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

中村　達哉† 　白川　真澄† 　原　隆浩† 　西尾章治郎†

† 大阪大学大学院情報科学研究科　〒 565–0871 大阪府吹田市山田丘 1-5

E-mail: †{nakamura.tatsuya,shirakawa.masumi,hara,nishio}@ist.osaka-u.ac.jp

あらまし　本稿では，ソーシャルメディアのテキスト集合から言語横断的に話題を抽出するためのエンティティリンキング手法について述べる．ソーシャルメディアのような短いテキストから異なる言語間で比較可能な話題を抽出する場合，エンティティリンキングにより，テキスト中に出現するエンティティを多言語な知識体系のエントリに紐付けることが有効である．このとき，エンティティリンキングの精度およびトピック情報として付与されるエントリの意味的な粒度を均一化することが言語間のエントリの比較において重要となる．提案手法では，任意の言語で記述されたテキスト中に出現するエンティティに対して同じ言語の Wikipedia の記事をトピック情報として紐付けた後，言語間リンクにより英語の記事に変換することで言語空間を統一する．また，入力のテキスト集合中において出現するキーワードの周辺情報を用いた不適切なキーワード抽出の抑制と，記事タイトルやカテゴリ情報を用いた記事集約による意味的な粒度の均一化を行う．評価実験の結果から，既存手法と比較して提案手法が話題抽出の観点から有効であることを確認した．

キーワード　エンティティリンキング, ソーシャルメディア, 言語横断解析, 話題抽出

1. はじめに

文書集合に含まれるトピックを抽出する研究は数多く行われているが，近年，Twitter に代表されるソーシャルメディアがその対象として注目を集めている．その理由として，ソーシャルメディアの即時性 (リアルタイム性) が挙げられる．ソーシャルメディアでは，様々な人が実世界の出来事や自身の興味・関心についての情報を常時発信している．また最近では，官庁や報道機関等の公的な組織もソーシャルメディア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．このようなソーシャルメディアのテキストを解析することで，即時性が高いトピック情報を抽出できる．

ソーシャルメディアのもう一つの特徴として多言語性が挙げられる．例えば，Twitter は公式に 44 言語 (注1) に対応しており，ユーザの使用言語や居住地域に応じたトレンド情報 (話題になっている語句) をサービスとして提供している．この特徴は，多くの言語で話題になっているトピックや，自身の言語でのみ話題である (あるいは話題でない) トピック等，言語の壁を超えたトピック情報をソーシャルメディアから抽出できる可能性を示している．ソーシャルメディアから言語横断的にトピックを抽出することができれば，自分が使用できない言語のトピック情報を，その言語の知識なしに得ることができる．ソーシャルメディアのような，ユーザが自身の言語で情報発信を行う多言語なメディアにおいて，言語横断的にトピックを抽出することは有益であると考えられる．

しかし，このようなソーシャルメディアのテキストから言語横断的にトピックを抽出する場合，どのようにして異なる言語のテキスト集合からトピック情報を抽出するかが問題となる．言語によって使用される文字の種類が異なるため，テキスト中に出現する語句の統計情報を用いるような従来のトピック抽出手法によりそれぞれの言語について個別にトピック情報を抽出できたとしても，それらを異なる言語間で比較することは困難である．言語間でトピック情報を比較可能にするには，トピックの言語空間を統一する必要がある．

そこで筆者らは先行研究 [13] において，多言語な知識体系を用いたソーシャルメディアからの言語横断的な話題抽出手法を提案した．この手法では，テキスト中に出現するエンティティを表す語句を抽出しその語句を対応する知識体系のエントリに紐付けるエンティティリンキングとよばれる技術を，多言語な知識体系である Wikipedia を対象として用いる．一つトピックを一つの Wikipedia の記事として定義し，任意の言語で記述されたソーシャルメディアのテキスト中に出現する Wikipedia のアンカーテキストをキーワードとして抽出する．そして，各キーワードに対して入力テキストと同じ言語の Wikipedia の記事を紐付け，言語間リンクにより英語の記事に変換することで，言語空間が英語に統一されたトピック情報を付与する．これにより，異なる言語間でトピック情報を比較することが可能となる．エンティティリンキングにより言語横断的な話題抽出を実現する場合，誤ったキーワード抽出の抑制とテキストに付与される記事 (トピック) の意味的な粒度を考慮する必要がある．話題抽出は入力のテキスト集合中で話題になっている (出現回数の多い) トピックを抽出することが目的であるが，キーワード抽出において，キーワードとして不適切な語句の抽出を繰り返すと，話題でないトピックの出現回数が増加し，話題抽出の精

(注1)：2015 年 1 月時点．ユーザ情報設定画面において確認 (Beta 版含む)．

度が低下する．この問題は特に，本来のキーワードの部分文字列を誤って抽出する場合に発生する．また，エンティティリンキングではキーワードに対して正解となるリンク先の記事候補が粒度の違いにより複数存在することがある．同じ話題を表すキーワードに対して意味的な粒度が異なる記事が付与されると，それぞれが異なるトピックを表す記事として扱われてしまい，本来話題であるトピックを抽出できないという問題が生じる．その結果，言語間でトピック情報を比較する際に，言語間で共通の話題があったとしても，その話題を別々のトピックとして抽出してしまう．先行研究では既存の単一のテキストを対象としたエンティティリンキング手法をそのまま用いたため，これらの問題が考慮されていなかった．

そこで本研究では，これらの問題に対応したエンティティリンキング手法を提案する．提案手法では，キーワード抽出において，テキスト中のアンカーテキストの前後に出現する文字の統計情報を事前に集計することで，テキスト中のキーワードがアンカーテキストとして定義されていない場合に，本来のキーワードの部分文字列をキーワードとして誤って抽出することを抑制する．また，記事タイトルや Wikipedia のカテゴリ情報を用いてエンティティリンキングの対象となる記事を集約することで，テキストに紐付けられる記事の意味的な粒度の均一化を目指す．

