Wikipediaのページビュー数を用いたTwitter上の話題語に対するエンティティリンキング

中村 達哉[1,a)]　白川 真澄[1,b)]　原 隆浩[1,c)]　西尾 章治郎[1,d)]

概要：本稿では，ある期間に投稿されたツイートの集合において頻出する語句を話題語として抽出し，対応するWikipediaの記事に話題語を紐付けるエンティティリンキングについて述べる．Twitter上の話題の意味解析や異なる言語間での話題の比較を行う場合，エンティティリンキングにより話題語を知識体系のエントリとして表現することが有効であるが，エンティティリンキングを高精度で行う必要がある．そこで提案手法は，話題語の出現頻度とWikipediaのページビュー数の相関性に着目し，同じ期間に多く閲覧されているWikipediaの記事に話題語を紐付ける．評価実験の結果から，提案手法がTwitter上で話題である度合いが高い語句ほど，より高精度にエンティティリンキングを行えることを確認した．

1. はじめに

テキスト集合に含まれるトピック情報を抽出する研究は数多く行われているが，近年，Twitter[*1]に代表されるマイクロブログがその対象として注目を集めている．その理由として，マイクロブログのリアルタイム性が挙げられる．マイクロブログでは，様々な人が実世界の出来事や興味・関心についての情報を短いテキストという形式で常時発信している．実際にTwitterでは，ツイートと呼ばれる最大140文字の短いテキストが1日に5億回以上投稿されている[*2]．また最近では，官庁や自治体等の行政機関や報道機関といった公的な組織もマイクロブログを通じてリアルタイムな情報発信を積極的に行っている．このようなマイクロブログのテキストを解析することで，現在注目を集めている出来事やその出来事に対する人々の反応といった即時性が高いトピック情報を抽出できる [1], [3].

マイクロブログのような短いテキストからトピック情報を抽出する場合，エンティティリンキングと呼ばれる技術により，テキスト中に出現するエンティティを表す語句を対応する知識体系のエントリ（本研究ではWikipedia[*3]の記事）に紐付けることが有効である [6], [12]．エンティティリンキングは，テキストが持つ意味情報を知識体系のエントリとして明示的に表現できるため，抽出したトピック情報について知識体系内の情報を用いたさらなる意味解析が可能となる．また，Wikipediaのような多言語で展開されている知識体系を用いることで，異なる言語間の比較を行うことができる．このような背景から，近年，マイクロブログのような短いテキストを対象としたエンティティリンキングに関する研究が盛んに行われている [6], [8], [9], [12], [17]．特に，Wikipediaのページビュー数を用いるアプローチは，ツイートに対するエンティティリンキングにおいて大幅に性能を改善できることが明らかにされている [19]．これは，ツイートに出現する語句に対応するWikipediaの記事のページビュー数が増加する傾向にあることを利用している．

本研究では，Wikipediaのページビュー数を用いるアプローチを，Twitter上の話題語に対するエンティティリンキングに用いることを考える．Twitterからの話題抽出におけるエンティティリンキングでは，Twitter上で話題となっている語句についてのみ，高い精度で語句に対応する知識体系のエントリを発見できれば良い．Wikipediaのページビュー数を用いることによる精度向上は，Twitter上で話題になっている語句ほど顕著であると考えられるため，より高精度なエンティティリンキングが可能になると考えられる．また，話題抽出の対象として，ある短い期間に投稿されたツイートの集合を入力としたとき，そのツイート集合中の曖昧性を持つ話題語は，特定の話題を指し示している可能性が高い．そのため，ある期間に投稿されたツイート集合中で出現頻度が多く話題となっている語句

[1] 大阪大学大学院情報科学研究科
[a)] nakamura.tatsuya@ist.osaka-u.ac.jp
[b)] shirakawa.masumi@ist.osaka-u.ac.jp
[c)] hara@ist.osaka-u.ac.jp
[d)] nishio@ist.osaka-u.ac.jp
[*1] https://twitter.com/
[*2] http://www.sec.gov/Archives/edgar/data/1418091/000119312513390321/d564001ds1.htm
[*3] https://www.wikipedia.org/

に対して，まとめて一つのエントリを紐付けることができると考えられる．

そこで本研究では，Twitter 上の話題を表す語句に対して高精度にエンティティリンキングを行う手法を提案する．提案手法ではまず，ツイート集合において頻出する語句を話題語として抽出する．そして，話題語に対する Wikipedia の記事の候補の中から，入力のツイート集合と同じ期間に多く閲覧されている Wikipedia の記事を話題語に紐付けることにより，Twitter 上の話題語に対する高精度なエンティティリンキングを実現する．評価実験を行い，Twitter 上で話題である度合いが高い語句ほど，より高精度に対応する Wikipedia の記事を発見できることを実証する．また，Twitter 上での語句の出現頻度と Wikipedia のページビュー数の相関性を考慮して，不適切な話題語の除去や対応する Wikipedia の記事が存在しない話題語の特定が可能かどうかについても検証する．

2. 関連研究

エンティティリンキングに関する研究は Wikify! [13] を発端として，以降急速に研究対象としての認知度が高まっている．一般的にエンティティリンキングの処理は，キーワード抽出，エンティティの曖昧性解消の順に行われる．

