DBSJ Journal, Vol.12, No.2
October 2013

Wikipedia の記事を素性としたベクトルの生成による短文間の言語非依存な関連度計算

Language-independent Short Text Similarity Measurements by Generating the Vector of Wikipedia Articles

中村 達哉♠ 白川 真澄♣
原 隆浩♣ 西尾 章治郎♣

Tatsuya NAKAMURA Masumi SHIRAKAWA
Takahiro HARA Shojiro NISHIO

本論文では，多言語 Web 事典である Wikipedia を用いた言語空間の統一により，異なる言語で記述された短文間の関連度を計算する手法を提案する．最近では，情報発信の即時性や地域性が重要視されており，世界中の人がそれぞれの言語でその地域に関する情報を常時発信している．しかし，このようなテキストは短文かつ様々な言語で記述されるという特徴を持っているため，テキスト間の関連度を計算することは困難である．提案手法ではこの問題に対し，任意の言語で記述された短文を，ある 1 つの言語の Wikipedia の記事を素性としたベクトルにより表現し，このベクトルを用いて関連度を計算することで解決する．英語，日本語およびタイ語の Twitter のデータを用いた評価実験の結果から，既存手法と比較して提案手法が有効であることを確認した．

In this paper, we proposed a language-independent method to measure the similarity between short texts written in different languages by unifying the language space using Wikipedia. In recent years, immediacy and locality of information dissemination have been regarded as important, and people around the world have been continuously transmitting information about their local area in their own languages. However, measuring the similarity between these texts is difficult because they are short and written in various languages. Our method solves this problem by representing short texts written in any languages using the vector of a certain language Wikipedia articles. From the experimental results using English, Japanese and Thai Twitter data, we confirmed that our method significantly outperformed comparative methods.

1. はじめに

自然言語で記述されたテキスト間の関連度を計算する手法は，意味を考慮したテキスト解析の重要な基盤技術である．特にWebは様々な種類のテキストが大量に存在するため，関連度計算を始めとする意味解析により，情報をまとめたり分類したりすることが必要である．また最近では，Web上での情報発信の即時性や地域性が重要視されており，世界中の人がそれぞれの言語でその地域に関する情報を常時発信することも増えている．例えばNew York Times[1]では，記者個人のソーシャルメディアのアカウントを自社の公認アカウントとし，現地の記者によるリアルタイムな情報発信を行っている．

このような即時性や地域性を考慮したWeb（あるいはソーシャルメディア）のテキストは，情報としての価値は高いが，短文かつ様々な言語で記述されるという特徴を持っているため，テキスト間の関連度を計算することは困難である．まず，短文は語句の絶対数が少なく，テキスト自身が持つ情報量が少ない．短文間の関連度を計算するためには，短文が持つ意味情報を拡張し，拡張した意味情報を用いて関連度計算を行う必要がある．また，異なる言語で記述されたテキストを対象とした関連度計算では，対象とするテキストの言語空間を1つに統一する必要がある．しかし，言語空間を統一するためには各言語間の対訳関係が必要であり，解析の対象とする言語数の増加に従い，必要な対訳関係の数は膨大になる．

そこで本研究では，異なる言語で記述された短文間の関連度を計算する手法を，Wikipediaとベイズ理論を用いて構築する．提案手法では，短文に対してWikipediaの記事（エンティティ）の付与を行う手法[10]を，Wikipediaの言語間リンクにより拡張することで，短文に対する意味情報の拡張と意味情報の言語空間の統一を同時に実現する．具体的には，任意の言語で記述されたテキストに対して，ある1つの言語のWikipediaの記事を用いたベクトルを作成する．この際，ベクトルに用いる記事の言語として，他の言語からの言語間リンクによって繋がっている記事が多い英語を採用する．そして，作成したベクトル同士の比較によって，異なる言語で記述された短文間の関連度を計算する．提案手法では，入力テキストの言語の記事から英語の記事への変換（マッピング）を確率として定義することにより，言語間リンクを持たない記事についても英語の記事へのマッピングを可能としている．

2. 関連研究

Wikipedia を知識抽出の対象とする研究（Wikipedia マイニング）は 2006 年に注目を集め，以降急速に研究対象としての認知度が高まっていった．Wikipedia は，Wiki をベースにした大規模 Web 百科事典であるため，誰でも Web ブラウザを通じて記事内容を変更できることが大きな特徴である．Wikipedia は，記事の網羅性や即時性だけでなく，密な記事間リンク，質の高いアンカーテキスト，URL による語義の一意性，280 以上の言語サポートと言語間リンクによる対訳関係の定義など，知識抽出のコーパスとして有利な性質を数多く持っている．