2. 関 連 研 究

エンティティリンキングに関する研究は Mihalcea らの Wikify! [8] 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，1) キーワード抽出，2) エンティティの曖昧性解消の順に行われる．

Wikify! [8] では，ある語句が Wikipedia においてリンクとして出現する度合いを表すスコア (keyphraseness) を定義し，keyphraseness が TF–IDF [11] などの語句の重み付け手法よりも高い精度で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では，Lesk アルゴリズム [6] を用いた手法と，キーワードの品詞やその前後に出現する 3 語などを素性とした Naive Bayes を用いた手法を組み合わせることで，高精度なエンティティリンキングを実現している．Cucerzan の研究 [2] 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 [10] は，機械学習を用いたエンティティリンキング手法 Wikipedia Miner(注2) を提案している．はじめに，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) アンカーテキスト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 [9] 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，Wikify!より高い精度でエンティティリンキングを達成している．

Kulkarni らの研究 [5] 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おり，評価実験において Cucerzan らの手法や Milne らの手法より高い精度のエンティティリンキングを実現している．Hoffart ら [4] 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 AIDA(注3) を提案している．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た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，抽出した部分グラフを用いて，各キーワードに対応するエントリを決定することでエンティティリンキングを実現している．

ソーシャルメディア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．Ferragina ら [3] は，短いテキストを対象としたエンティティリンキング手法 TAGME(注4) を提案している．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シンプルな処理で，高速かつ精度の高いエンティティリンキングを実現している．Meij らの研究 [7] 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，またそれら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 のツイートを用いた評価実験において，Wikipedia Miner や TAGME と比較して高い精度を達成している．

短いテキストを対象とした手法の特徴として，エンティティリンキングの処理速度が挙げられる．一般的なエンティティリンキングに関する研究では，手法の精度を確保するために品詞分類や最適化問題を用いているため，短文を対象とした手法よりも低速である [1]．例えば AIDA は，15 個のキーワードを含むテキストに対して 2 秒以上の処理時間を必要とする [4]．一方，TAGME は一つのキーワードあたり 2 ミリ秒以下と非常に高速である [3]．エンティティリンキングの処理速度は，日々大量に投稿されるソーシャルメディアのテキストを処理する上で非常に重要な要素である．また，本研究で想定している様々な言語のテキストに対するエンティティリンキングを実現する場合，品詞分類や機械学習を用いる手法では，対応する言語の増加に従って，品詞分類や機械学習のための教師データを用意することが困難になる．そこで本研究では，品詞分類や機械学習を必要とせず，かつ，高速な手法である TAGME を，Wikipedia の言語間リンクを用いて多言語的に拡張することで，任意の言語

(注2)：http://wikipedia-miner.cms.waikato.ac.nz/

(注3)：http://www.mpi-inf.mpg.de/yago-naga/aida/

(注4)：http://tagme.di.unipi.it/

で記述されたテキストに対するエンティティリンキングを実現する．

3. 提案手法

本章ではまず，単一言語の短いテキストを対象としたエンティティリンキング手法である TAGME [3] について説明する．その後，ソーシャルメディアのテキスト集合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を実現する上で問題となる点について述べる．そして，TAGME を Wikipedia の言語間リンクにより拡張し，任意の言語で記述されたテキストに出現するエンティティに対して英語の Wikipedia の記事を紐付ける手法について述べる．

3.1 TAGME

TAGME [3] は，単一言語で記述された短いテキストに対して，テキスト中に出現するエンティティを入力テキストと同じ言語の Wikipedia の記事に紐付ける手法である．TAGME は，1) キーワード抽出，2) キーワードの曖昧性解消，3) 確信度の低いエンティティの除去の各処理によって高速なエンティティリンキングを実現している．以下では，TAGME の各処理について詳しく説明する．

3.1.1 キーワード抽出

TAGME は，テキストを入力として受け取ったあと，テキスト中に出現する Wikipedia のアンカーテキストとして用いられている全ての語句をキーワードとして抽出する．ここで，あるキーワード a_1 が別のキーワード a_2 の部分文字列である場合，それぞれのキーワードが Wikipedia の記事中でアンカーテキストとして使われる確率を $lp(a_1)$，$lp(a_2)$ として，次の処理を行う．

- $lp(a_1) < lp(a_2)$ の場合，a_2 のみをキーワードとして抽出する．
- $lp(a_1) \geqq lp(a_2)$ の場合，a_1 と a_2 の両方をキーワードとして抽出する．

ここで，$lp(a)$ はキーワード a のリンク確率と呼ばれ，以下の式から算出される．

$$lp(a) = \frac{link(a)}{freq(a)} \tag{1}$$

$link(a)$ は，Wikipedia の全ての記事においてキーワード a がアンカーテキストとして出現する回数，$freq(a)$ は Wikipedia 内でキーワード a が出現する回数である．

提案手法では，リンク確率の代わりに keyphraseness [8] を用いる．Wikipedia では，記事中にアンカーテキストの候補となる語句が複数回出現する場合，最初に出現した語句のみアンカーテキストとして定義し，それ以降に出現する語句はアンカーテキストとして定義しないことが一般的である．そのため，語句 a が重要であるほど，記事中に語句 a がアンカーテキストとして出現する回数 $link(a)$ に対して語句 a が出現する回数 $freq(a)$ が大きな値となり，実際には重要な語句であるにも関わらず，リンク確率 $lp(a)$ が低くなるという問題がある．一方，keyphraseness は，語句 a がアンカーテキストとして出現する記事数 $lf(a)$ および語句 a が出現する記事数 $df(a)$ を用いて以下の式

$$keyphraseness(a) = \frac{lf(a)}{df(a)} \tag{2}$$

により算出されるため，記事中で語句 a がアンカーテキストとして出現し，かつ，複数回出現するような場合であっても，keyphraseness はその影響を受けにくい．