Wikify! [13] では，ある語句が Wikipedia の記事中でアンカーテキストとして用いられる度合いを表すスコア keyphraseness を定義し，keyphraseness が TF-IDF [18] などの語句の重み付け手法よりも高い精度でテキスト中のキーワードを抽出できることを示した．また，エンティティの曖昧性解消においては，Lesk アルゴリズム [11] を用いた手法と Naive Bayes を用いた手法を組み合わせた手法を提案している．Cucerzan の研究 [4] では，Wikipedia から抽出したエンティティに関するコンテキスト情報やカテゴリ情報などを用いた手法を提案している．この手法では，入力テキストと Wikipedia の記事を Wikipedia から抽出した情報によりベクトル化し，ベクトルの内積（類似度）に関する最大化問題を解くことで，入力テキスト中の各キーワードに対応する記事のリストを求めている．Milne ら [15] は，機械学習を用いた手法 Wikipedia Miner[*4] を提案している．まず，Wikipedia において曖昧性を持たない（リンク先の記事の候補が一つのみ存在する）語句によってリンクされる記事を収集する．そして，収集した記事とそれ以外の記事について，どのような記事間関連度 [14] を持ち，また，同時にリンクされやすいかを学習することで，エンティティリンキングを実現している．

Kulkarni らの研究 [10] では，エンティティの曖昧性解消において，入力テキストの各キーワードに対する局所的なスコアと大局的なスコアを導入した手法を提案している．AIDA[*5] は知識体系のリンク構造に対してグラフ理論を用いたエンティティリンキング手法である [7]．AIDA では，Mention-Entity Graph と呼ばれる，キーワード（Mention）と知識体系のエントリ（Entity）をノード，キーワード・エンティティ間および異なるエンティティ間の類似度をエッジとして重み付き無向グラフを定義している．このグラフから，入力テキスト中に出現するキーワードのノードを全て含んだ高密度な部分グラフを抽出し，抽出したグラフを用いて，各キーワードに対応するエントリを決定している．

マイクロブログのテキストのような短いテキストを対象とした手法も提案されている．TAGME[*6] は短いテキストを対象とした高速なエンティティリンキング手法である [6]．TAGME では，入力テキストから Wikipedia のアンカーテキストとして用いられている語句をキーワードとして抽出し，それぞれのキーワード (アンカーテキスト) によってリンクされる記事の候補の中から，互いに関連性の高い記事を付与するという処理で，高速なエンティティリンキングを実現している．Meij らの研究 [12] では，入力テキスト中のキーワードの長さや keyphraseness，候補となる記事が持つリンク数やカテゴリ数などを組み合わせた合計 33 の素性を用いて機械学習を行い，Twitter のツイートに対するエンティティリンキングを実現している．Yamada ら [19] は，テキスト中の語句のミススペルや省略を考慮したキーワード抽出手法，および，候補となる記事が持つリンク数やページビュー数などを素性に用いた機械学習によるエンティティリンキング手法を提案している．Yamada らの手法は，Web 研究に関する世界最大の国際会議 WWW2015（World Wide Web）で開催されたエンティティリンキングに関するコンペティション Microposts2015 において 2 位に大差をつけて優勝しており，Yamada らの手法で用いられている素性，とりわけ他のエンティティリンキング手法では使われていない Wikipedia のページビュー数がエンティティリンキングに対して有効であることを示している．

このようにエンティティリンキングに関する多くの研究が行われているが，これらの研究では，任意のトピックについて記述されたテキストを入力として，エンティティを表す語句とそれに対応する知識体系のエントリを網羅的に抽出する，汎用的かつ高精度なエンティティリンキングの実現を目指している．本研究は Twitter からの話題抽出をアプリケーションとして想定したエンティティリンキングを目指しており，既存研究とは目的が異なる．なお，本研究の提案手法は Yamada らの手法 [19] と同様に，Wikipedia のページビュー数を利用するが，本研究では Twitter 上の話題語に焦点を当てることで，エンティティリンキングの性能をさらに向上させられることを示す．

[*4] http://wikipedia-miner.cms.waikato.ac.nz/

[*5] http://www.mpi-inf.mpg.de/yago-naga/aida/

[*6] http://tagme.di.unipi.it/

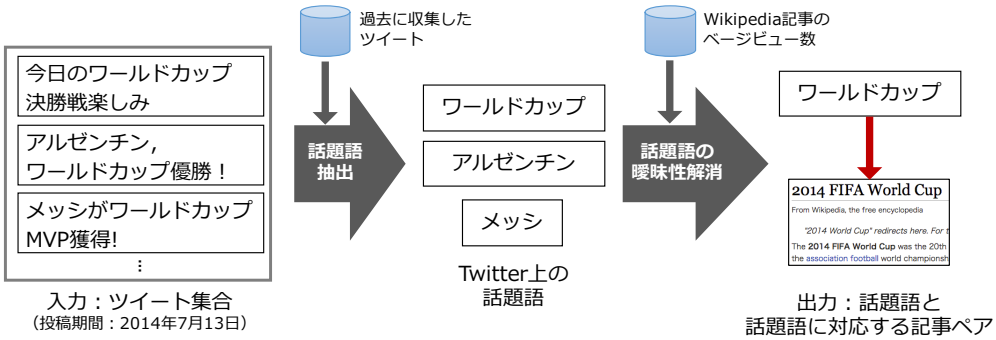

図 1　提案手法の流れ

3. 提案手法

本章ではまず，Twitter 上の話題語に対してエンティティリンキングを行う提案手法の概要について述べる．そして，Twitter 上の語句の出現頻度と Wikipedia のページビュー数の相関性を用いて，Twitter 上で話題を表す語句を抽出し，その語句に対応する Wikipedia の記事を紐付ける手法について述べる．