最近では，Wikipedia を用いて意味情報を拡張することにより，マイクロブログやニュースフィードなどに代表される短いテキストを解析する研究が注目を集めている．例えば，Meij らの研究[4]や Ferragina らの研究[1]では，Wikipedia から得られた情報を利用することにより，高精度で短文に対する曖昧性解消タスクを達成している．白川らの研究[10]で

♠ 学生会員 大阪大学 大学院情報科学研究科
nakamura.tatsuya @ist.osaka-u.ac.jp
♣ 正会員 大阪大学 大学院情報科学研究科
{shirakawa.masumi, hara, nishio}@ist.osaka-u.ac.jp

[1] http://www.nytimes.com/twitter

は，ナイーブベイズを拡張した手法により，短文の入力に対して関連する記事とその関連度を取得している（3.3 節で後述）．これらの研究では，統計的な手法では対応が困難な情報量の少ない短文に対して，Wikipedia という基盤知識を利用して意味情報を拡張するアプローチがとられている．これらの研究では，単一の言語を対象としている．

Wikipedia の言語間リンクを用いた言語横断に関する研究も数多く行われている．Navigli ら[8]は，Wikipedia，WordNet，および機械翻訳により，大規模かつ高品質な多言語意味ネットワーク BabelNet を自動で構築する手法を提案している．Navigli らはこれに続く研究として，BabelNet を用いた異なる言語の語句間の関連度計算手法である BabelRelate! [9]を提案している．しかし，BabelRelate!ではテキスト間の関連度計算は対象としていない．

Sorg ら[11]は，異なる言語で記述された語句あるいはテキスト間の関連度計算手法として Cross-Lingual Explicit Semantic Analysis（CL-ESA）を提案している．CL-ESA は単一言語を対象とした関連度計算手法である Explicit Semantic Analysis（ESA）[2]を言語間リンクによって拡張した手法である．それぞれの言語において ESA ベクトルを作成した後，ベクトルの基底を特定の言語への言語間リンクを持つ記事に制限することで，異なる言語で記述された語句およびテキスト間の関連度計算を実現している．Wikipedia を用いた関連度計算手法は数多く存在するが，その中でも ESA は任意の長さのテキストを対象としていることや比較的安定した性能が得られることから，現在最も良く使われる手法の 1 つとなっている．しかし，短文に対しての性能は前述の白川らの研究に劣っており，ESA は短文に対してうまく機能しないと考えられる．したがって，CL-ESA についても短文に対して精度の低下が生じることが考えられる．また，CL-ESA では言語間リンクを持つ記事のみを用いて ESA ベクトルを作成するため，言語に特有なトピックについて記述されたテキストや言語間リンクの少ない言語のテキストに対して，言語間リンクを持つ記事のみでテキストの意味情報を精度よく表現することが困難である．

3. 提案手法

本章ではまず，複数の言語で記述された短文間の関連度を計算する上で問題となる点について述べる．次に，提案手法の概要を紹介した後，単一言語を対象とした関連記事取得手法[10]について説明する．その後，単一言語を対象とした手法を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対して関連する英語の記事の付与を行う手法について述べる．

3.1 考慮すべき問題

異なる言語で記述された短文間の関連度を計算するタスクを考える．このとき，主に 2 つの問題が存在する．

第一の問題は，入力テキストが短いことである．短文は語句の絶対数が少なく，テキスト自身が持つ情報量が少ない．短文間の関連度を計算するには，短文が持つ意味情報を拡張し，拡張した意味情報を用いて関連度計算を行う必要がある．提案手法では，単一言語を対象とした短文に対する関連記事取得手法[10]を用いることでこの問題を解決する．

第二の問題は，入力テキストが様々な異なる言語で記述されていることである．異なる言語で記述されたテキスト間の関連度を計算するには，入力テキストの言語空間を 1 つの言語空間に統一する必要がある．そこで，付与する記事の言語空間を統一するため，言語間リンクにより入力テキストの言語の記事を別の言語の記事に変換（マッピング）することが求められる．

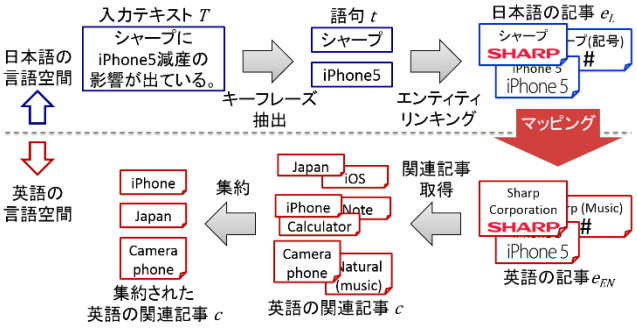


図 1　提案手法の流れ
Fig. 1　Example of our proposed method.