3.1.2 キーワードの曖昧性解消

次に，テキスト中の各キーワード $a \in A$ について，そのキーワードによってリンクされる記事の集合 $Pg(a)$ のうち，キーワードがどの記事 $p_a \in Pg(a)$ を表しているかを決定する．キーワード a が記事 p_a にリンクされる確信度を表すスコアは，voting scheme と呼ばれる以下の式によって算出される．

$$rel_a(p_a) = \sum_{b \in A \setminus \{a\}} \frac{\sum_{p_b \in Pg(b)} rel(p_b, p_a) \cdot Pr(p_b|b)}{|Pg(b)|} \tag{3}$$

ここで，$rel(p_b, p_a)$ は記事間関連度 [9] を表し，$Pr(p_b|b)$ はキーワード b がアンカーテキストとして使われる際に記事 p_b にリンクされる確率を表しており commonness と呼ばれる．各キーワードによってリンクされる記事すべてを対象として式 (3) を算出することは，計算コストの面で非効率的であるため，$Pr(p_b|b) > \tau$ を満たす記事のみを対象とする．式 (3) は，記事 p_a が他のキーワードによってリンクされる記事と関連が強いほど高い値となる．そして，キーワード a によってリンクされる各記事について，式 (3) のスコアを高い順に並べた際の上位 ϵ% の記事のうち，commonness の最も高い記事 p_a をキーワード a が示す記事として採用する．提案手法では，キーワードの曖昧性解消におけるしきい値 τ および ϵ として，TAGME [3] で用いられている $\tau = 0.02$，$\epsilon = 30\%$ を用いた．

3.1.3 確信度の低いキーワード・記事ペアの除去

3.1.2 項までの処理によって，テキスト中に出現する全てのキーワードに対応する Wikipedia の記事を紐付けることができる．しかし，入力テキスト中に出現する全てのアンカーテキストをキーワードとして抽出しているため，入力テキストの内容にあまり関係のない語句がキーワードとして抽出されている可能性がある．そこで，3.1.2 項までの処理によって得られたキーワード・記事ペア (a, p_a) について，

$$\rho(a, p_a) = \frac{1}{2}(lp(a) + coherence(a, p_a)) \tag{4}$$

を算出し，最終的に $\rho(a, p_a) > \rho_{NA}$ を満たすキーワード a のみに対して記事 p_a をリンクする．$coherence(a, p_a)$ は次式によって算出される．

$$coherence(a, p_a) = \frac{1}{|S| - 1} \sum_{p_b \in S \setminus \{p_a\}} rel(p_b, p_a) \tag{5}$$

S は式 (4) の計算の対象となる全ての記事集合であり，式 (5) は候補の記事が互いに関連しているほど高い値となる．提案手法では 4 章で説明する評価実験により，しきい値 ρ_{NA} を決定した．

3.2 考慮すべき問題

本研究では，多言語なソーシャルメディアからの言語横断的なトピック抽出を最終目標として，言語間で比較可能なトピック情

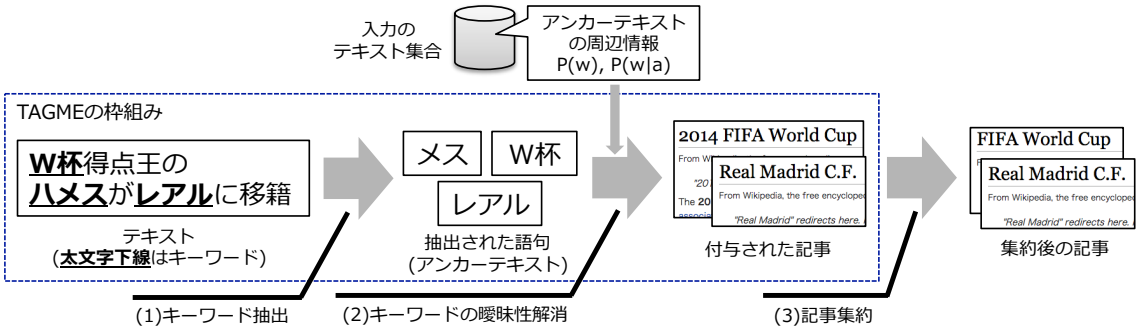


図 1　提案手法の流れ

報として Wikipedia の記事を用いることを考える．Wikipedia では同一のエンティティを表す異なる言語の記事は言語間リンクで繋がっているため，エンティティリンキングにより各言語のテキストに記事を付与した後，言語間リンクを用いることで容易にトピックの言語空間を統一できる．エンティティリンキングにより，任意のテキスト集合に対して言語空間が統一されたトピック (本研究では英語の Wikipedia の記事) を付与し，テキスト集合中の話題を抽出する上でいくつかの問題が存在する．

まず，不適切なキーワードによる誤った話題の抽出が挙げられる．TAGME では，キーワード抽出において，Wikipedia のアンカーテキストなど事前に準備されたキーワード辞書を用いているため，辞書に含まれないキーワードを抽出することができない．辞書に含まれない語句がテキスト中にキーワードとして出現する場合，そのキーワードを抽出できないか，もしくは，そのキーワードの部分文字列を誤ったキーワードとして抽出してしまうため，エンティティリンキングの精度が低下する．例えば日本語では，サッカー選手の「ハメス・ロドリゲス」を表す語句として「ハメス」が用いられるが，日本語の Wikipedia において「ハメス」はアンカーテキストとして定義されていない．そのためキーワード抽出において,「ハメス」の部分文字列であり，かつ，アンカーテキストとして定義されている「メス」がキーワードとして誤って抽出される問題が発生する．特に,「ハメス」が話題として頻出する場合,「メス」が大量に抽出されてしまうため,「メス」が誤って話題として抽出される．このように，話題として抽出されるべきキーワードを誤って抽出することは，話題抽出において大きな問題となる．

