マイクロブログを対象にした エンティティリンキングにおける語義曖昧性解消

浦田 智昭† 前田 亮‡

†立命館大学情報理工学研究科 〒525-8577 滋賀県草津市野路東 1-1-1
‡立命館大学情報理工学部 〒525-8577 滋賀県草津市野路東 1-1-1

†is0156ss@ed.ritsumei.ac.jp, ‡amaeda@is.ritsumei.ac.jp

あらまし 近年，読者が興味のある情報を取得するために，Web 上で Twitter などのマイクロブログから記事を読む機会が増えてきている．しかし，文書中に出現した人名や組織名などの情報に読者が興味を持ち，詳しい情報を知りたいと思っても，その情報に関する情報へのリンクが存在しない場合が多く，Google などの検索エンジンで逐一調べるのは面倒である．本研究では，Twitter などの文字数が少ない文書であるマイクロブログ上で、地名・団体名など何らかの実体を指すエンティティの情報を得る際に，対象エンティティに曖昧性が存在していても，正しい記事を抽出できるエンティティリンキングの手法を提案する．評価実験として，Twitter から 50 件の語義曖昧性が存在するエンティティが含まれるツイート文を用いて提案手法とベースラインの手法による有効性を検証した結果，精度は 84%であった．

キーワード マイクロブログ，固有表現抽出，語義曖昧性解消，Wikipedia

1. はじめに

近年，読者が興味のある情報を取得するために，Web 上で Twitter などのマイクロブログから記事を読む機会が増えてきている．しかし，文書中に出現した人名や組織名などの情報に読者が興味を持ち，詳しい情報を知りたいと思っても，その情報に関する情報へのリンクが存在しない場合が多く，Google などの検索エンジンで逐一調べるのは面倒である．

本研究では，Twitter などの文字数が少ない文書であるマイクロブログ上でエンティティの情報を得る際に，対象エンティティに曖昧性が存在しても，正しい記事を抽出できるエンティティリンキングの手法を提案する．

エンティティリンキングとは，テキスト中で何らかの「実体（エンティティ）」を指示する記述を抽出し，外部のデータベースや知識ベースの項目と対応づける操作のことである．エンティティリンキングに関する研究は，情報検索や自然言語処理の分野で活発に行われている．従来の研究において，対象テキストは文書量の多い新聞記事や Web 文書が対象とされていることが多く，エンティティに曖昧性が存在したときに，複数の記事候補から正しい記事を絞り切れないため，正しい記事にリンクを付与できない問題がある．本研究では投稿文字数が 140 文字と制限されているツイートの文書を対象にし，文書量の少ない文中においても曖昧性のあるエンティティに対して正しくリンクを付与することができるエンティティの曖昧性解消の手法を提案する．ここで，本研究が目指しているエンティティリンキングの例を図 1 に示す．

図 1 エンティティリンキングの例

図 1 中には，2 つのツイート文があり，上部のツイートには魚を意味する「スズキ」，下部のツイートでは企業を意味する「スズキ」が文中に表れている．提案手法では，このようにエンティティに曖昧性の問題が発生しても正しいリンクを自動的に付与する．

また，テキスト中では，人名や組織名などの固有表現がエンティティとして重要な役割を持つことから，エンティティリンキングでは固有表現が重視される場合が多い．固有表現抽出を行う研究において，人手で作成した規則に基づく手法や，機械学習を用いた手法が提案されている．本研究では，新語や固有表現に強い辞書を使用することによってエンティティを抽出する．

2. 関連研究

2.1. エンティティリンキング

エンティティリンキングにおけるエンティティの対象物は，主に人名，組織名，団体名，地名などがあるが，本研究では，語義曖昧性が存在する単語をエンティティとして取得する．

古川[1]らは，日本語の技術的な文書中の専門的な用語から，英語 Wikipedia 項目へのリンクを付与するための語義曖昧性解消手法を提案している．用語抽出には，与えられた日本語の学術文献の記事から辞書を用いて専門用語を抽出している．知識ベースは英語版 Wikipedia 記事の記事およびセクションから主題, 副題, URL，文書，内部リンクを抽出して構築している．従来の辞書を使用すると，マイクロブログ上で頻繁に出現する新語や固有表現の抽出が上手く行えず，正しいエンティティの抽出が困難になる．そこで，本研究ではエンティティの抽出を行う際に辞書を用いるが，固有表現や新語に特化した辞書を使用することによってエンティティを抽出する．