マッピングを行うタイミングについても考慮する必要がある．記事数や出現する語句，記事間リンク数などは各言語の Wikipedia によって異なるため，同じ意味のテキストであっても，言語によって付与される記事は異なることが予想される．そのため，本来は高い関連度が得られるべき同じ意味のテキストのペアに対して，得られる関連度が低くなるという問題が生じる．このような問題を回避するには，テキストに対する関連記事の付与において，早い段階でマッピングを行い，マッピング先の言語の Wikipedia から得られる情報を用いて関連記事の付与を行う必要がある．

また，単純に言語間リンクを持つ記事のみをマッピングするだけでは，短文の意味情報の拡張に用いる記事の数が言語間リンクの数に制限される．この場合，言語間リンクが少ないトピックや言語のテキストに対して意味情報を十分に拡張できない可能性がある．短文に対して言語空間が統一された意味情報を十分に拡張し，かつ言語間リンクが少ないトピックや言語に対応するには，言語間リンクを持たない記事についてもマッピングを行う必要がある．

3.2 手法の概要

3.1 節で説明した異なる言語で記述されたテキスト間の関連度計算に関する問題に対し，短文に対する意味情報の拡張と意味情報の言語空間の統一という 2 つの問題を同時に解決し，統一された言語空間で意味情報を付与する手法を提案する．

提案手法では，3.3 節で後述する単一言語を対象とした関連記事取得手法に言語空間の統一という処理を組み込むことで，短文間の言語非依存な関連度計算を実現する．このとき，Wikipedia が対応する言語の中で最も規模が大きく，他の言語の Wikipedia からの言語間リンクによって繋がっている記事が多い英語版 Wikipedia に言語空間を統一する．

図 1 は提案手法の処理の流れを表している．まず，対象とする言語（図 1 では日本語）で記述されたテキストからキーフレーズとなる語句を抽出し，各語句がどのような記事（エンティティ）を意味しているかを決定する．そして，言語間リンクを介して英語の記事にマッピングする．このとき，言語間リンクを持たない記事についても，類似した意味を持つ別の記事へのマッピングを行う．その後，関連する記事を取得する．ここまでの処理を確率的に行い，関連記事とそのスコア（確率）を決定する．

提案手法により，任意の言語で記述された短文に対して言語空間が統一された記事を付与できる．そして，付与した記事を素性としたベクトルを，短文の意味を補足する情報として用いることで，異なる言語で記述された短文間の関連度計算が可能となる．

3.3 単一言語を対象とした関連記事取得手法

単一言語を対象とした関連記事取得手法[10]（以下，単一言語手法）は，Wikipedia の記事の付与によるテキストの意味情報の拡張を目的としている．また，付与した記事を素性としたベクトルを用いることで，単一の言語で記述されたテキスト間の関連度計算を実現している．単一言語手法では，関連記事の取得における個別の問題に対して，既存の Wikipedia を用いた手法をベイズ理論の枠組みで再定義し，拡張ナイーブベイズを用いて統一的に解決することにより，ESA よりも高い精度で短文間の関連度計算を実現している．具体的には，Wikipedia から取得可能な情報をもとに，以下の式を用いて，入力テキストのキーフレーズ集合Tからテキストに関連する記事cとその確率$P(c|T)$を算出する．

$$P(c|T) = \frac{\sum_{k=1}^{K}\left(P(t_k \in T)P(c|t_k) + \left(1 - P(t_k \in T)\right)P(c)\right)}{P(c)^{K-1}} \quad (1)$$

$P(t_k \in T)$は，語句t_kがアンカーテキストとして記事中に出現する確率[5]であり，語句t_kがテキスト中のキーフレーズである確率を意味する．テキストからのキーフレーズ候補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．トライ木を用いて抽出するキーフレーズ候補は，Wikipedia で用いられている記事タイトルおよびアンカーテキストとする．