付与する記事の意味的な粒度についても考慮する必要がある．例えば Wikipedia では FIFA ワールドカップについて，FIFA ワールドカップという概念と実際に開催された FIFA ワールドカップがそれぞれ「FIFA World Cup」と「2014 FIFA World Cup」のように個別の記事として定義されている．それぞれの記事が FIFA ワールドカップに関するトピックとして適切であっても，異なるエンティティを表すものとして扱われるため，それらの記事を同一のトピックに関する記事として扱うことができない．また，このような記事は内容が類似しており，エンティティリンキングにおいて区別して記事を付与することが困難である．テキストに対して意味的な粒度が異なる記事が分散して付与されると，本来話題であるトピックを話題として抽出することが難しくなる.「FIFA World Cup」と「FIFA World Cup 2014」のように，一方が他方を包含するようなエンティティを表す記事である場合,「FIFA World Cup 2014」を「FIFA World Cup」に集約しても元の情報はほとんど失われないため，話題抽出においては記事の集約により意味的なばらつきを抑えたほうが良い．

3.3　提案手法の概要

提案手法では，3.1 節で説明した TAGME を Wikipedia の言語間リンクによって拡張することで，任意の言語で記述されたソーシャルメディアのテキストに対して，英語の Wikipedia の記事の付与を実現する．図 1 に提案手法の流れを示す．提案手法による多言語拡張では，テキスト集合中のあるテキストに対し，同じ言語の Wikipedia の記事を TAGME により付与した後，その記事が英語の記事への言語間リンクを持っている場合は英語の記事に変換し，言語間リンクを持っていない場合は言語特有のトピックを表すものとしてそのまま用いる．これにより，任意の言語で記述されたソーシャルメディアのテキストに対し，英語の Wikipedia の記事を付与できる．

また，3.2 節で述べた二つの問題に対し，キーワードの周辺情報を用いたキーワード抽出の改良および記事集約によるトピックの意味的な粒度の均一化を行う．以下では，これらの処理について詳しく説明する．

3.4　キーワードの周辺情報を用いたキーワード抽出の改良

提案手法では，入力のテキスト集合中の各キーワード候補について，その前後に出現する文字列の統計情報を事前に集計し，テキスト中のキーワード候補が適切であるかどうかを判断するスコアを導入する．これにより，誤ったキーワード抽出を抑制する．

再度，サッカー選手「ハメス・ロドリゲス」を表す語句「ハメス」を例に説明する．入力のテキスト集合中に出現するアンカーテキスト「メス」の前後の文字の統計情報を集計し,「メス」の前一文字に「ハ」が高い確率で出現しており，かつ，それ以外で「ハ」がアンカーテキストの前一文字として出現する確率が低い場合,「ハ」と「メス」により異なるキーワードが構成されていると考えられる．つまり，入力テキスト中に語句「ハメス」が出現する場合,「メス」をキーワードとして抽出するのは不適切であると判断できる．このような処理は非常にシンプルであるが，言語に非依存的な処理として，様々な言語のテキストに対して容易に適用できるという利点がある．

入力のテキスト集合全体において，ある語 w[(注5)] がアンカー

(注5)：ここで語とは，英語のような分かち書きされている言語の場合は一単語，日本語のような分かち書きされていない言語の場合は一文字を表す．

テキスト a の前 (あるいは後) の語として特徴的に出現する度合い $con(w,a)$ は，語 w がアンカーテキスト a の前 (あるいは後ろ) の語として出現する確率 $Pr(w|a)$ と語 w が任意のアンカーテキストの前 (あるいは後) の語として出現する確率 $Pr(w)$ から算出できる．

$$con(w,a) = \frac{Pr(w|a)}{Pr(w)} \tag{6}$$

なお，式 (6)は $\sum_w con(w,a) = 1$ となるように正規化する．$Pr(w|a)$ および $Pr(w)$ はそれぞれ以下の式から求められる．

$$Pr(w|a) = \frac{count(w,a)}{\sum_{w'} count(w',a)} \tag{7}$$

$$Pr(w) = \frac{\sum_a count(w,a)}{\sum_{w',a} count(w',a)} \tag{8}$$

ここで $count(w,a)$ は，語 w がアンカーテキスト a の前 (あるいは後) の語として出現する回数である．式 (6) の値は，語 w がアンカーテキスト a の前 (あるいは後) の語として特徴的に出現しやすいほど高い値となる．つまり，式 (6) の値が大きいとき，アンカーテキスト a は，語 w を含むあるキーワードの部分文字列として出現している可能性が高い．

最終的に，式 (4) に式 (6) を組み込み，入力テキスト中のアンカーテキスト a とその前 (あるいは後) の語が w であるとき，

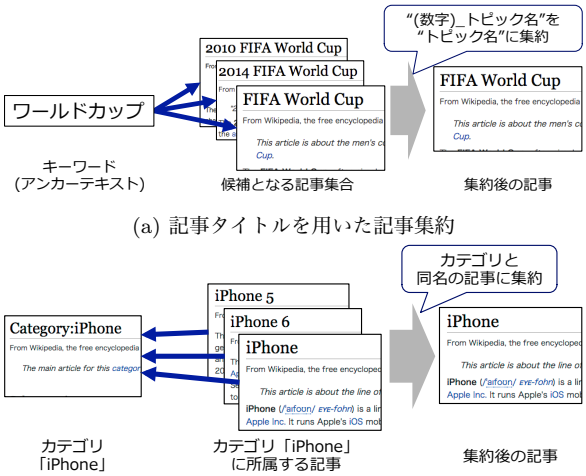
$$\rho(w,a,p_a) = \frac{lp(a) + coherence(a,p_a) + (1 - con(w,a))}{3} \tag{9}$$