3.1 提案手法の概要

図 1 に提案手法の流れを示す．提案手法はまず，ある短い期間内に投稿されたツイートの集合を入力とし，入力のツイート集合中で頻出する語句を話題語（話題を表している語句）として抽出する．そして，抽出した話題語からリンクされ，かつ，同じ期間にページビュー数が多い Wikipedia の記事を，話題語に対応する記事として紐付ける．提案手法によって抽出した話題語とそれに対応する記事を用いることで，抽出した話題語を含む入力集合の各ツイートに対するエンティティリンキングを実現できる．以下では，これらの処理について詳しく説明する．

3.2 話題語の抽出

提案手法では，まず入力のツイート集合の各ツイート中から Wikipedia のアンカーテキストとして用いられている語句をキーワードとして抽出する．ツイートからのキーワードの抽出はトライ木を用いて行い，最長一致の語句のみを採用する．そして，キーワードの出現頻度について総計を取り，入力集合の中で頻出しているキーワードを話題語として抽出する．具体的には，入力のツイートが投稿された期間より以前に投稿されたツイートの情報を用いて，あるキーワード a が入力のツイート集合でのみ特に出現頻度が多いかどうかの度合いを表すスコア Trendiness ϕ_{term} を導入する．

$$\phi_{term}(a) = \frac{AvgCount_{input}(a)}{AvgCount_{input}(a) + AvgCount_{past}(a)} \quad (1)$$

$AvgCount_{input}(a)$ は入力のツイート集合において語句 a を含むツイート数の日平均，$AvgCount_{past}(a)$ は過去に投稿されたツイート集合において語句 a を含むツイート数の日平均であり，ϕ_{term} は 0 から 1 の実数値を取る．$\phi_{term}(a) > 0.5$ のとき，入力集合における語句 a を含むツイートの平均数が，過去のツイート集合における語句 a を含むツイートの平均数より多いことを表している．また，$\phi_{term}(a) = 1.0$ のとき，語句 a は入力集合において初めて出現した語句であることを表している．提案手法ではあるしきい値 τ について，$\phi_{term}(a) \geq \tau$ を満たす語句を話題語として抽出する．Trendiness ϕ_{term} のような，入力のテキスト集合における出現頻度と比較用のテキスト集合における出現頻度の比から語句の特徴度を算出する方法は，新語抽出や特徴語抽出に関する研究で用いられている [2], [5], [16]．本研究では Trendiness ϕ_{term} のスコアを話題語の条件として定義する．すなわち，τ 未満の ϕ_{term} を持つ語句は話題語にはならない．

表 1 に，Twitter Streaming API[*7] を用いて収集した，2015 年 6 月 8 日に投稿された日本語のツイート集合を入力としたときの語句の Trendiness ϕ_{term} および出現頻度の一例を示す．ϕ_{term} の計算では，過去 1 週間（2015 年 6 月 1 日から 7 日まで）のツイートを用いた．2015 年 6 月 8 日は，Apple の開発者向けイベント Worldwide Developers Conference の開催日であるため，イベントを表す語句「WWDC」の ϕ_{term} が高く，入力のツイート中で特に頻繁に出現していることがわかる．他にも，同日に販売が再開された商品を表す語句「ペヤングソース焼きそば」や，従業員が顧客の情報をマイクロブログ上で流出させた問題に関する語句「りそな銀行」，同日に開催されたトニー賞の授賞式で受賞を逃した日本人の俳優に関する語句「渡辺謙」など，出現頻度の絶対数に関わらず入力のツイート集合の期間中に起きた出来事を表す語句の ϕ_{term} が高くなっていることが確認できる．一方，「おはよう」や「時間」，「今日」など出現頻度は多いが話題を表さないような語句については，ϕ_{term} が低くなっていることが確認できる．

[*7] Public stream の sample を使用，`https://dev.twitter.com/streaming/overview`

表 1 2015 年 6 月 8 日に投稿されたツイート集合における語句の Trendiness ϕ_{term} と出現頻度

語句	ϕ_{term}	出現頻度
ペヤングソース焼きそば	1.0	11
りそな銀行	0.987	67
渡辺謙	0.970	74
WWDC	0.961	184
おはよう	0.501	5,143
時間	0.476	4,114
今日	0.453	9,288

3.3 話題語の曖昧性解消

3.2 節の処理によって抽出した各話題語 a について，その話題語によってリンクされる記事の集合 $Page(a)$ のうち，話題語がどの記事 $p_a \in Page(a)$ を表しているかを決定する．ここで，話題語とそれが指し示す記事に対して考えられる性質について整理する．まず，短い期間内に投稿されたツイートの集合中で話題になっている語句は，ある特定の話題の発生に伴って出現頻度が増加したと考えられる．そのため，出現するツイートごとにその語句の意味は変わらないと仮定できる．また，Wikipedia のページビュー数と Web 検索頻度との間に強い相関があり，Wikipedia のページビュー数も話題の発生に伴って増加する現象が見られる [20]．このことから，ページビュー数が多い Wikipedia の記事が，その記事を表しうる Twitter 上の話題語の内容を表していることが示唆される．これらを考慮すると，Twitter 上の話題語（ϕ_{term} が高い語句）について，出現したツイートに関わらず，ページビュー数の多い Wikipedia の記事がリンク先の候補として適切である可能性が高い．