$P(c|t_k)$は，語句t_kから関連する記事cが連想される確率であり，語句t_kがアンカーテキストとして記事eにリンクされる確率$P(e|t_k)$[7]と，記事eから隣接リンク（フォワードリンクおよびバックワードリンク）をたどって記事cに到達する確率$P(c|e)$を用いて次式により算出できる．

$$P(c|t_k) = \sum_{e \in W} P(c|e)P(e|t_k) \quad (2)$$

Wは Wikipedia で定義されている記事集合である．$P(c)$は，Wikipedia の総リンク数に対する記事cの隣接リンク数の割合であり，記事cの一般度を意味する．

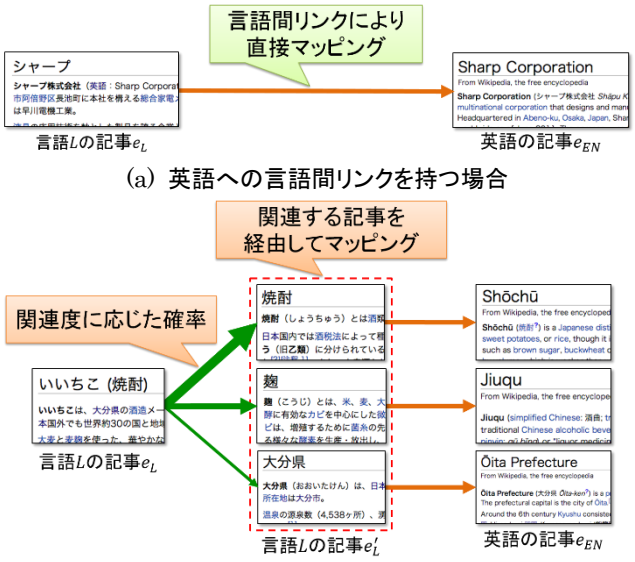
最終的に，式(1)により入力テキストTから関連記事cとその確率$P(c|T)$が取得できる．また，上位の関連記事を素性としたベクトルを用いることで，テキスト間の関連度を計算できる．

3.4 任意の言語のテキストに対する英語の記事付与

以下では，提案手法で導入するマッピング確率について述べた後，単一言語手法にマッピング確率をどのように組み込むかについて説明する．

3.4.1 マッピング確率

マッピング確率は，入力テキストの言語Lの記事e_Lが英語の記事e_{EN}への言語間リンクを持つ場合と持たない場合に分けて定義する．始めに，言語Lの記事e_Lが英語の記事e_{EN}への言語間リンクを持つ場合を考える．言語間リンクによって繋がっている英語の記事e_{EN}は言語Lの記事e_Lと同一の意味を持つ記事であるため，この場合はe_Lをe_{EN}に直接マッピング



(a) 英語への言語間リンクを持つ場合

(b) 英語への言語間リンクを持たない場合

図 2　マッピングの例

Fig. 2　Example of mapping process.

できる（図 2(a)）．このときのマッピング確率$P(e_{EN}|e_L)$を以下の式により定義する．

$$P(e_{EN}|e_L) = \begin{cases} 1 & , e_{EN}\ is\ linked\ from\ e_L \\ 0 & , otherwise \end{cases} \quad (3)$$

ここで，e_Lはただ 1 つのe_{EN}についてのみ言語間リンクを持つため，$\sum P(e_{EN}|e_L) = 1$である．

次に，言語Lの記事e_Lが英語への言語間リンクを持たない場合を考える．提案手法では，入力テキストに関連のある記事が取得できればよいため，厳密に対訳関係にある英語の記事にマッピングする必要がない．そこで，英語への言語間リンクを持つ言語Lの別の記事e'_Lを経由させて言語Lの記事e_Lを英語の記事e_{EN}へマッピングする．このとき，記事e_Lと記事e'_Lとの関連度に応じた確率を定義することで，できる限り類似した意味を持つ英語の記事へマッピングされるようにする（図 2(b)）．なお，全ての記事ペアに対して関連度を算出しようとすると膨大な計算量になるため，英語への言語間リンクを持ち，かつ記事e_Lのフォワードリンクによって繋がっている記事$e'_L \in R_L$のみを対象として関連度を算出する．

言語Lの記事e_Lとe'_Lとの関連度を$Sim(e_L, e'_L)$とし，記事e'_Lが持つ言語間リンクによって繋がっている英語の記事をe_{EN}としたとき，e_Lからe_{EN}へマッピングされる確率$P(e_{EN}|e_L)$を以下の式により定義する．

$$P(e_{EN}|e_L) = \frac{Sim(e_L, e'_L)}{\sum_{e'_L \in R_L} Sim(e_L, e'_L)} \quad (4)$$