を算出し，$\rho(w,a,p_a) > \rho_{NA}$ を満たすキーワード a のみに対して記事 p_a をリンクする．$con(w,a)$ には語 w がアンカーテキストの前一語と後一語の二通りがあるが，ここでは誤ったキーワード抽出を回避できればよいため，前一語と後一語のうち $con(w,a)$ の値が大きい方を用いればよい．ただし，入力のテキスト集合中において出現回数が少ないアンカーテキストは，その周辺に出現する語の統計情報を十分に収集できないため，式 (6) の値を用いることが不適切である場合がある．提案手法では，入力のテキスト集合中で出現回数が 20 回以上のアンカーテキストについて式 (6) を算出し，式 (9) を用いてエンティティリンキングを行う．出現回数が 20 回未満のアンカーテキストについては，TAGME の式 (4) をそのまま用いる．

3.5 記事集約によるトピックの意味的な粒度の均一化

エンティティリンキングでは，テキスト中のキーワードに対し，正解として適切な記事の候補が，意味的な粒度の違いにより複数存在することがある．これらの粒度を均一化し一つの記事に集約することで，同一のトピックに対して付与される記事が分散しないようにする．

このような記事の例として，「FIFA World Cup」と「2014 FIFA World Cup」のように定期的に開催されるイベントに関する記事や，「Microsoft Windows」や「iPhone」のような様々なバージョンが存在する製品に関する記事が挙げられる．前者の例では，記事タイトルが「(数字)␣トピック名」という形式を取ることが多い．後者の例では，製品の概要に関する記事のタイトルと同名のカテゴリが存在し，製品の各バージョンに関する記事はそのカテゴリに属している傾向にある．本研究ではこれらの特徴を利用し，英語の Wikipedia において，記事タイトルを用いた記事集約と，記事が所属するカテゴリ情報を用いた記事集約を行う (図 2)．集約された記事情報を 3.4 節で説明した処理の出力に対して適用することで，トピック情報の意味的な粒度を均一化する．



(a) 記事タイトルを用いた記事集約

(b) カテゴリ情報を用いた記事集約

図 2　記事集約の例

3.5.1 記事タイトルを用いた記事集約

記事タイトルを用いた方法では，図 2(a) に示している記事「2014 FIFA World Cup」のように「(数字)␣トピック名」の形式を取る記事を集約の対象とする．ここで，ある記事 p_x のタイトルを $title_x$，「(数字)␣トピック名」という記事タイトルに対してマッチする正規表現を $re_{withyear}$，西暦を除いた記事タイトルを返す関数を $extract(title_x)$ としたとき，集約の対象となる記事の条件は以下の通りである．

（1） 記事 p_x は，正規表現 $re_{withyear}$ にマッチする記事タイトル $title_x$ を持つ．

（2） $title_x$ から西暦を除いた記事タイトル $extract(title_x) = title_y$ を持つ記事 p_y が存在する．

条件 (1) および (2) を満たすとき，記事 p_x を記事 p_y に集約する．最終的に，35,847 記事が 8,240 記事に集約された (注6)．なお，記事タイトルを用いた記事集約において集約の対象とならなかった記事は，次項で説明するカテゴリ情報を用いた記事集約の方法を適用する．

3.5.2 カテゴリ情報を用いた記事集約

Wikipedia では各記事が所属するカテゴリ情報が定義されており，カテゴリ情報を子カテゴリから親カテゴリの方向へ辿ることで，ある記事が所属する様々な意味的な粒度のカテゴリ情報を収集できる．また，記事タイトルと同名のカテゴリを持つような記事は，同名のカテゴリに属する他の記事を子として持つような親記事であると考えることができる (図 2(b))．

そこで本研究では，記事タイトルを用いた記事集約に加えて，カテゴリ情報を用いた記事集約により，エンティティリンキングにより付与される記事の意味的な粒度の均一化を図る．なお，

(注6)：本研究では，2014 年 11 月 06 日に公開されたバージョンの英語の Wikipedia を用いた．

表 1　カテゴリを用いた記事集約の結果

ホップ数 k	集約の対象となる記事数	集約後の記事数
1	439,955	73,478
2	815,458	83,518
3	1,076,751	84,587

単純にカテゴリ情報を用いると，著名人やスポーツ選手など，記事タイトルと同名のカテゴリを持たないが記事単体でトピックとして適切な粒度を持つ記事を，所属カテゴリと同名のタイトルを持つ記事に集約してしまう．過剰な集約を避けるため，記事が所属するカテゴリのうち，記事タイトルの一部を含むカテゴリのみを，集約先の記事を探索する際の候補として用いる．記事 p_x が所属するカテゴリ集合を $categories(p_x)$，カテゴリ c_x のカテゴリ名を $cattitle_x$，c_x の親カテゴリを $c_{parent}(c_x)$ としたとき，記事 p_x に対して以下の処理を行う．

（1） 記事 p_x が所属するカテゴリ集合 $categories(p_x)$ 中に，$title_x$ の一部をタイトル名に含むカテゴリ $c_x \in categories(p_x)$ が存在する．