そこで提案手法では，入力のツイート集合の投稿期間においてページビュー数が多い記事をリンク先として決定する．しかし，話題語の抽出と同様に，単純にページビュー数の多い記事をリンク先の記事として決定してしまうと，日常的に閲覧回数の多い記事が話題語に紐付けられてしまう．そこで，記事 p_a のページビュー数についても，過去のページビュー数の情報を用いて，記事の閲覧回数の増加の度合いを表すスコア Trendiness ϕ_{page} を算出する．

$$\phi_{page}(p_a) = \frac{AvgView_{input}(p_a)}{AvgView_{input}(p_a) + AvgView_{past}(p_a)} \quad (2)$$

$AvgView_{input}(p_a)$ は入力のツイート集合の投稿期間における記事 p_a の日平均閲覧回数，$AvgView_{past}(p_a)$ は過去のツイート集合の投稿期間における記事 p_a の日平均閲覧回数である．

リンク先の記事の候補として近い Trendiness ϕ_{page} の値を持つ記事が複数存在する場合も考慮する必要がある．このような場合，提案手法では，話題語の語句からリンクされやすい記事ほどその語句との関連性が高いという仮説をもとに，話題語によってリンクされる確率が高い記事をリンク先の記事として決定する．語句 a がアンカーテキストとして記事 p_a へリンクする確率 $P(p_a|a)$ は commonness と呼ばれ [15]，以下の式で定義される．

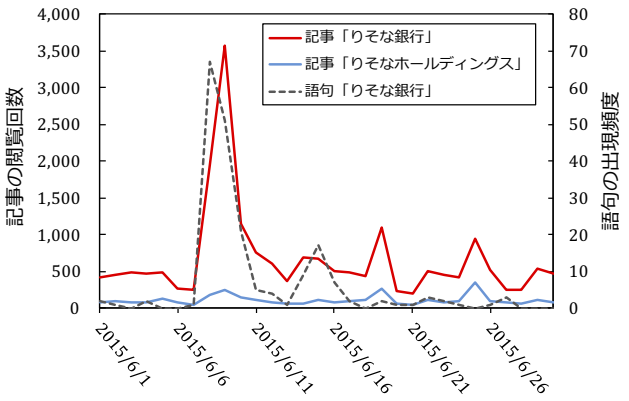

図 2 2015 年 6 月における語句「りそな銀行」の日別出現頻度と語句「りそな銀行」からリンクされる記事の日別閲覧回数

$$P(p_a|a) = \frac{CountAnchortexts(a, p_a)}{\sum_{p_i \in Page(a)} CountAnchortexts(a, p_i)} \quad (3)$$

ここで，$CountAnchortexts(a, p_a)$ は話題語 a がアンカーテキストとして記事 p_a にリンクしている回数である．

最終的に話題語 a とそのリンク先の候補の記事 p_a について次のスコアを計算し，各話題語について最大のスコアを持つ記事をその話題語のリンク先の記事として決定する．

$$\rho(a \mapsto p_a) = \frac{1}{2}(\phi_{page}(p_a) + P(p_a|a)) \quad (4)$$

提案手法では，記事の Trendiness $\phi_{page}(p_a)$ とリンク確率 $P(p_a|a)$ の平均を用いて話題語のリンク先の記事を決定しているが，$\phi_{page}(p_a)$ を特に重視してリンク先を決定する等，$\phi_{page}(p_a)$ と $P(p_a|a)$ の重みを考慮した方法も考えられる．

図 2 に，2015 年 6 月における，Twitter 上での語句「りそな銀行」の日別出現頻度と，語句「りそな銀行」からリンクされる記事の日別閲覧回数を示す．語句「りそな銀行」はリンク先の記事として「りそな銀行」と「りそなホールディングス」の二つを持つ．2015 年 6 月 8 日ごろの回数を比べると，語句「りそな銀行」の出現頻度の急増と同時に記事「りそな銀行」の閲覧回数が急増していることがわかる．記事「りそなホールディングス」も閲覧回数が増加しているが，増加の割合は記事「りそな銀行」よりは低く，また，語句「りそな銀行」からリンクされる確率も低い．この場合，提案手法は語句「りそな銀行」のリンク先として記事「りそな銀行」を選択する．

4. 評価実験

4.1 実験環境

提案手法の性能を評価するために，実際のツイートを用いた実験を行った．提案手法により，入力のツイート集合

表 3 0.6 以上の Trendiness ϕ_{term} を持つ語句の数の分布（括弧内の数値はその割合を表す）

ϕ_{term}	2015 年 6 月 8 日のツイート集合中の語句数	評価用データセット中の語句数
0.6 以上 0.7 未満	2,746 (65.9%)	661 (66.1%)
0.7 以上 0.8 未満	959 (23.0%)	225 (22.5%)
0.8 以上 0.9 未満	318 (7.6%)	82 (7.6%)
0.9 以上	147 (3.5%)	32 (3.2%)