また, 山田ら[2]は, Wikipedia などの知識ベースから Twitter 上の文書中のエンティティと対応させ，また, Wikipedia だけでなく, いくつかの知識ベースから学習データを用意し，それをもとにエンティティを抽出する固有表現抽出の手法を提案している．また，Wikipedia 記事タイトル，Wikipedia 上のアンカーテキストを用いて語義曖昧性解消を行う手法を提案している．本研究では，エンティティを抽出する際に辞書を用い，記事中に語義が曖昧なエンティティが出現した際に，そのエンティティ周辺の語義が曖昧ではないエンティティの情報を使用することで，語義曖昧性解消を行う．

Mihalcea[3]らは, 英語版 Wikipedia を使用し, キーワード抽出と語義曖昧性解消を行う手法を提案している. キーワード抽出には Wikipedia 記事タイトルをキーワード候補として抽出し，キーワードに重要度を付与しランク付けを行い，キーワードを絞り込んでいる．キーワードの重要度の算出には, その単語が Wikipedia 記事集合でリンクテキストとして扱われた数が多ければ多いほど，そのキーワードは重要であるとして値が大きくなるようにし, Wikipedia 記事でリンクテキストとして多く扱われたキーワードを抽出している．語義曖昧性解消にはナイーブベイズによる機械学習の手法を用いて曖昧性の解消を行っている．キーワードの品詞とそのキーワード前後 3 語の品詞を素性としている．しかし，記事タイトルに新語があるとき，新語を正しく抽出することは難しい．本研究では，それらの単語の抽出が可能な辞書を使用することによってエンティティを抽出する．

2.2. 固有表現抽出

固有表現抽出は，地名・人名・組織名などの固有名詞や日時・時間・通貨などの数値表現を文書から抽出する技術である．IREX（Information Retrieval and Extraction）というワークショップの日本語固有表現タスクでは，組織名，人名，地名，固有物名，日付表現，時間表現，金額表現，割合表現の 8 種類の固有表現を定義している．固有表現情報を抽出する手法は，主に２つに分かれる．1 つ目は人手で作成したルールに基づく手法[4]であり，２つ目は機械学習による手法である．機械学習の手法として，SVM（Support Vector Machine），最大エントロピー法，条件付き確率場に基づく手法などが提案されている．機械学習を用いた固有表現抽出では，入力文を解析単位（トークン）に分割し，固有表現を構成する 1 つもしくは複数のトークンをまとめる手法が一般的である．

山田ら[5]は SVM を用いた固有表現抽出を行っている．単語を解析の単位として，単語自身，品詞分類，文字種などを素性として利用している．実験には CRL（郵政省通信総合研究所）固有表現データを使用している. CRL 固有表現データは, 毎日新聞 95 年度版 1,174 記事, 約 11,000 文に対して固有表現が付与されている. 提案手法による抽出実験では，F 値で約 83%という高い精度の結果を得ている．

この研究を含めて，前後 2 トークンの情報を素性として使用することが一般的である．しかし，固有表現の構成要素数が多い場合は十分な素性が与えられず，解析誤りが起こりやすくなる．そこで，中野ら[6]は，文節区切りを行い文節内外の情報を素性として使用する手法を提案している．CRL 固有表現データを用いた固有表現抽出の実験結果は，F 値で 89%という結果を得ている．

しかし，これらの機械学習を用いた固有表現抽出の手法で対象にしている文書は，文が正しく整えられている新聞記事などに対する手法であり，Twitter のような文が整えられていない文書に対して正しく固有表現を抽出するのは難しいと考えられる．そこで，本研究では，正しく整えられていない文書から，固有表現に強い辞書を使用することによって固有表現抽出を行う．

2.3. mecab-ipadic-NEologd

佐藤[7]らは，形態素解析器が対応できなかった未知語や新語に対して定期的に辞書を更新していくことで未知語や新語を正しく形態素解析できる手法を提案した．この辞書は「NEologd」と呼ばれ，文を形態素単位に分かち書きするための辞書ではなく,形態素解析エンジンである MeCab と共に使用する単語分ち用の辞書である．この辞書には，主に 3 つの特徴がある．1 つ目は MeCab に標準搭載されている辞書である IPADIC