（2） $cattitle_x = title_y$ を満たす記事 p_y が存在する場合，p_x を p_y に集約する．

（3） 存在しない場合，しきい値を k として，k ホップ目まで幅優先探索により上記の処理を $c_{parent}(c_x)$ に対して再帰的に行う．k ホップ目までに見つからない場合，p_x は集約しない．

なお，同じ階層に集約先の候補が複数見つかる場合，Wikipedia 内におけるカテゴリ ID 番号が最も小さい候補を集約先として選択する．

表 1 に，ホップ数のしきい値 k を変化させた時の集約結果について示す．

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，Twitter のデータ (ツイート) を用いたエンティティリンキングの精度について実験を行った．提案手法により各ツイートに対してエンティティリンキングを行い，

- 誤ったキーワードの抽出を抑制できているか
- テキスト中のキーワードに対して適切な Wikipedia の記事を付与できているか

という観点から評価を行い，提案手法におけるキーワード抽出の改善が有効に機能しているか，および記事タイトルとカテゴリ情報を用いた記事集約がエンティティリンキングの性能を低下させていないかを検証した．

本実験では，Shirakawa らの研究 [12] において 2014 年 11 月 1 日から 2015 年 1 月 15 日にかけて収集された，メディアやジャーナリストが発信している英語と日本語のツイートを用いた．本実験で使用したツイート集合の統計情報を表 2 に示す．

収集したツイート集合から，英語と日本語のツイートをそれぞれ 300 件抽出したものを評価用データセットとして用いた．このとき，提案手法におけるキーワード抽出の改善が，話題になっている (入力のツイート集合の中で出現頻度の高い) キーワードに対して有効に機能するかを評価するために，ツイート集合中のアンカーテキスト (キーワード候補) の出現頻度を考慮してデータセットを作成した．具体的な作成手順として，それぞれ言語において，ツイート集合中に出現するアンカーテキストの出現頻度分布を対数軸上で三つの区間 (高頻度，中頻度，低頻度) に等分割し，各区間からその区間に属するアンカーテキストを含むツイートをランダムに 100 件ずつ抽出した．

表 2　使用したデータセットの統計情報

言語	ツイート数	1 ツイート当たりの平均語数 (注5)
英語	647,937	13.0
日本語	91,219	55.1

本実験におけるエンティティリンキングの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．はじめに，データセット中の各ツイートに対して，評価者らにより手動でツイート中のキーワードを抽出した．ここで，ツイート中のどの語句をキーワードとして定義するかは評価者によって異なるため，本評価では正解キーワードについて評価者間での集約を行った．具体的には，まず，評価者らが抽出したキーワードの中から最長のキーワードを順に選択する．次に，他の評価者が抽出したキーワードの中から，最長のキーワードと出現位置が重複するものを全て列挙する．そして，それらのキーワードを語単位 (注5) に分割し，二名以上の評価者によりキーワードとして抽出された語が連続する領域を正解キーワードとして定義した．次に，データセットの各ツイートに対して提案手法および比較手法を適用し，ツイート中の各キーワード候補 (アンカーテキスト) に付与された記事が適切かどうかを同じ三名の評価者らがラベル (正解，不正解) 付けした．また，キーワード候補に付与された記事が不正解であると判定された場合，その理由が，付与された記事が不適切であるためか，または，キーワード候補がツイート中のキーワード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:キーワード) 付けした．最終的に，評価者らにより手動で抽出したキーワード，および，各手法の出力において「正解」または「誤:記事」とラベル付けされたキーワード候補をデータセットにおけるキーワードの正解集合として定義した．エンティティリンキングの正解集合には，各手法の出力において「正解」とラベル付されたキーワード・記事ペアの集合と，各手法において抽出できなかった正解キーワードについてリンク先の記事を未定義としたキーワード・記事ペアの集合を用いた．

データセットの正解集合の作成において，エンティティリンキングに対する正解の記事を一意に定義しない理由は，曖昧性の高いキーワードに対して正解となる記事の定義が難しいことに加えて，提案手法と比較手法とで付与される記事の種類が異なるためである．例えば，2014 年の FIFA ワールドカップに関するツイート中に出現する「ワールドカップ」というキーワードに対して，記事「FIFA World Cup」と記事「2014 FIFA World Cup」のどちらがより適切であるかはアプリケーション依存であるため，本実験ではどちらも適切に付与された

記事として取り扱う．

提案手法として，キーワードの周辺情報のみを考慮し記事の集約を行わない手法 (hop0)，および，キーワードの周辺情報と記事集約の両方を考慮した際にカテゴリ情報を用いた記事集約におけるホップ数 k を $k = 1$ および $k = 2$ と変化させた手法 (hop1，hop2) を用いた．提案手法におけるキーワードの周辺情報に関するスコア (式 (6)) の算出には，表 2 に示した全ツイートを用いた．比較手法には，TAGME [3] の出力に対して言語間リンクのみを適用した手法について，リンク確率を用いた手法 (TAGME(LP)) と keyphraseness を用いた手法 (TAGME(KP)) を採用した．評価指標には，Micro における適合率 (Precision)，再現率 (Recall) を用いた．本実験では，提案手法と比較手法のそれぞれについて，式 (9) および式 (4) のしきい値 ρ_{NA} を 0 から 1 の間で変化させた際の各評価指標を算出した．

4.2 実験結果

実験結果を図 3 および図 4 に示す．図 3 は，各手法によって出力された各ツイートに対するキーワード・記事ペアについて，キーワードの正誤のみを考慮した場合の適合率と再現率の関係を表しており，各手法のキーワード抽出に関する性能を表している．キーワード抽出では，提案手法の hop0，hop1，hop2 の間で抽出されるキーワードに違いがない (付与される記事は異なる) ため，図 3 では hop1 の結果のみを示している．図 4 は，各手法の出力について，付与された記事の正誤についても考慮した場合の適合率と再現率の関係を表しており，エンティティリンキングにおける各手法の性能を表している．