また，関連度$Sim(e_L, e'_L)$は次式により定義する[6]．

$$Sim(e_L, e'_L) = \frac{\log(|W_L|) - \log(\min(|E_L|, |E'_L|))}{\log(max(|E_L|, |E'_L|)) - \log(|E_L \cap E'_L|)} \quad (5)$$

W_Lは言語Lの Wikipedia で定義されている記事集合である．

E_LおよびE'_Lはそれぞれ記事e_Lおよびe'_Lのバックワードリンクによって繋がっている記事集合である．

3.4.2 単一言語手法の拡張

3.4.1 項で定義したマッピング確率$P(e_{EN}|e_L)$を，単一言語手法に組み込む．3.1 節で説明したように，異なる言語で記述されている同一の意味のテキストに対して同一の記事を付与するためには，できる限り早い段階でマッピングを行う必要がある．そこで提案手法では，関連記事取得の前にマッピングを行い，英語版 Wikipedia の中で関連記事を取得することで，入力テキストの言語の Wikipedia と英語版 Wikipedia の差異による影響を軽減する．

マッピング確率$P(e_{EN}|e_L)$により，式(2)の語句tから関連記事cが連想される確率$P(c|t)$は，言語Lの語句tから英語の関連記事cが連想される確率に更新できる．

$$P(c|t) = \sum_{e_{EN} \in W_{EN}} P(c|e_{EN}) \sum_{e_L \in W_L} P(e_{EN}|e_L) P(e_L|t) \quad (6)$$

W_{EN}は英語の Wikipedia で定義されている記事集合である．ここで，確率$P(e_L|t)$および確率$P(c|e_{EN})$はそれぞれ言語Lおよび英語の Wikipedia から算出された確率である．

単一言語手法と同様に，以下の式を用いて言語LのテキストTに対して関連する英語の記事cとその関連度$P(c|T)$が取得できる．

$$P(c|T) = \frac{\sum_{k=1}^{K} \Big(P(t_k \in T) P(c|t_k) + \big(1 - P(t_k \in T)\big) P(c) \Big)}{P(c)^{K-1}} \quad (7)$$

ここで，確率$P(t \in T)$および確率$P(c)$はそれぞれ言語Lおよび英語の Wikipedia から算出された確率である．式(7)は式(1)と同一の式であり，提案手法は単一言語手法と同じ枠組みの中で言語空間が統一された記事の付与を実現している．

4. 評価実験

4.1 評価環境

提案手法の関連度計算の性能を評価するために，Twitter のデータ（ツイート）のクラスタリングを行った．テキストクラスタリングでは，同じクラスタリングアルゴリズムを用いた場合，どのようにテキスト間の意味的距離（関連度）を計測するかが性能に影響を与えるため，クラスタリングの性能により関連度計算の性能を評価できる．具体的な評価方法として，Twitter のハッシュタグを元にあらかじめ正解集合を定義しておき，提案手法による関連度に基づいて K-means クラスタリングを実行した．ハッシュタグとは，ツイートを発信するユーザが意図的に「#MLB」や「#iPhone」のようにキーフレーズの直前に「#」をつけたものであり，そのツイートが言及しているトピックを明示的に表現する役割を持っている[3]．そのため，ハッシュタグを用いて短文クラスタリングのための擬似的な正解データを生成できる[10]．ここでは，ハッシュタグによる正解データができるかぎり正しいクラスタとなるよう，ハッシュタグのキーフレーズとして，曖昧性が低く，互いにトピックが独立しそうなものを選択した．

評価に用いた 2 種類のデータセットを表 1 に示す．各データセットでは，類似したトピックの中で異なるコンテキストを持つクラスタを想定し，情報技術（IT）とスポーツ（Sport）からそれぞれハッシュタグを選択した．データセットの作成手順として，1）各ハッシュタグによる検索を行い，英語，日本語，およびタイ語で記述されたツイートをそれぞれ収集，2）同じデータセット内の別のハッシュタグを含むツイートを削除，3）リツイート（「RT」で始まり，他人のツイートの引用を表す），URL の除去，4）ツイートの末尾にあるハッシュタグは全て除去し，それ以外のハッシュタグは「#」のみを除去，5）3 単語以下の英語のツイートおよび 10 文字以下の日本語・タイ語のツイートを削除，の各処理を行った．各データセットの統計値を表 1 にまとめる．

表 1　評価に用いた二つのデータセットと統計値
Table 1　Two datasets for evaluation and their statistics.