で形態素解析する際に複数の形態素に分割されてしまう固有表現を採録している点である．2 つ目は IPADIC では，辞書の更新頻度が低いので，日々増えていく新語に対応ができないが，NEologd では週に 2 度以上の辞書更新をすることで，新語に対応している点である．3 つ目は Web 上の言語資源を活用して更新時に新しい固有表現を随時追加している点である．

Web 上の言語資源は，Web 上のニュース記事や，SNS での発言などを使用し，辞書に語彙を追加している．ニュース記事では，記事から抽出した新語・未知語・固有表現を追加し，SNS では流行した単語・慣用句・ハッシュタグを使用している．

本研究では，新語や固有表現が頻繁に表れるマイクロブログからエンティティリンキングを行うので，これらの方法で作成された辞書を使用し，エンティティの抽出を行う．

3. 提案手法

本手法は図 2 のように 3 つの手順に分かれる．

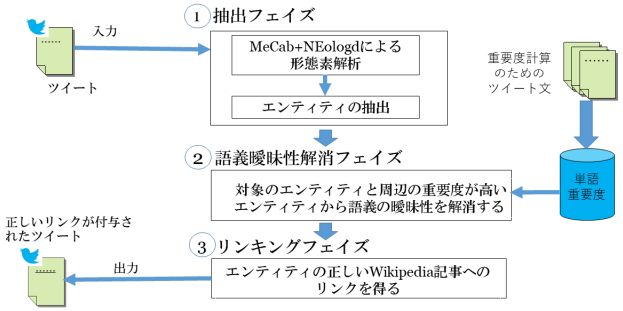

図 2 提案手法の流れ

1 つ目は，固有表現候補となるエンティティの抽出である．MeCab による形態素解析の際にエンティティを抽出するための固有表現辞書（NEologd）を形態素解析ツールに追加することでエンティティ候補を抽出する．

2 つ目は，取得したエンティティの語義曖昧性解消である．抽出したエンティティには，語義が曖昧なエンティティである場合がある．よって，その単語を対象のエンティティとして抽出し，対象エンティティの周辺の重要なエンティティの情報を使用して曖昧性を解消する

最後に，リンクが付与されていないツイートの文書に二つ目の流れで得た正しいエンティティのリンク先の情報を付与する．

3.1. 抽出フェイズ

Twitter などのマイクロブログ上では，新語や未知語がよく現れる．また，それらの単語はエンティティとして抽出しなければならない単語である場合が多い．よって，本研究では形態素解析エンジン MeCab と新語・固有表現に強い NEologd を使用することによってエンティティの抽出を行う．今回，エンティティとして抽出する単語は，Wikipedia 上で記事として存在する単語であり，Wikipedia 上で曖昧性のある単語としての記事が存在する単語である．

3.2. 語義曖昧性解消フェイズ

本研究では，図 3 のように，語義曖昧性解消のために 7 つの手順を踏み，対象エンティティの曖昧性解消を行う．

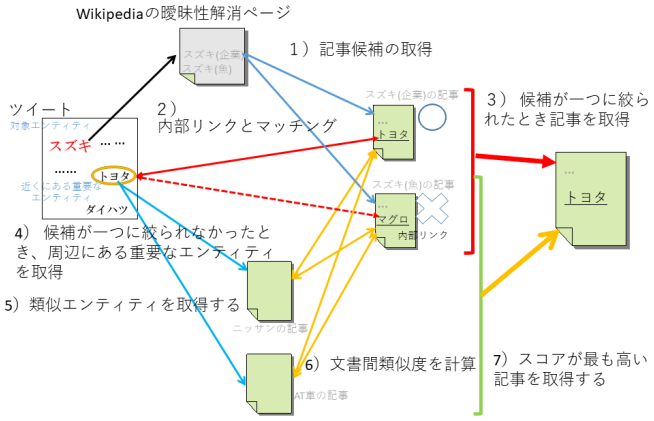

図 3 語義曖昧性解消の手順

3.2.1 正解記事候補の取得

Wikipedia 上では，図 4 のように曖昧さ回避のページが存在する．図 4 中のツイート上では，「スズキ」という単語が Wikipedia 上で曖昧性のある対象エンティティである．これをクエリとして Wikipedia 上で検索することによって，エンティティの記事の候補が取得できる．この「スズキ」というエンティティの場合，企業名であるスズキと，魚の名前であるスズキが正解記事候補として取得できる．