キーワード抽出において，提案手法は低い再現率では TAGME と同等，高い再現率では TAGME より高い適合率を達成している．図 5 に，提案手法 (hop1) と TAGME(KP) について，しきい値 ρ_{NA} を変化させた際の評価ラベル数の変化を示す．TAGME では，誤りのラベルが付けられた出力のキーワード・記事ペア数が減少するに従って，正解ラベルが付与されたペア数も減少している．一方，提案手法では，$\rho_{NA} = 0.2$ からラベル「誤:キーワード」が付与されたキーワード・記事ペア数が減少しているが，正解ラベルのペア数や，ラベル「誤:記事」が付与されたペア数は減少していない．$\rho_{NA} = 0.3$ において，提案手法が出力の候補から削除した「誤：キーワード」のラベルが付けられた語句について見ると，表 2 のツイート集合中での出現回数が多い順に，日本語では「リタ」や「リタス」,「編集」,「すす」など，英語では「free mobile」や「pissed」「descend」といった，キーワードとして不適切な語句が実際に削除されていた．語句「リタ」や「リタス」はツイート中で「ポリタス」の部分文字列として出現していた.「ポリタス」は選挙に関する Web サイトであるが，Wikipedia のアンカーテキストとして定義されていない．また，語句「編集」は「○○編集部」,「すす」は「おすすめ」の形で多く出現しており，それぞれの語句の前後一語が特徴的に出現していたため，提案手法においてキーワードとしての重要性が低くなっていた．また,「free mobile」はツイート中で「free mobile app」の部分文字列として出現していた．この結果から，提案手法で導入したキーワードの周辺情報

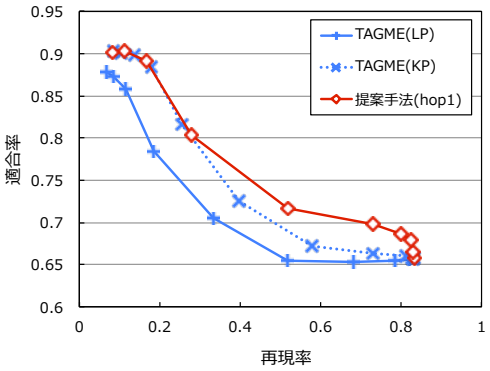


図 3　キーワードの正誤のみを考慮した場合の適合率・再現率

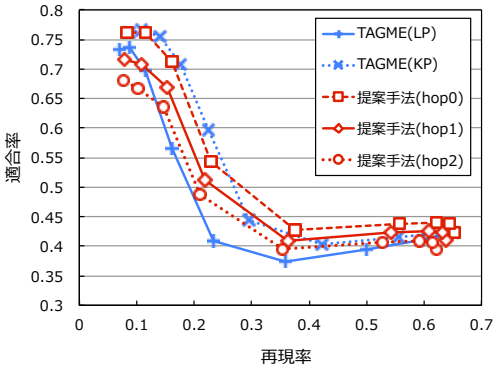
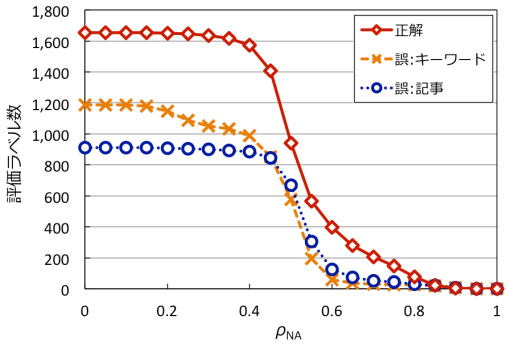


図 4　記事の正誤を考慮した場合適合率・再現率

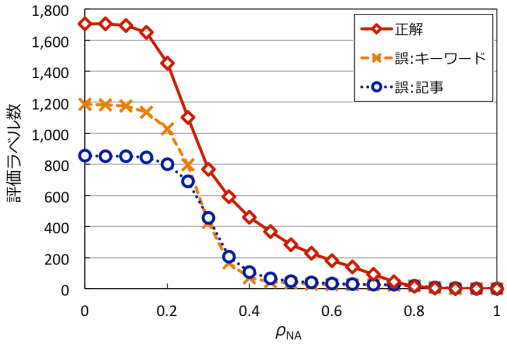
に関するスコアは，キーワードとして不適切な語句の抽出を抑制するのに有効に機能していると考えられる．TAGME(KP) および TAGME(LP) を比較すると，keyphraseness を用いた TAGME の方が適合率と再現率のどちらも高い値となっている．これは，キーワードの重み付けとして keyphraseness がリンク確率より優れていることを表している．

エンティティリンキングにおける評価では，再現率が高いときは提案手法 (hop0，hop1) が，再現率が低いときは TAGME(KP) が高い適合率を達成している．再現率が高い場合，提案手法は TAGME よりも，キーワードとして不適切な語句の抽出を抑制できているため，エンティティリンキングにおいても高い適合率を達成できている．図 5 におけるラベル数の変化を見ると，提案手法はキーワードが適切である場合，付与された記事が誤りであったとしても式 (9) のスコア ρ が高い傾向にある．これは，式 (6) のキーワードの周辺情報に関するスコア con は，キーワードが適切な場合，紐付けられた記事に関わらずその値が小さくなることが原因である．そのため，提案手法は再現率が低い場合でも，ラベル「誤:記事」が付与された候補が多く残っており，TAGME(KP) より低い適合率になったと考えられる．キーワードの曖昧性解消において，提案手法は TAGME と同じ処理を行っているため，キーワードの曖昧性解消の処理を改善できれば，誤ったキーワード抽出を抑制できる提案手法の方が高い性能を達成できると考えられる．