データセット名	情報技術(IT)	スポーツ(Sports)
タグ/ツイート数 （英/日本/タイ）	#Google / 2,493 (1,366/612/515) #iPhone / 2,642 (1,147/1,358/137) #Python / 956 (904/52/0) #Microsoft / 2,094 (973/796/325) #Firefox / 815 (551/30/234)	#F1 / 2,784 (1,394/1,217/173) #Golf / 3,553 (1,685/1,602/266) #MLB / 3,070 (1,931/1,139/0) #NBA / 2,954 (1,838/997/119) #SerieA / 1,524 (1,032/492/0) #Boxing / 1,341 (868/159/314)
総ツイート数 （英/日本/タイ）	9,000 (4,941/2,848/1,211)	15,226 (8,748/5,606/872)
ツイートあたりの テキストの情報量	英：12.7 語 日本：58.0 文字 タイ：69.2 文字	英：13.5 語 日本：44.7 文字 タイ：60.2 文字

比較手法として，日本語およびタイ語のテキストから，単一言語手法により，それぞれ日本語およびタイ語の記事を取得した後，英語への言語間リンクを持つ記事のみを英語の記事に置き換えて付与する手法（単純マッピング），および Sorg らの CL-ESA [11]によって得られた関連記事を素性とする手法（CL-ESA）を採用した．提案手法および比較手法では，それぞれ関連する記事の上位 10，20，50，100，200，500，1,000 を素性ベクトルとしてクラスタリングを行った．クラスタリングの評価指標には正規化相互情報量（NMI）[12]を用いた．NMI はすべてのクラスタの状態を情報理論的な解釈によって表現した指標であり，最もよく使われる指標の 1 つである．NMI のスコアは 0 から 1 までの値をとり，値が大きいほどクラスタリングの性能が高いことを意味する．評価実験では，それぞれの手法において初期値を変えて 20 回ずつ K-means クラスタリングを実行した時の平均値を採用した．

4.2 評価結果

ツイートのクラスタリング結果を表 2 に示す．表 2 の各列は，データセットごとに英語・日本語・タイ語の全ツイートをクラスタリングした結果（全言語），全言語のツイートをクラスタリングした後，それぞれの言語のツイートだけを抽出した場合の結果（英語，日本語，タイ語）を示している．

表 2　クラスタリングの結果
Table 2　The result of clustering.

データセット		IT				Sports			
言語		全言語	英語	日本語	タイ語	全言語	英語	日本語	タイ語
CL-ESA	（上位 10）	0.119	0.198	0.054	0.006	0.049	0.078	0.033	0.006
CL-ESA	（上位 20）	0.123	0.195	0.056	0.005	0.054	0.087	0.035	0.009
CL-ESA	（上位 50）	0.140	**0.239**	**0.065**	**0.007**	0.069	0.107	0.035	**0.015**
CL-ESA	（上位 100）	**0.144**	0.237	0.063	**0.007**	0.075	0.111	0.042	**0.015**
CL-ESA	（上位 200）	0.132	0.210	0.060	**0.007**	0.084	0.122	0.049	0.011
CL-ESA	（上位 500）	0.118	0.160	0.059	0.003	**0.094**	**0.132**	**0.057**	0.001
CL-ESA	（上位 1,000）	0.119	0.143	0.056	0.000	0.093	0.123	**0.057**	0.002
単純マッピング	（上位 10）	0.103	0.107	0.141	0.117	0.094	0.121	0.101	0.057
単純マッピング	（上位 20）	0.134	0.155	0.148	**0.123**	0.162	0.195	0.176	0.061
単純マッピング	（上位 50）	0.176	0.221	0.187	0.088	0.230	0.294	0.227	0.118
単純マッピング	（上位 100）	0.195	0.245	0.223	0.099	0.269	0.348	0.255	**0.128**
単純マッピング	（上位 200）	0.203	0.244	0.240	0.086	0.289	0.376	0.291	0.097
単純マッピング	（上位 500）	**0.217**	**0.303**	**0.254**	0.074	**0.311**	**0.395**	**0.330**	0.077
単純マッピング	（上位 1,000）	0.207	0.283	0.199	0.053	0.291	0.385	0.302	0.068
提案手法	（上位 10）	0.094	0.123	0.116	0.075	0.121	0.145	0.141	0.041
提案手法	（上位 20）	0.156	0.196	0.183	0.097	0.192	0.217	0.215	0.051
提案手法	（上位 50）	0.217	0.243	0.292	0.132	0.268	0.286	0.300	0.070
提案手法	（上位 100）	0.271	0.303	0.340	0.145	0.331	0.355	0.360	0.087
提案手法	（上位 200）	0.275	0.293	0.352	0.134	**0.369**	**0.373**	0.419	0.103
提案手法	（上位 500）	**0.340**	**0.362**	**0.407**	**0.179**	0.368	0.365	**0.427**	0.128
提案手法	（上位 1,000）	0.332	0.361	0.391	0.178	0.348	0.351	0.395	**0.162**