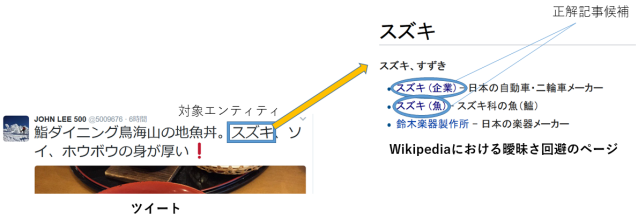

図 4 Wikipedia における曖昧さ回避のページ例

3.2.2 内部リンクとのマッチング

語義曖昧性解消のキーワードを得るために，対象エンティティを除く，Twitter 上に出現し，語義曖昧性が存在せず，かつ Wikipedia 記事が存在するエンティティを全て抽出し，正解記事候補の Wikipedia ページの内部リンクと，それらのエンティティがマッチングするかどうか判定を行う．図 3 を例に詳しく説明をすると，対象エンティティが「スズキ」で，ツイート本文中には他に「トヨタ」，「ダイハツ」などのエンティテ

ィが出現する．これらのエンティティと「スズキ」の記事候補の Wikipedia 記事と内部リンクのマッチングを行う．企業の「スズキ」の Wikipedia 記事には「トヨタ」という内部リンクがあり，魚の「スズキ」の Wikipedia 記事にはツイート上のエンティティが存在しない．よって企業の「スズキ」の記事にマッチングしたということになる．

3.2.3 内部リンクによる正解記事の取得

3.2.2 節で正解記事候補を 1 つに絞られた場合，その記事を正解記事として抽出し，手順を終了する．1 つに絞られなかった場合，3.2.4 節の手順に移る．

3.2.4 周辺の重要なエンティティの取得

事前に収集した大量のツイート文から OkapiBM25+を用いてエンティティの重要度を算出しておく．

OkapiBM25 は，情報検索の分野で TF-IDF よりも精度が高いとされる重みづけ手法であり，TF-IDF を拡張したものである．ある単語の出現回数が同一である 2 つのドキュメントがあるとき，「ある単語の重要度は 2 つのドキュメントに対して同等ではなく，短い文章に対する重要度がより高くなる」という考えを，TF-IDF に追加している．OkapiBM25+[8]は，OkapiBM25 を拡張したものであり，短い文書に高い重要度を付与するよう改良されている．

提案手法ではエンティティの重要度を OkapiBM25+に加え，曖昧性のあるエンティティからの形態素距離が近くなるほど重要度が大きくなるように重みづけを行う．OkapiBM25+の式は(1)式の$w(e,t)$，(1)式中の $idf(e)$は(2)式で計算する．

ここで，(1)式，(2)式の説明をする．$tf(e,t)$は，ツイート上における対象エンティティ e の出現頻度を示す．$df(e)$は対象エンティティが事前に収集したツイート集合中に現れるツイートの数である．$idf(e)$は，$df(e)$の逆数に，事前に収集したツイート数を表す$|N|$を掛け，log をとった値である．$len(t)$は，ツイート t の長さ，$avgtl$ は全ツイートの平均の長さを示す．k_1とbとδは調整パラメータであり，k_1は 1.2 から 2.0 の値で，b は 0.75，δは 1.0 をパラメータとして設定する．

$$w(e,t) = idf(e) \cdot \left(\frac{tf(e,t) \cdot (k_1 + 1)}{tf(e,t) + k_1 \cdot \left(1 - b + b \cdot \frac{len(t)}{avgtl}\right)} + \delta \right) \tag{1}$$