提案手法の hop0，hop1，hop2 をそれぞれ比べると，記事の集約を行わない hop0 が最も高い性能を達成している．そこで，hop1 と hop2 について，どのような記事が性能を低下させる原因となっているかを調査した．hop1 では，例えば，アイス



(a) 提案手法 (hop1)



(b) TAGME(KP)

図 5　ρ_{NA} を変化させた場合の評価ラベル数の変化

スケートに関するツイート中に出現する語句「ice rink」に記事「Ice Hockey」が紐付けられていたために誤りとなっていたケースがあった．これは，提案手法の記事集約において，集約先の候補が複数ある場合，Wikipedia 内における ID が最も小さいカテゴリと同名の記事を集約後の記事として一意に決定していることが原因である．一方，ツイート中の語句「Xperia Z3」に対して記事「Sony Xperia」が付与されるなど，適切に記事を集約できている場合もあった．hop2 では，集約の前後の記事で意味的な粒度が大きく変わっているために不正解とラベル付けされたケースが見られた．

実際のツイート集合に対して，どの程度の記事 (トピック) を集約できているかについて，表 2 の全ツイートを用いて調査した．その結果，記事集約を行わない hop0 では 163,564 記事，集約を行う hop1 および hop2 ではそれぞれ 152,288 記事と 139,680 記事となっており，実際のツイート集合に対しても数万の記事を集約できていることがわかる．この結果から，前述の記事集約の問題を解決できれば，精度を落とすことなく同一トピックに付与される記事の分散を抑えられるため，記事集約が話題抽出に対してより有効に機能すると考えられる．

5. ま と め

本研究では，多言語な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を提案した．提案手法では，任意の言語で記述されたソーシャルメディアのテキストについて，テキスト中の各キーワードに対応する英語の Wikipedia の記事を言語間で比較可能なトピック情報として付与する．このとき，入力のテキスト集合中に出現する各キーワードについて，その前後に出現する語の統計情報を事前に収集することで，キーワードとして不適切な語句の抽出を抑制する．また，記事タイトルや記事が所属するカテゴリ情報を用いて意味的に似た記事を一つの記事に集約することで，テキストに付与される記事の意味的な粒度を均一化する．Twitter のデータを用いた評価実験により，提案手法が話題抽出に向けたエンティティリンキングとして有効であることを確認した．

今後の課題として，カテゴリ集約手法を再検討することが挙げられる．また，提案手法におけるキーワードの曖昧性解消の処理を改善し，エンティティリンキングの精度を向上させることや，話題抽出のタスクにより提案手法の有効性を検証することを検討している．

謝　　辞

本研究の一部は，文部科学省研究費補助金・基盤研究 A(26240013)，および，文部科学省国家課題対応型研究開発推進事業 －次世代 IT 基盤構築のための研究開発－「社会システム・サービスの最適化のための IT 統合システムの構築」(2012 年度–2016 年度) の助成による．

文　　献

[1] M. Cornolti, P. Ferragina, and M. Ciaramita, “A Framework for Benchmarking Entity-annotation Systems,” In WWW, pp.249–260, 2013.
[2] S. Cucerzan, “Large-Scale Named Entity Disambiguation Based on Wikipedia Data,” In EMNLP-CoNLL, pp.708–716, 2007.
[3] P. Ferragina, and U. Scaiella, “Fast and Accurate Annotation of Short Texts with Wikipedia Pages,” IEEE Software, vol.29, no.1, pp.70–75, 2011.
[4] J. Hoffart, M.A. Yosef, I. Bordino, H. Fürstenau, M. Pinkal, M. Spaniol, B. Taneva, S. Thater, and G. Weikum, “Robust Disambiguation of Named Entities in Text,” In EMNLP, pp.782–792, 2011.
[5] S. Kulkarni, A. Singh, G. Ramakrishnan, and S. Chakrabarti, “Collective Annotation of Wikipedia Entities in Web Text,” In KDD, pp.457–466, 2009.
[6] M. Lesk, “Automatic Sense Disambiguation Using Machine Readable Dictionaries: How to Tell a Pine Cone from an Ice Cream Cone,” In SIGDOC, pp.24–26, 1986.
[7] E. Meij, W. Weerkamp, and M. de Rijke, “Adding Semantics to Microblog Posts,” In WSDM, pp.563–572, Feb. 2012.
[8] R. Mihalcea, and A. Csomai, “Wikify!: Linking Documents to Encyclopedic Knowledge,” In CIKM, pp.233–242, 2007.
[9] D. Milne, and I.H. Witten, “An Effective, Low-cost Measure of Semantic Relatedness Obtained from Wikipedia Links,” In AAAI Workshop on Wikipedia and Artificial Intelligence, pp.25–30, July 2008.
[10] D. Milne, and I.H. Witten, “Learning to Link with Wikipedia,” In CIKM, pp.509–518, 2008.
[11] G. Salton, and C. Buckley, “Term-weighting Approaches in Automatic Text Retrieval,” Information processing & management, vol.24, no.5, pp.513–523, 1988.
[12] M. Shirakawa, T. Hara, and S. Nishio, “MLJ: Language-Independent Real-Time Search of Tweets Reported by Media Outlets and Journalists,” In VLDB, vol.7, no.13, pp.1605–1608, 2014.
[13] 中村達哉, 白川真澄, 原隆浩, 西尾章治郎, “Wikipedia を用いたソーシャルメディアからの言語横断的な話題抽出システムの試作,” 情報処理学会研究報告, vol.2014-DBS-160, no.11, pp.1–9, 2014.