表 2 収集した各投稿日における日本語のツイート数

投稿日	ツイート数
2015 年 6 月 1 日	525,217
2015 年 6 月 2 日	571,996
2015 年 6 月 3 日	566,492
2015 年 6 月 4 日	507,012
2015 年 6 月 5 日	518,380
2015 年 6 月 6 日	597,668
2015 年 6 月 7 日	608,719
2015 年 6 月 8 日	518,864

から話題語とそれに対応する Wikipedia の記事を抽出し，

- 話題語がツイート中で言及している話題を表す語句として適切か否か
- 話題語に紐付けられた記事は適切か否か

という観点から評価を行い，提案手法が Twitter 上の話題語とそれに対応する適切な Wikipedia の記事を抽出できるかを検証した．

本実験では，Twitter Streaming API を用いて収集した 2015 年 6 月 1 日から 2015 年 6 月 8 日にかけて投稿された日本語のツイートを用いた．このうち，2015 年 6 月 8 日に投稿されたツイートの集合を提案手法の入力とし，直前 7 日間の 2015 年 6 月 1 日から 2015 年 6 月 7 日の間に投稿されたツイートの集合を Trendiness ϕ_{term} の算出のために用いた．収集したツイートの統計情報を表 2 に示す．

本実験では，以下の手順により，2015 年 6 月 8 日のツイート集合から 1,000 件の話題語とそれを含むツイートのペアを抽出し，評価用データセットとして用いた．

(1) 2015 年 6 月 8 日に投稿されたツイート集合において 5 回以上出現する語句の Trendiness ϕ_{term} を算出
(2) $\phi_{term}(a) \geq 0.6$ である語句 a を話題語としてランダムに 1,000 件抽出
(3) 抽出した各話題語について，話題語を本文中に含むツイートをランダムに 1 件抽出

表 3 に，2015 年 6 月 8 日のツイート集合および評価用データセットにおいて，0.6 以上の ϕ_{term} を持つ話題語の数の分布を示す．

本実験における正解集合を定義するために，作成したデータセットに対して三名の評価者による正解データの作成を行った．具体的には，データセットの各話題語・ツイートペアと提案手法によって話題語に紐付けた Wikipedia の記事を提示し，ツイート中の話題語に紐付けられた記事が適切かどうかを評価者がラベル（適切，不適切）付けした．また，ツイート中の話題語に紐付けられた記事が不適切であると判定された場合，その理由が，話題語に紐付けられた記事が不適切であるためか，または，話題語がツイート中で言及している話題を表す語として誤っているためか，のどちらであるのかについてもラベル（誤：記事，誤：話題語）付けした．最終的に，データセットの各ツイート・話題語・記事のデータについて，二名以上の評価者らにより付与されたラベルをそのデータのラベルとして用いた．

評価ではまず，データセット中で「適切」または「誤：記事」とラベル付けされた話題語を正解と定義し，提案手法が出力した話題語のうち，正解の話題語をいくつ含むかにより，提案手法の話題語抽出の適合率を算出した．次に，データセット中で「適切」とラベル付けされた話題語・記事ペアを，話題語に対応する記事の正解と定義した．そして，正解の話題語を含む提案手法の出力した話題語・記事ペアのうち，正解の話題語・記事ペアをいくつ含むかにより，話題語の曖昧性解消の適合率を算出した．最後に，提案手法のエンティティリンキングとしての性能を評価するために，データセット中で「適切」または「誤：記事」とラベル付けされた話題語・記事ペアを話題語とそれに対応する記事の正解として定義し，適合率および再現率を算出した．なお，ラベル「誤：記事」が付与された話題語・記事ペアについて，正解の記事は未定義（提案手法では対応する記事を決定できない）とした．本実験では，提案手法が特に ϕ_{term} の高い語句に対してエンティティリンキングを高精度に達成できているかを確認するために，話題語の定義に関するしきい値 τ を 0.6 から 0.9 まで 0.1 ずつ変化させ，しきい値 τ 以上の ϕ_{term} を持つ語句を話題語として評価を行った．

4.2 実験結果

4.2.1 話題語抽出の結果

図 3 は，提案手法が抽出した話題語の正誤のみを考慮した場合の適合率を表している．話題語抽出において，提案手法が話題語として抽出した語句は Trendiness ϕ_{term} が高い語句であるほど適合率が向上している．これは，ϕ_{term} の高い語句はツイート中で言及している話題を表す語として適切な語句であることがわかる．しかし，ϕ_{term} のし

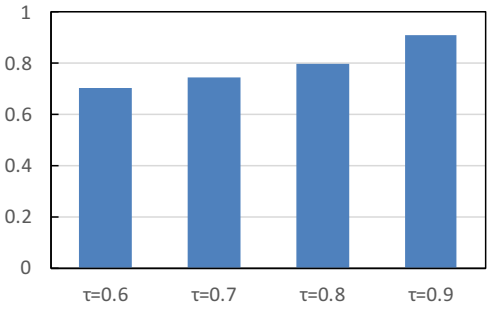

図 3　話題語抽出の適合率

表 4　ラベル「誤：話題語」が付与された語句

語句	ϕ_{term}	ラベル「誤：話題語」が付与された語句を含むツイート中の語句
バゴ	0.947	バゴォーン，バゴーン
えいさ	0.906	うんえいさん，はくえいさん
アダチ	0.897	アダチン
ツェレ	0.854	ツェレオ