また，各手法および言語におけるスコアの最大値は太字で表している．

はじめに，両データセットにおける全言語の結果について考察する．CL-ESA や単純マッピングと比較して，提案手法が高いスコアを達成している．CL-ESA は，短文に対してうまく機能しない ESA をベースにしていることや，言語間リンクの数により付与できる記事が制限されることから，異言語のツイート間の関連度をうまく計算できていないと考えられる．単純マッピングでも CL-ESA と同様に言語間リンクの数の制約を受けているために性能が低下していると予測できる．

両データセットにおけるそれぞれの言語において，提案手法と単純マッピングのスコアの平均値の差について両側t検定を行ったところ，データセット Sports の英語のスコアとの差を除いて$p < 0.01$であり，差が有意であることがわかった．これは，データセット Sports においては，提案手法と単純マッピングの間で英語のツイートに関する性能に差が無い一方，それ以外のすべての場合において提案手法の性能が優れていることを意味している．

各データセットの日本語とタイ語のスコアに関してみると，提案手法は単純マッピングに比べて両方のデータセットでスコアが高くなっている．これは，単純マッピングにおいては，入力テキストの言語空間で関連記事を取得した後，英語への言語間リンクを持つ記事のみを用いているため，言語によって付与される関連記事が異なっていることが原因であると考えられる．一方，提案手法では，入力テキスト中のキーフレーズが意味する記事が得られた直後に英語の記事にマッピングし，英語の言語空間の中で関連記事を取得している．そのため，同じ意味を表すツイートに対して，言語に関わらず同一の関連記事が付与されやすくなり，日本語とタイ語のスコアが向上したと考えられる．

提案手法により英語のツイートに付与される関連記事は，単純マッピングと全く同じであるのに対して，提案手法におけるデータセット IT の英語のスコアは単純マッピングよりも高くなっている．これは，データセット IT の各トピックについて，各言語の正解クラスタの差異が小さく，日本語での精度向上が英語にも影響したためであると考えられる．一方，データセット Sports では，同じハッシュタグを付けていても言語によって言及しているトピックがずれているために各言語の正解クラスタの差異が大きく，提案手法により日本語の精度は向上しても英語の精度には影響が無かったものと考えられる．

タイ語のクラスタリングの性能は，他の言語と比較して全体的に低くなっている．提案手法における英語，日本語，およびタイ語の結果について，最も正解の多いクラスタのみを考慮した単純な正解率で比較した場合，データセット IT においては，それぞれ 53%，74%，55%，データセット Sports においては，54%，65%，42%である．データセット IT において，タイ語と英語の正解率は同程度である一方，タイ語の NMI のスコアは英語の約半分である．つまり，タイ語で

は，最も正解の多いクラスタ以外に残りのツイートが分散しており，うまくクラスタリングできていないことがわかる．そこで，タイ語のツイートについて調査したところ，ツイート中のキーフレーズの表記が揺らいでいることがわかった．また，タイ語の Wikipedia の記事数は約 8 万と少ないため，キーフレーズの候補となる語句数が少ない．そのため，タイ語のツイートからキーフレーズを取得できないことが多く，適切な関連記事を付与できないという問題が生じている．

5. まとめと今後の課題

本研究では，Wikipedia の記事を素性としたベクトルによる短文間の言語非依存な関連度計算手法を提案した．具体的には，任意の言語で記述された短文に対する意味情報として，関連する英語の記事（エンティティ）を付与する．そして，付与した英語の記事を素性としたベクトルを用いて，異なる言語で記述された短文間の関連度を計算する．提案手法では付与する記事の言語空間を統一するため，Wikipedia の言語間リンクを用いて入力テキストの言語の記事を英語の記事に変換（マッピング）する．このとき，マッピングを確率として扱うことで，言語間リンクを持たない記事についても，関連する別の記事を経由することで英語の記事にマッピングすることを可能にしている．マッピング確率を，単一言語を対象とした関連記事取得手法に組み込むことで，短文に対する意味情報の拡張だけでなく，意味情報の言語空間の統一，言語非依存な関連度計算を 1 つの枠組みの中で実現した．Twitter のデータを用いた評価実験により，提案手法が複数の言語で記述された短文間の関連度計算において有効であることを確認した．

今後の課題として，語句を抽出する処理をよりロバストにすることが挙げられる．提案手法では，キーフレーズの候補として Wikipedia の記事タイトルおよびアンカーテキストを採用していたが，アンカーテキスト以外の記事中に出現する語句についても記事を表すキーフレーズの候補として用いることでタイ語のような Wikipedia のサイズが小さい言語への対応を強化できると考えられる．

[謝辞]

本研究の一部は，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] P. Ferragina, and U. Scaiella: "TAGME: On-the-fly Annotation of Short Text Fragments (by Wikipedia Entities)," in Proceedings of ACM Conference on Information and Knowledge Management (CIKM), pp.1625-1628 (2010).