$$idf(e) = \log\left(\frac{|N|}{df(e)}\right) \tag{2}$$

$$weight(e) = 1 - k \cdot |l| \tag{3}$$

$$entity = \arg\max_e \left(w(e,t) \cdot weight(e)\right) \tag{4}$$

重みづけの式は(3)式の$weight(e)$で計算する．(3)式の$|l|$は，対象エンティティからの形態素の距離を表し，k は調整パラメータを表す．例えば，「私は海釣りをしてタイとスズキとカンパチを釣りました」というツイートがあったとき，NEologd で形態素解析を行った結果，「私-は-海釣り-を-し-て-マダイ-と-スズキ-と-カンパチ-を-釣り-まし-た」という形態素に分けられる．この時に対象となる曖昧性のあるエンティティは「スズキ」であり，ツイート文中で表れるエンティティは「海釣り」，「マダイ」，「カンパチ」である．このとき$|l|$の値は「海釣り」は 6，「マダイ」は 2，「カンパチ」は 2 となる．

最終的に重要なエンティティは，(4)式の$entity$によって計算する．(4)式では，(1)式から(3)式で得た，収集したエンティティから見たエンティティの重要度である$w(e,t)$，対象エンティティからの形態素の距離による重みづけである$weight(e)$を掛け合わせたものの中から最大の値を取るエンティティを，対象エンティティから近くにあり，かつ重要なエンティティとして抽出する．このエンティティを取得する際，曖昧性の有無は考慮しない．

3.2.5 類似するエンティティの取得

ツイートは分量が 140 文字に制限されており，新聞記事やブログなどの文書に比べて文書中に出現するエンティティの数は少ない．よって，語義曖昧性解消のキーワードとなるエンティティの数を増やすために，曖昧性のあるエンティティに近くにある重要なエンティティで，かつ曖昧性のないエンティティが，対象エンティティに類似するエンティティであるだろうと想定し，Word2vec[9]を用いて 3.2.4 節で取得したエンティティに類似するエンティティを取得する．Word2vec の学習には日本語 Wikipedia 記事の全データを使用する．

Word2vec は，Mikolov[9]らによって提案された，ニューラルネットワークを用いた単語の分布表現(単語の概念を表す低次元の密なベクトルによる表現)を獲得する手法である．テキストコーパスを入力すると，コーパスにある単語の特徴量ベクトルが出力される．単語の特徴や意味構造を含めてベクトル化することができる．このことによって，意味的に近い単語は，空間上で近くに存在するベクトルとして表現され，入力した単語に類似する単語の取得に用いられている．本実験では，入力するテキストコーパスとして，日本語版 Wikipedia 記事の全記事を用い，対象エンティティに最も近い曖昧性のないエンティティの類似する単語を抽出した．

3.2.6 文書間類似度の計算

3.2.5 節の手順で抽出した，対象エンティティの近く

にある重要なエンティティに類似するエンティティの Wikipedia 記事と，対象エンティティの Wikipedia 記事との文書間類似度を(5)式によってそれぞれ計算する．

$$cos\,(A,B) = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}} \qquad (5)$$

A，B は，それぞれ，エンティティ候補の Wikipedia 記事と，Word2vec が出力した類似エンティティの Wikipedia 記事中の名詞の頻度ベクトルであり，n は A，B の記事中の名詞の総数である．本研究では Word2vec によって類似エンティティを 5 件取得し，語義曖昧性解消の候補エンティティの Wikipedia 記事と，類似エンティティ 5 件の Wikipedia 記事との文書間類似度の平均をスコアとして算出する．

3.2.7 正解記事の抽出

3.2.6 節の手順で計算した候補エンティティの Wikipedia 記事と類似エンティティの Wikipedia 記事との文書間類似度の平均スコアが最も高い記事を曖昧性のある対象エンティティの正解記事として抽出する．

3.3. リンキングフェイズ

3.2.3 節と 3.2.7 節の手順で取得した正解記事を対象エンティティのリンクとして付与し，正しいリンクが付与されたツイート文を出力する．

4. 実験・考察

本章では，提案手法の有効性を検証するための評価実験とその考察について述べる．評価実験には，Twitter から取得した 50 件の語義曖昧性が存在するエンティティが含まれるツイート文を用いて提案手法の有効性を検証した．

4.1. 実験データ

実験データとして，Twitter から対象エンティティをキーワードとして検索し，出力されたツイートを実験データとして使用した．本実験ではツイート文を 50 件，エンティティの種類は人名，動物名，地名の三種類を，TwitterAPI を用いて収集した．今回実験を行ったのは，2 件の人名の 20 ツイート，2 件の動物名の 20 ツイート，1 件の地名の 10 ツイートである．