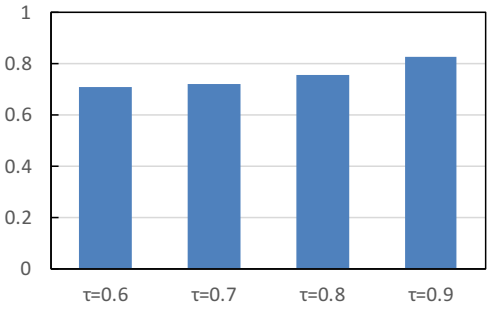

図 4　話題語の曖昧性解消の適合率

表 5　話題語とそれに誤って紐付けられた記事

話題語	ϕ_{term}	記事
大阪駅前	1.0	大阪駅・梅田駅周辺バスのりば
ハローグッバイ	1.0	ハロー!グッバイ
フェンリル	0.962	フェンリル

きい値 τ を増加させても適合率が 1 にならないことから，ϕ_{term} が高いにも関わらずテキスト中のキーワードとして不適切な語句が存在していることを示している．表 4 に，データセットにおいて ϕ_{term} は高いがラベル「誤：話題語」が付与された語句の一例を示す．ラベル「誤：話題語」が付与された語句について，実際にツイート中にどのような形で出現しているかを確認したところ，いずれのケースもツイート中のある語句の部分文字列であることがわかった．例えば，「バゴ」は「バゴォーン」あるいは「バゴーン」と呼ばれるカップ焼きそばの商品名，「アダチ」は東京都足立区のキャラクター「アダチン」の部分文字列であった．これは Wikipedia において，これらの名称に関する記事やアンカーテキストが定義されていないために生じた問題である．

4.2.2　話題語の曖昧性解消の結果

図 4 は，各話題語に対して提案手法が紐付けた記事の適合率を表している．話題語抽出と同様に，話題語の曖昧性解消における評価結果でも，語句の Trendiness ϕ_{term} のしきい値 τ の増加に応じて適合率が向上している．これは提案手法が，ϕ_{term} の高い話題語であるほど，その話題語に対応する適切な記事を決定できていることを意味している．しかし，話題語抽出の結果と同様に，しきい値 τ を増加させても適合率が 1 にならないことから，ϕ_{term} が高い話題語であっても，それに対応する適切な記事を決定できない話題語が存在していることを示している．

表 5 に，適切な記事を紐付けられなかった話題語と誤って紐付けた記事の例を示す．表 5 に示した語句について，実際にツイート中でどのような話題を表しているかを確認した．語句「大阪駅前」を含むツイートでは，大阪駅前の地下道にある老舗串かつ店の閉店について述べられていた．ツイートの内容は大阪駅前に関することであるため，データセット中では話題を表す語として正解とラベル付けされていたが，リンク先の記事が話題の内容と異なっていたために誤りと判定されたと考えられる．語句「大阪駅前」のリンク先の別の候補として，適切だと思われる「大阪駅」や「梅田」も存在した．しかし，記事の $Trendiness$ は記事「大阪駅・梅田駅周辺バスのりば」の 0.465 に対して，記事「大阪駅」は 0.509，記事「梅田」は 0.551 と差が小さく，また，リンク確率も低かったため，これらの記事にリンクされなかった．記事の $Trendiness$ のみを考慮した場合，記事「梅田」をリンク先として決定できるため，式 (4) で定義したリンク先の記事を決定するためのスコアの算出について，話題語によって記事の Trendiness とリンク確率の重みを考慮した方法を考案する必要がある．

語句「ハローグッバイ」や語句「フェンリル」は，ツイート中ではそれぞれ，歌手新山詩織の新しいアルバム名，および，ソウルズアルケミストというゲームのキャラクター名を表しており，リンク先の記事が話題語が表す話題の内容と異なったため誤りと判定されたと考えられる．また，語句「ハローグッバイ」および語句「フェンリル」をアンカーテキストとして，話題の内容と合致する記事が存在するかについて確認したが，該当する記事は存在しなかった．提案手法は，話題語に対して対応する記事が Wikipedia に存在するという前提を置いていたため，これら話題語に対して，適切なリンク先を見つけることができなかった．この問題に対しては，リンク先の記事が存在する話題語であるかどうかを確認する処理を導入し，話題語としては抽出するが，対応するリンク先の記事は存在しないものとして出力する必要がある．

4.2.3　エンティティリンキングの結果

図 5 は，話題語の正誤および話題語に紐付けられた記事の正誤の両方を考慮した場合，すなわちエンティティリン

表 6 語句の Trendiness ϕ_{term} とそれに対応する記事の Trendiness ϕ_{page} の関係

語句	ϕ_{term}	記事	ϕ_{page}	$\phi_{term} - \phi_{page}$
りそな銀行	0.987	りそな銀行	0.827	0.160
渡辺謙	0.970	渡辺謙	0.962	0.008
バゴ	0.947	バゴ (競走馬)	0.470	0.477
えいさ	0.947	えいさ	0.411	0.536
アダチ	0.897	足立製作所	0.646	0.251
ツェレ	0.897	ツェレ	0.552	0.345
大阪駅前	1.0	大阪駅・梅田駅周辺バスのりば	0.465	0.535
ハローグッバイ	1.0	ハロー!グッバイ	0.659	0.341
フェンリル	0.962	フェンリル	0.496	0.466