[2] E. Gabrilovich, and S. Markovitch: "Computing Semantic Relatedness Using Wikipedia-based Explicit Semantic Analysis," in Proceedings of International Joint Conference on Artical Intelligence (IJCAI), pp.1606-1611 (2007).

[3] D. Laniado, and P. Mika: "Making Sense of Twitter," in Proceedings of International Semantic Web Conference (ISWC), pp.470-485 (2010).

[4] E. Meij, W. Weerkamp, and M. de Rijke: "Adding Semantics to Microblog Posts," in Proceedings of ACM International Conference on Web Search and Data Mining (WSDM), pp.563-572 (2012).

[5] R. Mihalcea, and A. Csomai: "Wikify!: Linking Documents to Encyclopedic Knowledge," in Proceedings of ACM Conference on Information and Knowledge Management (CIKM), pp.233-242 (2007).

[6] D. Milne, and I.H. Witten: "An Effective, Low-cost Measure of Semantic Relatedness Obtained from Wikipedia Links," in Proceedings of AAAI Workshop on Wikipedia and Artificial Inteligence (WIKIAI), pp.25-30 (2008).

[7] D. Milne, and I.H. Witten: "Learning to Link with Wikipedia," in Proceedings of ACM Conference on Information and Knowledge Management (CIKM), pp.509-518 (2008).

[8] R. Navigli, and S.P. Ponzetto: "BabelNet: Building a Very Large Multilingual Semantic Network," in Proceedings of Annual Meeting of the Association for Computational Linguistics (ACL), pp.216-225 (2010).

[9] R. Navigli, and S.P. Ponzetto: "BabelRelate! A Joint Multilingual Approach to Computing Semantic Relatedness," in Proceedings of National Conference on Artificial Intelligence (AAAI), pp.22-26 (2012).

[10] 白川真澄，中山浩太郎，原隆浩，西尾章治郎："Wikipedia とナイーブベイズを用いた自然文に対する関連語句取得手法," データ工学と情報マネジメントに関するフォーラム (DEIM), (2012).

[11] P. Sorg, and P. Cimiano: "Cross-lingual Information Retrieval with Explicit Semantic Analysis," in Proceedings of Working Notes for the Cross-Language Evaluation Forum 2008 Workshop, (2008).

[12] A. Strehl, and J. Ghosh: "Cluster Ensembles - a Knowledge Reuse Framework for Combining Multiple Partitions," Journal of Machine Learning Research, vol.3, pp.583-617 (2002).

中村 達哉 Tatsuya NAKAMURA
大阪大学大学院情報科学研究科博士前期課程在学中. Web マイニングに関する研究に興味を持つ. 情報処理学会学生会員.

白川 真澄 Masumi SHIRAKAWA
大阪大学大学院情報科学研究科特任助教. 2013 年大阪大学大学院情報科学研究科博士後期課程修了. 博士（情報科学）. Web マイニングに関する研究に従事. 情報処理学会，自然言語処理学会各会員.

原 隆浩 Takahiro HARA
大阪大学大学院情報科学研究科准教授. 1997 年大阪大学大学院工学研究科博士前期課程修了. 工学博士. データベースシステム，分散処理の研究に従事. 本会より上林奨励賞を受賞. IEEE, ACM, 情報処理学会，電子情報通信学会各会員.

西尾 章治郎 Shojiro NISHIO
大阪大学大学院情報科学研究科教授，サイバーメディアセンター長. 1975 年京都大学工学部卒業. 1980 年同大学院工学研究科博士後期課程修了，工学博士. 京都大学工学部助手等を経て，1992 年大阪大学工学部教授となり，現職に至る. 文部科学省科学官，大阪大学理事・副学長等を歴任. データ工学の研究に従事. 本会理事，監事を歴任し，現在，会長を務める. 紫綬褒章を受章し，本会より功労賞，論文賞を受賞.