提案手法の Twitter 上の単語重要度を計算するために，2018 年 1 月 9 日の 4:00 から 2018 年 1 月 10 日 21:00 までの日本語ツイートを，TwitterStreamingAPI[1] を使用し，取得した．ツイート総数は，1,046,967 ツイートで，平均単語数は 21.3 単語であった．

[1] https://developer.twitter.com/en/docs/tutorials/consuming-streaming-data

4.2. 実験

本実験は，提案手法との比較のため，3 つのベースライン手法を用いて実験を行った．

1 つ目は，ツイート文中のエンティティと，曖昧性のある対象エンティティの Wikipedia 記事の内部リンクとのマッチングによって正解記事を抽出する方法である．実験結果の表中では，「Mtg」という名の列で示す．

2 つ目は，対象エンティティに最も近いエンティティに対して，Word2vec を用いて出力した 5 つの類似エンティティの Wikipedia 記事と，語義曖昧性のある対象エンティティの候補の Wikipedia 記事との文書間類似度を計算し，文書間類似度の平均が最も高い対象エンティティの候補の Wikipedia 記事を正解記事として抽出する方法である．実験結果の表中では，「Cos」という名の列で示す．

３つめは提案手法において，対象エンティティからの距離を考慮せず，OkapiBM25+の数値の結果のみ用いて，最も重要度が高いエンティティをツイート文中の重要エンティティを取得し，そのエンティティに対して，Word2vec を用いて出力した 5 つの類似エンティティの Wikipedia 記事と，語義曖昧性のある対象エンティティの候補の Wikipedia 記事との文書間類似度を計算し，文書間類似度の平均が最も高い対象エンティティの候補の Wikipedia 記事を正解記事として抽出する方法である．実験結果の表中では，「BM25」という名の列で示す．

それらの手法の精度を，取得した 50 件のツイートで比較する．提案手法の実験結果は表中では，「Pro」という列で示す．また，実験結果の表中の「マッチ単語」という列では，マッチングの手法により，Wikipedia の内部リンクと，ツイート文中のエンティティがマッチングした単語を表記している．

4.3.1 対象エンティティが人名の場合の結果

対象エンティティが人名の語義曖昧性のあるエンティティである場合について実験を行った．今回対象エンティティにした人名は，「田中理恵」，「ジョン・ウィリアムズ」という 2 件のエンティティであり，それぞれ 10 ツイートずつ，マッチングによる手法，コサイン類似度による手法，提案手法で正しい記事が取得されるか検証した．「田中理恵」には 2 種類の正解記事候補があり，一人目の候補は声優の田中理恵で，二人目の候補は体操選手の田中理恵である．「田中理恵」に対する実験結果を表 1 に記す．また，「ジョン・ウィリアムズ」には，5 種類の曖昧性候補があり，1 人目の候補

は，

作曲家，2 人目の候補はギタリスト(Gr)，3 人目の候補は俳優，4 人目の候補は政治学者，5 人目の候補は宣教師の「ジョン・ウィリアムズ」である．「ジョン・ウィリアムズ」に対する実験結果を表 2 に示す．

表 1 「田中理恵」に対する実験結果

	Mtg	Cos	BM25	Pro	マッチ単語
T1(声優)	○	○	○	○	声優
T2(声優)	×	○	○	○	
T3(声優)	×	○	×	×	
T4(体操)	×	○	○	○	
T5(声優)	×	○	○	○	
T6(声優)	○	○	○	○	アニメ
T7(声優)	×	○	○	○	
T8(声優)	×	×	×	×	
T9(声優)	○	○	○	○	声優
T10(声優)	×	○	×	×	
精度	0.3	0.9	0.7	0.7	

表 2 「ジョン・ウィリアムズ」に対する実験結果

	Mtg	Cos	BM25	Pro	マッチ単語
T1(作曲家)	○	○	○	○	映画音楽
T2(作曲家)	○	○	○	○	映画音楽
T3(作曲家)	×	×	○	×	
T4(作曲家)	○	○	○	○	映画音楽
T5(作曲家)	○	×	○	○	映画音楽
T6(作曲家)	×	○	×	×	
T7(Gr)	×	○	×	×	
T8(Gr)	×	○	×	×	
T9(俳優)	○	×	×	○	俳優
T10(俳優)	○	×	×	○	俳優
精度	0.6	0.6	0.5	0.6	