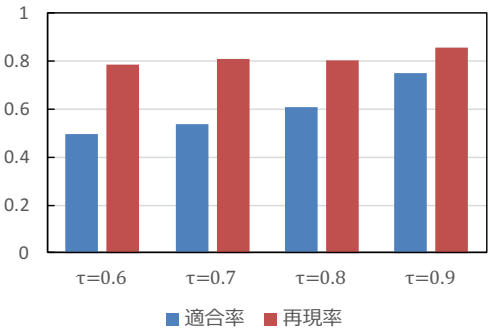

図 5 話題語に対応するエンティティリンキングの適合率と再現率

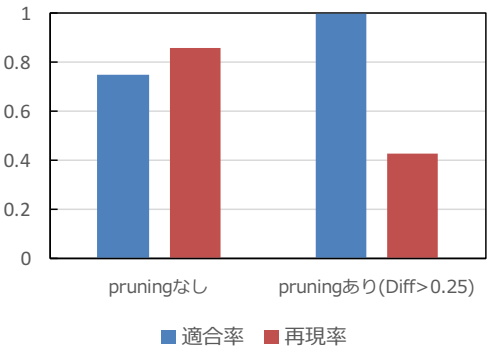

図 6 $\tau = 0.9$ における，ϕ_{term} と ϕ_{page} の差を用いた不適切な候補の除外（pruning）の結果

キング全体での処理の適合率と再現率を表している．4.2.1 項，4.2.2 項，および，図 5 の結果から，提案手法は，語句の Trendiness ϕ_{term} のしきい値 τ が大きくなればなるほど，Twitter 上の話題語とそれに対応する Wikipedia の記事を紐付けるエンティティリンキングをより高い精度で達成できることを確認した．また，表 4 に示すような話題語として不適切な語句や，表 5 のような話題語の候補やリンク先の記事を誤った候補，リンク先の記事が存在しない候補を出力から除外することができれば，提案手法の適合率をさらに向上できると考えられる．

4.2.4 不適切な出力の除外方法の検討

表 4 および表 5 に示した語句について，さらに詳しく調査すると，表 6 に示すように，語句の Trendiness ϕ_{term} とそれに紐付けられた記事の Trendiness ϕ_{page} が大きく異なることがわかった．一方，表 1 に示した，話題語とそれに対応する記事については，ϕ_{term} と ϕ_{page} の差が小さい傾向にあることがわかった．そこで，次の式で定義する ϕ_{term} と ϕ_{page} の差 $Diff$ を用いて，$Diff$ の値が大きい話題語とそれに対応する記事を誤った出力として除外 (pruning) した場合の提案手法の性能について評価した．

$$Diff(a \mapsto p_a) = \phi_{term}(a) - \phi_{page}(p_a) \qquad (5)$$

図 6 に，0.9 以上の ϕ_{term} を持つ話題語について，$Diff$ による pruning を行わない場合と $Diff > 0.25$ という条件で pruning を行った場合のエンティティリンキングの評価結果（適合率と再現率）を示す．$Diff$ を用いて pruning を行った場合，適合率が 100% になっており，話題語として不適切な語句，リンク先を誤った候補，および，リンク先の存在しない候補を全て除去できていることがわかる．一方，pruning を行わない場合と比較して，再現率が大幅に低下している．これは，データセットの正解集合に含まれる話題語・記事ペアについて，ϕ_{term} ほど ϕ_{page} が高くならないようなペアが多数存在することを意味している．

例えば，話題語「震度 4」は，2014 年 6 月 8 日に栃木県南部で発生した地震に関するツイート中に出現しており，ϕ_{term} も 0.955 と高い．しかし，リンク先として選ばれた記事「震度」は話題語「震度 4」のリンク先として適切であるラベル「適切」が付与された正解データであるが，記事の ϕ_{page} は 0.500 であった．これは，話題語について人々が興味を持ち，それについて調べたいと思うか否か，という話題語の特性に基づくと考えられる．例えば，地震の情報を伝えるツイートには，地震の発生日時や場所，規模についてすでに書かれており，そのツイートを見るだけでその地震に関する情報を把握できる．また，記事「震度」は震度の概念や定義について記述されているだけであるため，実際に起きた地震が話題になり，その地震の震度を調べる人が増えたとしても，記事「震度」は閲覧しないと考えられる．4.2.2 項で述べた話題語「大阪駅前」についても，記事の Trendiness のみを考慮して適切な記事「梅田」を紐付けたとしても，同様の理由により，出力から除外される．この

ような ϕ_{term} と ϕ_{page} の関係をもつ正解データを pruning の処理によって出力から除外した結果，pruning を行う場合の再現率が低下した．pruning を行うことで話題語として不適切な語句を全て除外できているため，ϕ_{term} と比較して ϕ_{page} が低いケースを分類することができれば，より高精度な話題語に対するエンティティリンキングを実現できると考えられる．

5. おわりに

本研究では，Twitter 上で話題となっている語句を抽出し，それに対応する Wikipeida の記事を紐付けるエンティティリンキング手法を提案した．提案手法では，ある期間に投稿されたツイートの集合を入力とし，入力のツイート集合の中で頻出する語句を話題を表す語句として抽出する．このとき，入力のツイート集合の投稿期間以前に投稿されたツイートの情報を用いることで，入力のツイート集合において特に頻出している語句を抽出する．そして，抽出した各話題語に対して，話題語からリンクされており，かつ，入力のツイート集合と同じ期間に多く閲覧されている Wikipedia の記事を紐付ける．Twitter のデータを用いた評価実験により，Twitter 上で話題である度合いが高い語句ほど，より高精度に対応する Wikipedia の記事を発見できることを確認した．

