用し，教師ありモデルを用いてそれらの用例を学習し，未知エンティティの発見を行った．

今後の課題としては，今回は普及過程の初期段階を示唆する手がかり表現を n-gram で用いたが，それだけでなくメンションの係り受け関係などの構造的な素性を設計し改めて実験を行い，分類精度の改善を図る必要がある．また，今回は用例単位での評価を行ったが，実際にはエンティティ単位でどれだけ正確に発見ができるかが重要であるため，エンティティ単位での評価基準を設計し，それにより評価することも検討している．また，今回は正例と負例が均等に成るようにエンティティとメンション候補を収集しデータセットを構築したが，エンティティの頻度や種類等を実際の分布に即したものになるよう評価データを作成し，それを用いて実験することが重要である．最後に，未知エンティティの検出においては即時性が重要なため，リアルタイムで次々と情報が投稿される Twitter などのテキストストリームに対応できるようなモデルの設計も検討していく必要がある．

謝　　辞

本研究の一部は JSPS 科研費 16K16109 と 16H02905 の助成を受けたものです．

文　　献

[1] Rada Mihalcea and Andras Csomai. Wikify!: linking documents to encyclopedic knowledge. *Proceedings of the Conference on Information and Knowledge Management (CIKM)*, pp. 233–242, 2007.

[2] Razvan Bunescu and Marius Pasca. Using Encyclopedic Knowledge for Named Entity Disambiguation. *Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, No. April, pp. 3–7, 2006.

[3] Xiaohua Liu, Yitong Li, Haocheng Wu, Ming Zhou, and Yi Lu Furu Wei. Entity Linking for Tweets. *Proceedings of the Assoociation for Computational Linguistics (ACL)*, pp. 1304–1311, 2013.

[4] Zhaohui Wu, Yang Song, and C Lee Giles. Exploring Multiple Feature Spaces for Novel Entity Discovery. *Proceedings of the Conference on Artificial Intelligence (AAAI)*, pp. 3073–3079, 2016.

[5] Thomas Lin, Mausam, and Oren Etzioni. No noun phrase left behind: detecting and typing unlinkable entities. *Proceedings of the Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, No. July, pp. 893–903, 2012.

[6] D. Nadeau. A survey of named entity recognition and classification. *Linguisticae Investigationes*, No. 30, pp. 3–26., 2007.

[7] 岩倉友哉. 固有表現抽出におけるエラー分析. 言語処理学会 第 21 回年次大会 ワークショップ, 2015.

[8] 平田亜衣, 小町守. 様々なジャンルのテキストに対する固有表現認識の分析. 言語処理学会 第 21 回年次大会 ワークショップ, 2015.

[9] Koby Crammer, Ofer Dekel, Joseph Keshet, Shai Shalev-Shwartz, and Yoram Singer. Online Passive-Aggressive Algorithms. *Journal of Machine Learning Research*, Vol. 7, No. Mar, pp. 551—-585, 2006.

[10] Naoki Yoshinaga and Masaru Kitsuregawa. Kernel Slicing: Scalable Online Training with Conjunctive Features. *Proceedings of the International conference on on Computational Linguistics (COLING)*, No. August, pp. 1245–1253, 2010.

[11] Yoon Kim. Convolutional Neural Networks for Sentence Classification. *Proceedings of EMNLP*, pp. 1746–1751, 2014.

[12] Diederik P. Kingma and Jimmy Lei Ba. Adam: a Method for Stochastic Optimization. *Proceedings of the International Conference on Learning Representations (ICLR)*, 2015.

[13] Seiya Tokui, Kenta Oono, Shohei Hido, and Justin Clayton. Chainer: a Next-Generation Open Source Framework for Deep Learning. *Proceedings of the Workshop on Machine Learning Systems in NIPS*, pp. 1–6, 2015.