4.3.2 対象エンティティが動物名の場合の結果

次に，対象エンティティが動物名の語義曖昧性のあるエンティティである場合について実験を行った．今回対象エンティティにした動物名は，「スズキ」，「ライオン」という 2 件のエンティティであり，それぞれ 10 ツイートずつマッチングによる手法，コサイン類似度による手法，提案手法で正しい記事が取得されるか検証した．「スズキ」には 2 種類の曖昧性候補があり，1 つ目の候補は魚の「スズキ」で，2 つ目の候補は企業の「スズキ」である．「スズキ」に対する実験結果を表 3 に示す．

また，「ライオン」には 7 種類の曖昧性候補があり，一つ目の候補はアニメソングのタイトルで，2 つ目の候補はメタルバンドのグループ名，3 つ目の候補は日本企業名，4 つ目の候補は戦艦の名前，5 つ目の候補は紋章学の象徴として使われている名前，6 つ目の候補は日本の遊助というアーティストの曲名，7 つ目の候補は動物名の「ライオン」である．「ライオン」に対する実験結果を表 4 に示す．

表 3 「スズキ」に対する実験結果

	Mtg	Cos	BM25	Pro	マッチ単語
T1(魚)	×	○	○	○	
T2(企業)	×	×	○	○	
T3(企業)	×	×	×	×	
T4(企業)	×	×	○	○	
T5(企業)	×	○	×	×	
T6(魚)	○	○	○	○	釣り
T7(企業)	×	○	○	○	
T8(企業)	×	×	×	○	
T9(魚)	×	○	○	○	
T10(魚)	×	○	○	○	
精度	0.1	0.6	0.7	0.8	

表 4 「ライオン」に対する実験結果

	Mtg	Cos	BM25	Pro	マッチ単語
T1(動物)	×	○	○	○	
T2(動物)	○	○	○	○	亜種
T3(動物)	○	○	○	○	ケニア
T4(動物)	×	×	×	×	
T5(動物)	×	○	○	○	
T6(動物)	×	○	○	○	
T7(動物)	×	○	×	×	
T8(動物)	○	○	○	○	トラ
T9 (アニソン)	○	×	○	○	マクロス F
T10(企業)	×	×	○	○	
精度	0.4	0.7	0.8	0.8	

4.3.3 対象エンティティが地名の場合の結果

最後に，対象エンティティが地名の語義曖昧性のあるエンティティである場合について実験を行った．今回対象エンティティにした地名は,「アフリカ」というエンティティである．収集した 10 ツイートをそれぞれの手法で正しい記事が取得されるか検証した．アフリカには 5 種類の曖昧性候補があり，1 つ目の候補は地域の名前，2 つ目の候補は戦艦の名前，3 つ目の候補は交響曲の名前，4 つ目の候補はアメリカのロックバンドの TOTO の曲名，5 つ目の候補は小惑星の名前の「アフリカ」である.「アフリカ」に対する実験結果を表 5 に記す．

表 5 「アフリカ」に対する実験結果

	Mtg	Cos	BM25	Pro	マッチ単語
T1(国)	○	○	○	○	ナイジェリア
T2(国)	○	○	○	○	アパルトヘイト
T3(国)	×	○	○	○	
T4(国)	×	○	○	×	
T5(国)	×	○	○	○	
T6(国)	×	○	○	○	
T7(国)	×	○	○	○	
T8(国)	×	○	○	○	
T9(国)	×	○	○	○	
T10(国)	×	○	○	○	
精度	0.2	1.0	1.0	0.9	

4.3. 考察

実験ツイート 50 件中，マッチングによる手法の精度は 32%，コサイン類似度による手法の精度は 76%，提案手法による精度は 84%であった．提案手法はマッチングによる手法の結果に依存しており，精度が低下する原因はマッチング手法の結果に影響していると考えられる．