今後の課題として，話題語のリンク先の記事を決定するためのスコアの算出方法や，ϕ_{term} に対して ϕ_{page} が低いケースを分類し，話題語として不適切な語句の候補のみを除外する方法について検討する．また，筆者らの先行研究 [21] では，ツイート集合中の文字列に関する統計情報を用いた不適切な語句の除去手法を提案しており，本研究の提案手法と組み合わせることを検討している．そして，提案手法を用いた話題抽出のアプリケーションにおいて，抽出した話題の有用性を検証し，話題抽出における提案手法の有効性を検証することを予定している．

参考文献

[1] Aiello, L. M., Petkos, G., Martin, C., Corney, D., Papadopoulos, S., Skraba, R., Goker, A., Kompatsiaris, I. and Jaimes, A.: Sensing Trending Topics in Twitter, *IEEE Transactions on Multimedia*, Vol. 15, No. 6, pp. 1268–1282 (2013).

[2] Bedathur, S. J., Berberich, K., Dittrich, J., Mamoulis, N. and Weikum, G.: Interesting-Phrase Mining for Ad-Hoc Text Analytics, *Proceedings on the VLDB Endowment*, Vol. 3, No. 1, pp. 1348–1357 (2010).

[3] Chen, Y., Amiri, H., Li, Z. and Chua, T.-S.: Emerging Topic Detection for Organizations from Microblogs, *In SIGIR*, pp. 43–52 (2013).

[4] Cucerzan, S.: Large-Scale Named Entity Disambiguation based on Wikipedia Data, *In EMNLP-CoNLL*, pp. 708–716 (2007).

[5] Damerau, F. J.: Generating and Evaluating Domain-Oriented Multi-Word Terms from Texts, *Information Processing & Management*, Vol. 29, No. 4, pp. 433–447 (1993).

[6] Ferragina, P. and Scaiella, U.: Fast and Accurate Annotation of Short Texts with Wikipedia Pages, *IEEE Software*, Vol. 29, No. 1, pp. 70–75 (2012).

[7] Hoffart, J., Yosef, M. A., Bordino, I., Fürstenau, H., Pinkal, M., Spaniol, M., Taneva, B., Thater, S. and Weikum, G.: Robust Disambiguation of Named Entities in Text, *In EMNLP*, pp. 782–792 (2011).

[8] Hua, W., Wang, Z., Wang, H., Zheng, K. and Zhou, X.: Short text understanding through lexical-semantic analysis, *In ICDE*, pp. 495–506 (2015).

[9] Hua, W., Zheng, K. and Zhou, X.: Microblog Entity Linking with Social Temporal Context, *In SIGMOD*, pp. 1761–1775 (2015).

[10] Kulkarni, S., Singh, A., Ramakrishnan, G. and Chakrabarti, S.: Collective Annotation of Wikipedia Entities in Web Text, *In KDD*, pp. 457–466 (2009).

[11] Lesk, M.: Automatic Sense Disambiguation Using Machine Readable Dictionaries: How to Tell a Pine Cone from an Ice Cream Cone, *In SIGDOC*, pp. 24–26 (1986).

[12] Meij, E., Weerkamp, W. and de Rijke, M.: Adding Semantics to Microblog Posts, *In WSDM*, pp. 563–572 (2012).

[13] Mihalcea, R. and Csomai, A.: Wikify!: Linking Documents to Encyclopedic Knowledge, *In CIKM*, pp. 233–242 (2007).

[14] Milne, D. and Witten, I. H.: An Effective, Low-Cost Measure of Semantic Relatedness Obtained from Wikipedia Links, *In WIKIAI*, pp. 25–30 (2008).

[15] Milne, D. and Witten, I. H.: Learning to Link with Wikipedia, *In CIKM*, pp. 509–518 (2008).

[16] P, D., Dey, A. and Majumdar, D.: Fast Mining of Interesting Phrases from Subsets of Text Corpora, *In EDBT*, pp. 193–204 (2014).

[17] Piccinno, F. and Ferragina, P.: From TagME to WAT: A New Entity Annotator, *In ERD*, pp. 55–62 (2014).

[18] Salton, G. and Buckley, C.: Term-weighting Approaches in Automatic Text Retrieval, *Information processing & management*, Vol. 24, No. 5, pp. 513–523 (1988).

[19] Yamada, I., Takeda, H. and Takefuji, Y.: An End-to-End Entity Linking Approach for Tweets, *Proceedings of WWW Workshop on Making Sense of Microposts* (2015).

[20] 吉田光男，荒瀬由紀，角田孝昭，山本幹雄：検索頻度推定のための Wikipedia ページビューデータの分析，人工知能学会全国大会，pp. 2I1–1 (2015).

[21] 中村達哉，白川真澄，原　隆浩，西尾章治郎：ソーシャルメディアからの言語横断的な話題抽出に向けたエンティティリンキング手法，第 7 回データ工学と情報マネジメントに関するフォーラム (2015).