図 8 のように，対象エンティティ「田中理恵」のツイート文中のエンティティの最も近いエンティティが「シャイニールミナス」というエンティティのとき，正解記事である声優の「田中理恵」の Wikipedia 記事中の内部リンクとして,「シャイニールミナス」というエンティティは存在していなかったが，声優の「田中理恵」の Wikipedia 記事中に，重要なエンティティのとき，太字の表記として「シャイニールミナス」というエンティティが存在していた．よって内部リンクだけでなく，重要なエンティティの記述であることが多い太字表記のエンティティもマッチングの対象にすることで，精度の向上が見込まれる．

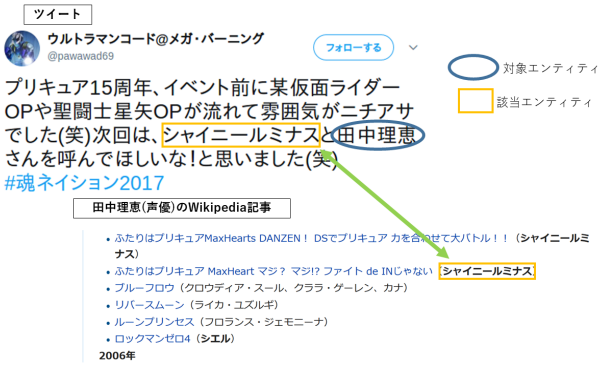

図 8 精度向上が見込まれるツイート

5. おわりに

本論文では，Twitter などの文字数が少ない文書であるマイクロブログ上で、地名・団体名など何らかの実体を指すエンティティの情報を得る際に，対象エンティティに曖昧性が存在していても，正しい記事を抽出できるエンティティリンキングの手法を提案した．

提案手法の実験結果は，精度は高い結果となった．しかし，本手法はツイート文中に語義曖昧性の存在するエンティティが 1 つと，それ以外のエンティティが文中に最低 1 回存在するときのみ適用できる手法である．

例えばツイート文に「スズキかっけえ」というツイート文が存在している場合，本手法では,「スズキ」という曖昧性の存在するエンティティは取得できるが，文中にエンティティが 1 つしか存在しないため，そのエンティティの正解候補記事を絞るためのキーワードが得ることが出来ない問題がある．その問題を回避するため，文中にエンティティが 1 つのみで，かつそのエンティティが曖昧性の存在するエンティティである場合は，対象ツイート文の前後のツイート文を，語義曖昧性解消のための対象文書とするなどの方法で曖昧性を解消していく必要がある．

参　考　文　献

[1] 古川竜也，相良毅，相澤彰子：言語横断エンティティリンキングのための語義曖昧性解消，情報処理学会誌，Vol.24，No.2，pp.172-177，(2014).

[2] Ikuya Yamada, Hideaki Takeda, Yoshiyasu Takefuji “Enhancing Named Entity Recognition in Twitter Messages Using Entity Linking” In Proceedings of the ACL 2015 Workshop on Noisy User-generated, pp. 136–140, (2015).

[3] Rada Mihalcea, Andras Csomai, “Wikify! :linking documents to encyclopedic knowledge ” In Proceedings of the 16th ACM Conference on Information and Knowledge management， pp. 233-242， Portugal，(2007).

[4] 渡辺一郎，桝井文人，福本淳一：固有表現抽出ツール NExT の精緻化とユーザビリティの向上，言語処理学会第 10 回年次大会 ,(2004).

[5] 山田寛康，工藤拓，松本裕治：Support Vector Machines による日本語固有表現抽出,情報処理学会論文誌，Vol.43，No.1，pp.44-53（2002）.

[6] 中野桂吾，平井有三：日本語固有表現抽出における文節情報の利用,情報処理学会論文誌, Vol.45，No.3，pp.934-941（2004）.

[7] 佐藤 敏紀，橋本 泰一，奥村 学：単語分かち書き辞書 mecab-ipadic-NEologd の実装と情報検索における効果的な使用方法の検討,言語処理学会第 23 回年次大会発表論文集，(2017).

[8] Y. Lv, C. X. Zhai “Lower-bounding term frequency normalization” In Proceedings of CIKM'2011, pp.7-16, (2011).

[9] Tomas Mikolov, Kai Chen, Greg Corrado, Jeffrey Dean, “Efficient Estimation of Word Representations in Vector Space” In Proceedings of the International Conference on Learning Representations (ICLR), (2013).